Aprendizaje de patrones relacionales para la extracción de información en apoyo a la toma de decisiones en medicina

José A. Reyes-Ortiz, Ana L. Jiménez, Jaime Cater, Cesar A. Meléndez, Patricia B. Márquez, Marlon García, Fernando Olvera, Juan C. Contreras, Gustavo Farfán

Health Digital Systems
Insurgentes Sur 617, 03810, Distrito Federal, México.
{alejandro.reyes,ajimenez,jaimecater,cmelendez,pmarquez,
marlon.garcia,fernando.olvera,ccontreras,
gfarfan}@saludhds.com.mx
http://www.saludhds.com.mx

(Paper received on June 30, 2013, accepted on August 15, 2013)

Resumen. La obtención de un diagnóstico definitivo en los sistemas de información basado en el Expediente Clínico Electrónico (ECE) se realiza por un médico, el cual analiza variables como síntomas y antecedentes del paciente. Este proceso puede ser realizado de manera incorrecta debido, entre otros factores, a la interpretación inexacta de la información descrita en lenguaje natural. Esto refleja la importancia de generar sistemas de extracción de información médica a partir de textos como apoyo a los médicos en el proceso de toma de decisiones diagnósticas. Estos sistemas se basan en patrones semánticos que localicen la información relevante. En este artículo se presenta un enfoque de aprendizaje automático para la construcción de patrones semánticos relacionales, los cuales extraen la información necesaria para apoyar el proceso generación de diagnósticos probables en el campo de la medicina. El aprendizaje utiliza un corpus de notas médicas para generar los patrones relacionales, generalizarlos y filtralos con base en su frecuencia de aparición.

Palabras clave: aprendizaje automático, diagnóstico médico, patrones semánticos relacionales, extracción de información.

1 Introducción

La extracción de información en el dominio médico ha despertado un interés creciente en los últimos años. Es necesario crear sistemas que identifiquen y realicen la extracción de información relevante de documentos médicos para un objetivo específico. El tipo de información extraída depende de su uso posterior, el cual determina su formato prefijado. Por ejemplo, un sistema que genera una lista de diagnósticos probables de manera automática necesita datos del paciente, síntomas, partes anatómicas, entre otras. Este sistema necesita de métodos que extraigan dicha información de los docu-

Special Issue in Information Technology
Research in Computing Science 64, 2013, pp. 33-42

mentos médicos generados en una consulta. Actualmente, esta información se regula por las reglas para generar sistemas de Expediente Clínico Electrónico [1] de cada paciente.

Los métodos de extracción de información a partir de textos necesitan patrones sintácticos, semánticos o lingüísticos para localizar los datos correspondientes. La generación de estos patrones se puede realizar de manera manual, sin embargo, esta opción conlleva diversos inconvenientes como el alto consumo de tiempo, los costos y se necesita de humanos para realizar todo el proceso. Una propuesta a este problema es el aprendizaje automático, el cual se encarga de proponer técnicas para que los dispositivos electrónicos *aprendan* a partir de ejemplos, también conocido como aprendizaje supervisado [2]. El aprendizaje automático se puede aplicar en la construcción automática de patrones, tarea que consiste en descubrir un conjunto de patrones textuales a partir de ejemplos proporcionados a la aplicación.

El diagnóstico médico es un procedimiento mediante el cual se identifica una enfermedad, síndrome o cualquier condición de salud de un paciente haciendo uso de diversas variables entre las que destacan la sintomatología y factores de riesgo.

Los diagnósticos pueden ser generados en los sistemas de Expediente Clínico Electrónico (ECE) de manera manual por un médico, el cual selecciona el diagnóstico definitivo del catálogo internacional de enfermedades llamado CIE10 [3]. Este proceso es tedioso, consume bastante tiempo y pueden ser generados de manera incorrecta. Se ha registrado que el porcentaje de errores médicos durante la atención hospitalaria es de entre 3.5% al 16.6% [4]. Estos problemas reflejan la importancia de generar sistemas de información que permitan apoyar a los médicos durante la toma de decisiones en el proceso de diagnóstico definitivo. Para que un sistema de este tipo sea una realidad es necesario el uso de técnicas de extracción de información en las notas médicas. Estas técnicas necesitan patrones los cuales se pueden generar de manera semiautomática.

Este artículo propone un enfoque de aprendizaje automático para generar un conjunto de patrones de extracción de información con la finalidad de alimentar un sistema de generación de diagnósticos probables. El aprendizaje presentado parte de un corpus de notas médicas anotadas con entidades nombradas que intervienen en un diagnóstico médico: anatomía, enfermedades, síntomas, intensidades, entre otras. A partir de estos datos el enfoque propuesto genera un conjunto de patrones de extracción enfocados en relacionar las entidades de síntoma con anatomía, paciente con síntoma, síntoma con evolución, paciente con antecedente y antecedente con familiar. Los patrones obtenidos se generalizan por sinonimia, se filtran por frecuencia y se validan por un experto médico.

El resto de este artículo se organiza de la siguiente manera. La sección 2 presenta las bases teóricas de la tarea de aprendizaje y se describen los elementos que intervienen en el proceso de obtención de un diagnóstico médico. En la sección 3 se presenta la descripción del corpus utilizado para el entrenamiento y las pruebas. El enfoque propuesto de aprendizaje de los patrones, el cual incluye una tarea de obtención de las raíces de las palabras, la generación de patrones y un filtrado se exponen en la sección 4. En la sección 5 se presenta la evaluación de los patrones en términos de precisión y su frecuencia de aparición en el corpus. Finalmente, en la sección 6 se muestran las conclusiones.

2 Bases teóricas

El trabajo presentado en el presente artículo se basa en un enfoque de aprendizaje de patrones con una intervención de expertos médicos para validar los resultados. Por ello, en esta sección se presentan las bases del aprendizaje de patrones, los elementos necesarios para realizar un diagnóstico médico y los trabajos relacionados en estas áreas.

2.1 Diagnóstico médico

El diagnóstico médico es un procedimiento mediante el cual se identifica una enfermedad, síndrome o cualquier condición de salud. Este proceso se lleva a cabo haciendo uso de diversos criterios, tales como: la sintomatología, factores de riesgo, signos vitales, datos personales del paciente como sexo y edad, estudios de laboratorio y estudios de gabinete.

Los criterios más relevantes para realizar un diagnóstico son los síntomas con su anatomía y evolución, los factores de riesgo como antecedentes y el familiar que lo padece. Los síntomas se identifican mediante el proceso de impresión diagnóstica, la cual consiste en un razonamiento clínico con el objetivo de explicar la enfermedad. Para consolidar un diagnóstico es necesario identificar los factores de riesgo internos y externos. Una vez determinado el diagnóstico definitivo se procede con el tratamiento, el control y seguimiento de la enfermedad.

En el área de la atención clínica, el proceso descrito anteriormente se realiza por un médico asignado a las consultas. Este proceso puede ser realizado de manera incorrecta, derivado de diversos factores entre los errores más frecuentes documentados son: mal uso de la relación médico-paciente, interrogatorio mal dirigido y la incorrecta interpretación clínica de la semiología [5]. Como apoyo a este proceso se han propuesto enfoques computacionales automáticos o semiautomáticos para la generación de diagnósticos probables, los cuales se discuten en la sección de trabajos relacionados.

2.2 Aprendizaje de patrones

La obtención automática de patrones es el principal obstáculo de un sistema de extracción de información a partir de texto. En el área de la generación de diagnósticos probables, el aprendizaje los patrones que relacionen la información del paciente con sus síntomas, los síntomas con su anatomía y evolución y los pacientes con sus antecedentes se han convertido en un reto para el aprendizaje automático.

El aprendizaje puede ser realizado de manera supervisada, el cual consiste en implementar algoritmos para que las máquinas *aprendan* a partir de ejemplos disponibles [10]. En la extracción de información para obtener un diagnóstico médico, la idea es que el aprendizaje ayude a construir el conjunto de patrones relevantes para dicha tarea.

El aprendizaje se puede aplicar para la construcción automática de conjuntos de patrones de extracción de información. En la literatura existen propuestas que hacen la construcción de diccionarios de patrones para la extracción de información en general. Estas propuestas y enfoques se discuten, más adelante, en la sección 3.

3. Trabajos relacionados

Las áreas de diagnósticos médicos y el aprendizaje de patrones convergen en este trabajo con la finalidad de obtener un conjunto de patrones semánticos relacionales para la obtención de información relevante a partir de los textos médicos. El objetivo de esta identificación es apoyar la toma de decisiones en los diagnósticos clínicos.

La generación de diagnósticos probables ha sido abordada por enfoques computacionales automáticos y semiautomáticos. En este aspecto existen dos tipos de enfoques: a) los enfoques estadísticos basados en aprendizaje, tales como [6] que presenta una actualización del algoritmo C4.5 para '*aprender*' a partir de ejemplos y poder determinar el diagnóstico, un enfoque estadístico mediante el algoritmo de Máquinas de Soporte Vectorial (MSV) es expuesto en [7]; b) finalmente, los enfoques basados en reglas de extracción de información a partir de textos, las cuales son generadas por expertos de manera manual[8] y [9].

Por su parte, el aprendizaje de patrones ha sido abordado desde diversas perspectivas para dominios ajenos a la medicina. De esta manera, se analiza la literatura correspondiente donde se propone la construcción automática de diccionarios de patrones de extracción de información, el cual se basa en corpus textuales sin anotaciones y consta de diversas etapas, de las cuales destaca la generación de patrones específicos, la generalización de patrones y el filtrado [11]. El aprendizaje de patrones de relación en presentado en [12] y [13] donde se expone un sistema para adquirir patrones a partir de documentos etiquetados con partes de la oración y clases semánticas. En el campo del aprendizaje de reglas de extracción de información en dominios genéricos se han propuesto diversos trabajos ([14], [15] y [16]).

Los patrones aprendidos de manera automática representan las estructuras semánticas del corpus de entrenamiento en cuestión de los elementos que caracterizan, de esta manera, se tienen patrones semánticos relacionales que generalizan las estructuras de los textos para la información que se desea extraer. Estos patrones sirven para detectar información relevante en textos futuros.

La investigación sobre el aprendizaje de patrones en la literatura existente es nula. Por lo tanto, nuestro enfoque aplica aprendizaje automático con la finalidad de generar los patrones semánticos relacionales, los cuales son necesarios para identificar a partir de los textos de las notas médicas la siguiente información: las relaciones de síntomas con anatomía y evolución; la relación de paciente con sus síntomas; la relación de paciente con antecedentes heredofamiliares y patológicos; y la relación de antecedentes con familiares que los padecen o padecieron.

Los patrones ayudan a identificar la información descrita anteriormente, la cual es utilizada como apoyo a la tarea del médico para la toma de decisiones en los procesos de diagnósticos clínicos. Para que los patrones puedan extraer el mismo tipo de información en nuevos textos, es necesario contar con un corpus de entrenamiento lo suficientemente extenso para generalizar los textos de las notas médicas del dominio deseado. Por ello, en la siguiente sección se describe el corpus creado y validado, manualmente, por expertos médico, y que finalmente es utilizado para el aprendizaje de los patrones semánticos relacionales.

4. Corpus de notas medicas

El corpus utilizado para nuestro enfoque consta de 2500 notas médicas que son utilizadas para generar un diagnóstico a partir de los datos que contienen (síntomas, anatomía, intensidad, antecedentes personales, antecedentes patológicos, entre otros). Este corpus está compuesto de 13,000 oraciones y 156,000 palabras. Este corpus es dividido en dos partes, 2000 notas médicas utilizadas para el aprendizaje de patrones de extracción y 500 notas médicas utilizadas para la fase de pruebas de los patrones.

El aprendizaje de los patrones se lleva a cabo con el corpus de entrenamiento de notas médicas. Por su parte, el extracto de corpus extraído para las pruebas es utilizado para la evaluación de los patrones extraídos, la cual se basa en determinar la precisión de los patrones, además de presentar un estudio de frecuencia de aparición de cada patrón.

Este corpus fue construido y validado manualmente por expertos médicos. Éste se utiliza para aprender los patrones que generalizan las estructuras semánticas en lenguaje natural de los elementos necesarios para llevar a cabo un diagnóstico. A partir de este corpus se aplica el enfoque propuesto, el cual se describe en la siguiente sección.

5. Enfoque propuesto

Las distintas aproximaciones presentadas en las secciones anteriores se basan en aprendizaje de reglas o patrones para dominios ajenos a la medicina. En la obtención de diagnósticos médicos probables, se han propuesto trabajos enfocados directamente en aprendizaje con estadística. El enfoque propuesto se basa en el aprendizaje de patrones relacionales de extracción de información utilizada para la obtención de diagnósticos probables en el área de medicina. El objetivo es reducir la intervención y esfuerzo del experto humano en la creación de los patrones de extracción. Para conseguir este objetivo el enfoque propuesto utiliza las raíces de las palabras para la obtención y generalización de patrones, esto elimina la intervención del experto para analizar el volumen de información. El experto interviene hasta la etapa de filtrado donde trabaja directamente con los patrones extraídos en lugar de trabajar con el corpus.

Nuestro enfoque se describe detalladamente en las siguientes secciones donde se expone el funcionamiento de cada fase.

5.1 Obtención de las raíces de las palabras

El proceso truncamiento (en inglés *stemming*) se lleva a cabo con la finalidad de reducir las palabras del corpus a sus raíces y con ello evitar la duplicidad de los patrones. Este proceso consiste en remover los sufijos comunes morfológicos e inflexionales de las palabras [17].

En nuestro enfoque se utiliza el algoritmo de *Porter* como un proceso de normalización mediante la obtención de la raíz de las palabras. Este algoritmo consiste en reglas para la eliminación de sufijos y acentos en las palabras [18]. Algunos ejemplos

de este proceso son: *embarazo confirmado:embaraz confirm; cirugía de rodilla:cirug de rodil; abdom:abdomen.*

Una vez que se realiza la obtención de las raíces de las palabras, el enfoque realiza el etiquetado de las entidades nombradas de los diagnósticos y la obtención de los patrones relacionales entre las entidades.

5.2 Etiquetado de las entidades nombradas

El etiquetado de entidades nombradas es el proceso de asignar una categoría semántica a una palabra o secuencias de palabras [19]. El objetivo de esta fase consiste en obtener un corpus etiquetado con entidades de los diagnósticos que servirá como base para la obtención de patrones relacionales.

En este enfoque se consideran las siguientes categorías semánticas de entidades: nombres de enfermedades, partes anatómica, síntomas, familiares, unidades de tiempo (día, mes, semana, hora, minuto) y cantidades (20, dos, una, 30). Para realizar el etiquetado semántico de entidades diagnósticas se utilizan listas de las entidades expresadas en su raíz. En la Tabla 1 se muestra el etiquetado mediante una lista de entidades completas y sus raíces de la categoría semántica "síntomas" para un texto de una nota médica.

Tabla 1. Lista de entidades de la categoría "síntomas"

Texto	Entidades	Raíces de las entidades
acude traído por la madre por presentar palpitaciones y agitación sin predominio de horaria, estreñimiento crónico, dolor cólico abdominal. rinorrea anterior, obstrucción nasal bilateral, prurito nasal, prurito faríngeo de evolución crónica.	palpitaciones	palpit
	agitación	agit
	estreñimiento crónico	estreñ cronic
	dolor cólico abdominal	dolor colic abdominal
	rinorrea anterior	rinorre anterior
	obstrucción nasal bilateral	obstruccion nasal bilateral
	prurito nasal	prurit nasal
	prurito faríngeo	prurit faringe

El corpus con las palabras en sus raíces y etiquetado con todas las categorías semánticas de los diagnósticos se utiliza para la obtención del conjunto de patrones relacionales.

5.3 Obtención de patrones relacionales

En esta fase el enfoque propuesto utiliza el corpus con las raíces de las palabras y con anotaciones de las entidades nombradas de los diagnósticos para obtener los patrones relacionales entre dichas entidades. El objetivo de esta fase es obtener un conjunto de patrones que serán validados y filtrados por los expertos médicos.

La obtención de patrones relacionales específicos se inicia con una identificación de las etiquetas semánticas en cada una de las sentencias del corpus de entrenamiento. Esta identificación proporciona los constituyentes semánticos (entidades de los diagnósticos) de las oraciones. Las palabras que se encuentren entre cada entidad se convierten en sentencias de los patrones específicos. Se les denomina patrones específicos debido a que sólo cubren la sentencia que representa.

En esta fase se obtienen patrones para relacionar las siguientes entidades: síntoma-anatomía, paciente-síntoma, síntoma-evolución, paciente-antecedente y antecedente-familiar. En la Figura 1 se muestra un ejemplo de un patrón relacional específico para una sentencia del corpus entre las entidades *síntoma* y *anatomía*.

Sentencia: *el paciente masculino de 45 años presenta edema en la pierna y temperatura local.*

Patrón relacional: <*síntoma:edem*> en <*parte_anatómica: piern*>

Figura 1. Patrón relacional específico para una sentencia del corpus.

Los patrones específicos obtenidos del corpus se generalizan, es decir, se obtiene una lista reducida en la cual se elimina la duplicidad de los mismos. Además se agrupan por el tipo de elemento que relacionan y se generalizan considerando la sinonimia de las entidades de los diagnósticos.

La generalización por sinonimia de las entidades consiste en hacer grupos de patrones a partir de la similitud de las entidades diagnósticas asociadas a las sentencias. A continuación se muestra un ejemplo de una similitud por sinonimia entre dos patrones específicos (1 y 2), los cuales pueden generalizarse en el mismo grupo (patrón generalizado).

Patrón relacional específico 1: <*síntoma:prurit*> en <*parte_anatómica: región ocul*>
Patrón relacional específico 2: <*síntoma:escoz*> en <*parte_anatómica: región ocul*>
Patrón relacional generalizado: <*síntoma*> en <*parte_anatómica*>

En este enfoque, el proceso de generalización por similitud se apoya de un diccionario de sinónimos de las entidades creado por expertos médicos.

El conjunto de patrones relacionales obtenidos puede resultar muy extenso aun cuando se ha realizado una generalización. Como propuesta para reducir los patrones relacionales se propone un proceso de filtrado.

5.4. Filtrado de patrones relacionales

Algunos patrones del conjunto generalizado pueden resultar poco útiles para la tarea de extracción de información de diagnósticos médicos. Por ello, se propone un proceso de filtrado de patrones basado en su frecuencia cuyo objetivo es eliminar los patrones irrelevantes.

El filtrado por frecuencia consiste en eliminar patrones generales que tienen poca o nula aplicabilidad en todo el corpus. El objetivo es obtener únicamente los patrones mínimamente útiles sin sacrificar la cobertura de la tarea de extracción.

Después de realizar el filtrado por frecuencia se lleva a cabo una validación por relevancia, con la diferencia de que este proceso es realizado por expertos médicos que determinan la relevancia y confiabilidad de cada patrón. Este proceso es apoyado por la precisión que tienen los patrones para extraer la información. La precisión de los patrones se mide en términos de los ejemplos relevantes que cubre cada patrón.

Los cinco patrones más frecuentes obtenidos para cada tipo de entidad relacionada se evalúan en términos de su precisión, los cuales se presentan en la siguiente sección.

Además, se presenta la frecuencia de aparición de cada patrón en el corpus de entrenamiento.

6. Evaluación de los patrones

La obtención de los patrones relacionales finales se realiza mediante la generalización, la validación por los expertos y el filtrado por frecuencia.

El filtrado por frecuencia ordena los patrones de mayor a menor representado el nivel de importancia en el corpus. El hecho de que un patrón relacional aparezca varias veces no significa que en cualquier dominio sea así, en este artículo, las cifras de frecuencia de aparición están altamente relacionadas al corpus de notas médicas.

La evaluación de los patrones más frecuentes se presente en términos precisión, es decir, la cantidad de relaciones correctamente extraídas por el patrón contra el total de relaciones extraídas.

En la Tabla 2 se muestra la frecuencia de aparición y la precisión de cada patrón relacional para los cinco más frecuentes en cada tipo de entidad.

Tabla 2. Frecuencia y precisión de los cinco patrones relacionales más frecuentes para cada entidad relacionada.

Entidades relacionadas	Patrón relacional	Frec.	Prec.
síntoma-anatomía	*<síntoma>* en *<parte_anatómica>*	1708	98.3
	<síntoma> *<parte_anatómica>*	486	98.3
	<síntoma> *de* *<parte_anatómica>*	387	88.3
	<síntoma> a nivel de *<parte_anatómica>*	101	94.0
	<síntoma> sobre *<parte_anatómica>*	27	96.2
paciente-síntoma	*<paciente>* refiere * *<síntoma>*	1987	95.6
	<paciente> presenta * *<síntoma>*	390	97.4
	<paciente> con *<síntoma>*	313	92.9
	<paciente> acude por *<síntoma>*	202	98.0
	<paciente> con presencia de *<síntoma>*	92	93.4
síntoma-evolución	*<síntoma>* de *<cantidad>* *<unidad_tiempo>* de evolución	727	100
	<síntoma> desde hace *<cantidad>* *<unidad_tiempo>*	646	99.0
	<síntoma> por *<cantidad>* *<unidad_tiempo>*	65	86.1
	<síntoma> inicio hace *<cantidad>* *<unidad_tiempo>*	49	91.8
	<cantidad> *<unidad_tiempo>* con *<síntoma>*	42	90.4
paciente-antecedente	*<paciente>* con antecedente de *<enfermedad>*	312	97.1
	<paciente> con *<enfermedad>*	172	94.1
	<paciente> presenta antecedente de *<enfermedad>*	79	100
	<paciente> tiene antecedente de *<enfermedad>*	45	100
	<paciente> sufrió de *<enfermedad>*	22	95.4
antecedente-familiar	*<familiar>* padece *<enfermedad>*	118	97.4
	<familiar> con *<enfermedad>*	95	89.4
	<familiar> positivo a *<enfermedad>*	35	91.4
	<familiar> finada por *<enfermedad>*	28	100
	antecedentes de *<enfermedad>* en *<familiar>*	17	100

7. Conclusiones y trabajo futuro

Este artículo ha propuesto un enfoque de aprendizaje automático para la creación de patrones relacionales de extracción de información con la finalidad de apoyar la toma de decisiones diagnósticas en la medicina. Con este enfoque se pretende minimizar la intervención del experto en la creación de estos patrones. Los patrones son obtenidos, generalizados y filtrados por el enfoque para que, finalmente, el experto realice una validación y evaluación.

El enfoque presentado consiste en una fase de preparación del corpus (obtención de las raíces de las palabras), un etiquetado semántico de las entidades de los diagnósticos médicos, una obtención automática de patrones y el filtrado semiautomático.

Como resultado del enfoque se obtiene un conjunto de patrones para cada tipo de las siguientes entidades relacionadas que intervienen para la obtención de diagnósticos probables: síntoma-anatomía, síntoma-anatomía, síntoma-evolución, paciente-antecedente y antecedente-familiar. Para el conjunto de los cinco patrones más frecuentes para cada entidad relacionada se presenta un estudio de relevancia por frecuencia y por precisión.

Debido a la generalización de patrones, su obtención del corpus médico y su inminente dependencia con el dominio, la precisión promedio de la información que extraen es de 95.3 para los patrones más frecuentes.

Los patrones obtenidos tienen como propósito final la extracción de información del área de medicina, la cual es útil para la obtención automática de diagnósticos probables. Esto apoya la generación del diagnóstico definitivo, actividad que es responsabilidad del experto médico.

El enfoque propuesto se ha probado con un corpus de medicina general, sin embargo, éste puede ser aplicado para cualquier conjunto de datos de entrenamiento etiquetado con entidades nombradas.

Referencias

1. La Secretaría de Salud de México: La Norma Oficial Mexicana NOM-024-SSA3-2010. Diario Oficial de la Federación, Distrito Federal, México (2010).
2. Mitchell, T.: *Machine Learning*, McGraw Hill, ISBN 0-07-042807-7, New York (1997).
3. Organización Mundial de la Salud: Clasificación Internacional de Enfermedades, décima versión CIE-10. Informe de la Conferencia Internacional para la Décima Revisión de la Clasificación Internacional de Enfermedades (1998).
4. Martínez, C. M.: Errores médicos en la práctica clínica, del paradigma biologicista al paradigma médico social. Revista Cubana Salud Pública 32(1). Ciudad de La Habana (2006).
5. Díaz, J., Gallego, B., León, A.: El diagnóstico médico: bases y procedimientos. Revista Cubana de Medicina General Integral. Revista Cubana Medicina General Integral 22(1). Ciudad de La Habana (2006).
6. Alexopoulos, E., Dounias, G. D., Vemmos, K.: Medical Diagnosis of Stroke using Inductive Machine Learning. Proceedings of Workshop on Machine Learning in Medical Applications, Advance Course on Artificial Intelligence ACAI99, pp. 91--101. Chania, Greece (1999).
7. Shashikant, U. G., Ghatol A. A.: Heart Disease Diagnosis Using Machine Learning Algorithm. Proceedings of the International Conference on Information Systems Design and Intelligent Applications. Advances in Intelligent and Soft Computing vol. 132, pp. 217--225. India (2012).
8. Drastal, G. A., Kulikowski, C. A.: Knowledge-Based Acquisition of Rules for Medical Diagnosis. Journal of Medical Systems 6(5), pp. 433--445. Springer, Netherlands (1982).

9. Jianwei, X., Ke, X.: A novel extracting medical diagnosis rules based on rough sets. The 2009 IEEE International Conference on Granular Computing, pp. 608--611. IEEE Press, Nanchang, China (2009).
10. Jiménez, C.: Razonamiento Aproximado y Adaptable en el Procesamiento de Consultas Vagas. Tesis Doctoral, Universidad Nacional de Colombia. Medellín (2008).
11. Catalá, N., Castell, N.: Construcción automática de diccionarios de patrones de extracción de información. Procesamiento del Lenguaje Natural No. 21, pp. 123--136. España (1997).
12. Califf, M. E., Money, R. J.: Relational Learning of Pattern-Match Rules for Information Extraction. Proceedings of the Sixteenth National Conference on Artificial Intelligence, pp. 328--334. Orlando, Florida (1999).
13. Califf, M. E., Money, R. J.: Bottom-up relational learning of pattern matching rules for information extraction. Journal of Machine Learning Research 39(2-3), pp. 177--210. Association for Computing Machinery (2003).
14. Soderland, S.: Learning Information Extraction Rules for Semi-Structured and Free Text. Machine Learning - Special issue on natural language learning 34 (1-3), pp. 233--272. The Netherlands (1999).
15. Huffman, S.: Learning Information Extraction Patterns from Examples. Proceeding of Connectionist, Statistical, and Symbolic Approaches to Learning for Natural Language Processing, pp. 246--260. Springer-Verlag, London (1996).
16. Riloff, E., Wiebe, J.: Learning extraction patterns for subjective expressions. Proceedings of the 2003 conference on Empirical methods in natural language processing, pp. 105--112. Stroudsburg, USA (2003).
17. Lovins, J. B.: Development of a Stemming Algorithm. Mechanical Translation and Computational Linguistics vol. 11, pp. 22--31. (1977).
18. Porter, M. F.: An Algorithm for Suffix Stripping, Readings in information retrieval, pp. 130--137. Morgan Kaufmann, San Francisco (1997).
19. Solorio, T.: Taking Advantage of Existing Named Entity Taggers by Machine Learning. Tesis de doctorado, National Institute of Astrophysics, Optics and Electronics. Puebla, México (2005).