

FACULTAD DE INGENIERÍA
UNIDAD DE POSGRADOS
PRESENTACIÓN PROPUESTA
TESIS DE MAESTRÍA

Andrés Romero Rodríguez, Código 299659
Maestría en Ingeniería de Sistemas y Computación

2 de octubre de 2006

1. Director Propuesto

Profesor Luis Fernando Niño Vásquez, Ph.D.
Departamento de Ingeniería de Sistemas e Industrial

Asesores: Profesor Gerardo Quintana
Departamento de Medicina

2. Título

Modelo de Obtención de Conocimiento Basado en Inmunología
Área: Sistemas Inmunes Artificiales
Línea de Investigación: Sistemas Inteligentes

3. Antecedentes y Justificación

Actualmente, el ritmo con el que se genera nueva información crece considerablemente a diario, de este modo, tanto las empresas como las personas, tienden a manejar volúmenes de datos enormes, ya sea en forma de documentos de trabajo, información corporativa, correo electrónico, etc. Es por esto que se hace necesario el contar con herramientas que permitan manejar esta cantidad de información de forma confiable, segura y eficiente de modo que este gran volumen se convierta en una ayuda importante en la cotidianidad de las personas en lugar de ser un dolor de cabeza por su difícil manejo.

Las herramientas con las que se cuenta actualmente, no solucionan del todo el problema de la organización de la información, debido a que simplemente proveen un medio para su almacenamiento ya sea en forma de bases de datos, sistemas de gestión bibliográficos o manejadores de correo electrónico; pero aunque el acceso a dicha información se ha hecho muy eficiente, aún no se cuenta con técnicas adecuadas que permitan aprovechar el conocimiento tácito contenido en dicha información almacenada para mejorar el desempeño de los usuarios en las tareas cotidianas que involucran la utilización de esta información. Es por esto que el manejo de estos volúmenes de datos se ha vuelto un problema serio no solo a nivel corporativo sino a nivel personal, puesto que las personas cada día tienen que lidiar más con esta información que se hace inmanejable, desde el mismo uso del correo electrónico hasta el manejo de información en su empresa.

3.1. Información Semántica

3.1.1. Representación del Conocimiento

La información semántica se refiere al conocimiento tácito que está contenido en los documentos, dicho conocimiento generalmente no está evidente dentro del documento y no es fácilmente accesible; se deben buscar técnicas que permitan encontrar y representar el conocimiento almacenado en los documentos, en [4] se definen 5 principios que debe cumplir una representación del conocimiento, estos son:

1. Una representación del conocimiento es un sustituto: esto significa que al interior de la entidad que almacena el conocimiento se presenta un proceso de razonamiento en el cual, este conocimiento almacenado es un sustituto de los objetos reales del mundo.
2. Una representación del conocimiento es un conjunto de acuerdos ontológicos: todas las representaciones que se puedan tener del conocimiento poseen cierto grado de error, el cual puede llevar a efectos no deseados en el proceso de manipulación del conocimiento. Es por esto que se debe llegar a un acuerdo acerca de las representaciones de modo que se trate de evitar ese error.
3. Una representación del conocimiento es una teoría fragmentada de razonamiento inteligente: esto se basa en la idea de que el conocimiento es la base para el razonamiento inteligente en los humanos. Al tener una representación de dicho conocimiento, esta no será del todo completa debido a las limitaciones inherentes de los sistemas de cómputo; por esto no se puede esperar un razonamiento completamente inteligente, sino que se debe fragmentar.
4. Una representación del conocimiento es un medio para computación eficiente: estas representaciones típicamente ofrecen un conjunto de ideas de modo que se facilite el proceso de inferencia, la efectividad de este proceso

depende completamente de la eficiencia con que pueda ser procesado el conocimiento almacenado.

5. Una representación del conocimiento es un medio de expresión humana: es el medio para expresarse acerca del mundo; el medio para expresar a las máquinas conceptos acerca del mundo.

Dadas las características que debe cumplir una representación del conocimiento, se debe obtener dicho conocimiento a partir de un conjunto de documentos que contengan información relacionada acerca de un dominio particular, para esto, se deben usar técnicas de agrupamiento y clasificación de los documentos, así como metodologías que permitan representar y procesar el conocimiento almacenado en tales documentos, esta información se representará típicamente mediante conjuntos de palabras clave, redes semánticas y ontologías; adicionalmente es útil contar con información previa en el contexto de interés posiblemente utilizando ontologías de dominio público como en [7].

En [3] se establecen los dos elementos principales en la adquisición de un lenguaje que sea de utilidad en la administración del conocimiento que se espera almacenar, estos son:

1. Construcción de un léxico, que es el conocimiento acerca de las palabras.
2. Generar relaciones entre las palabras del léxico mediante una red semántica o una ontología.

El conocimiento extraído de los documentos debe ser representado adecuadamente, de modo que se cumplan las características descritas anteriormente. La representación que tradicionalmente se ha utilizado para dicho conocimiento se basa en palabras clave, redes semánticas y ontologías.

3.1.2. Redes Semánticas

Las redes semánticas surgen como un medio para almacenar conocimiento semántico, en un enfoque inicial, este conocimiento se basaba en jerarquías de herencia, posteriormente, se extendió el concepto de modo que se cubriera cualquier tipo de representación gráfica [9]. Estas redes se basan en la conexión de nodos que están relacionados mediante alguna característica semántica. Los componentes de una red, se pueden dividir en 3 categorías como en [1]:

- Objetos
- Eventos
- Relaciones

En donde los objetos representan todas las entidades tanto físicas como abstractas (ideas, números, etc), para cada objeto, se define su propiedad principal como el tipo al cual pertenece. Los eventos representan acciones, además se definen predicados que modifican la descripción de un evento adicionando un rol.

Finalmente, las relaciones conectan objetos de diversos tipos, así como eventos que involucran dichos objetos.

Dichas redes son de gran utilidad puesto que proporcionan información acerca de los conceptos y las relaciones que existen entre ellos, estas relaciones pueden ser de diversos tipos, por ejemplo contención, herencia, etc. la característica común es que siempre representan una conexión semántica entre diversos conceptos que son de interés.

3.1.3. Ontologías

Una ontología se define como *una especificación formal y explícita de una conceptualización compartida* [2]; lo cual significa que esta representa un entendimiento común de algún dominio de modo que pueda ser comunicado tanto entre personas, como entre computadores. Las características que una ontología cumple, son las siguientes:

- Es un modelo abstracto de algún fenómeno del mundo, en el cual se identifican los conceptos relevantes.
- El tipo de conceptos usado, así como las restricciones que aplican sobre ellos, están definidos explícitamente.
- La ontología debe poder ser interpretada por un computador.
- La ontología captura conocimiento consensual, es decir, el resultado de un esfuerzo colaborativo de expertos en el dominio.

Las ontologías pueden ser expresadas utilizando diversos niveles de formalidad, sin embargo, existen cuatro categorías que son las formas más comunes de expresarlas [6]:

1. Altamente informal: escritas usando lenguaje natural
2. Semi-informal: Restringidas y estructuradas utilizando lenguaje natural
3. Semi-formal: Utilizando un lenguaje definido formalmente
4. Rigurosamente formal: semi-formal, incluyendo además teoremas y pruebas

Se debe encontrar el nivel de formalidad adecuado al momento de crear una ontología, puesto que el conocimiento representado en ella debe ser útil para los usuarios de la ontología y, además, esta debe ser usada para soportar el diseño y desarrollo de aplicaciones que hagan uso de este conocimiento, la mayoría de los cuales son sistemas basados en agentes para extracción de información en la web; dichos agentes deben ser capaces de compartir una ontología común para el acceso a la información.

Las ontologías pueden ser clasificadas de acuerdo al nivel de formalidad en el que están escritas [6]:

1. Catálogo: una lista simple de términos, sin definiciones ni axiomas
2. Catálogo con definiciones: un catálogo con definiciones en lenguaje natural
3. Taxonomía: una colección de conceptos con un orden parcial
4. Taxonomía Axiomatizada: una taxonomía con axiomas
5. Biblioteca contextual: un conjunto de taxonomías axiomatizadas con relaciones entre ellas.

3.2. Sistemas Inmunes Artificiales

El sistema inmune está compuesto de una gran variedad de moléculas, células y órganos distribuidos en el cuerpo. No existe un órgano central que controle el funcionamiento del sistema inmune. Su principal tarea es la de monitorear el cuerpo en busca de células con mal funcionamiento, que pueden ser propias del organismo o elementos extraños causantes de enfermedades; el sistema inmune es entonces capaz de distinguir entre lo que es propio y no propio del organismo [5].

Un conjunto de proteínas llamadas anticuerpos reconocen a los agentes externos, posteriormente, células llamadas fagocitos digieren dichos agentes. Si todos estos agentes permanecen en el mismo lugar, todo lo que tienen que hacer los fagocitos es acercarse al elemento que ha sido identificada por los anticuerpos y destruirlo [12]. El sistema inmune ha desarrollado la capacidad de generar un repertorio de detectores para los cuales, dada la exposición a un antígeno, se seleccionarán aquellos específicos que sean capaces de detectarlo adecuadamente. Este sistema presenta una capacidad casi ilimitada para detectar esencialmente cualquier agente químico, ya sea natural o artificial [11].

Los antígenos usualmente son proteínas o moléculas ajenas al organismo derivadas de patógenos o células malignas; estos antígenos consisten de componentes que los determinan llamados epítopes. En la definición de un antígeno, se deben tener en cuenta dos propiedades principales: *antigenicidad*, que es la capacidad del antígeno de ser reconocido por receptores específicos expresados en las células B y T; e *immunogenicidad*, que es la habilidad del antígeno de inducir una respuesta inmune [10].

Un sistema inmune artificial puede ser definido como un sistema computacional desarrollado utilizando ideas, teorías y componentes extraídos del sistema inmune natural.

El sistema inmune natural trabaja en tres niveles distintos: barreras físicas (como la piel), el sistema inmune innato y el sistema inmune adaptativo. Los sistemas inmunes artificiales están basados en el último nivel, el cual presenta las propiedades deseables para un sistema de inteligencia computacional tales como aprendizaje y memoria [13].

La función principal del sistema inmune de proteger el cuerpo puede ser vista, de manera general, como la clasificación de agentes en el cuerpo en dos clases: *propios*, que pertenecen al organismo; y *extraños*, que no pertenecen

al cuerpo y son potencialmente peligrosos. Esta clasificación se lleva a cabo utilizando un gran número de células T capaces de reconocer proteínas, las cuales son producidas por su propio *algoritmo de aprendizaje* [14]. Este algoritmo de aprendizaje inherente al sistema inmune ha sido utilizado con éxito en tareas de clasificación de texto y de extracción de información semántica con miras a dicho proceso de clasificación [8, 14].

4. Planteamiento del Problema

El problema que surge es el de la administración de la información tanto personal como corporativa, de modo que los usuarios no tengan que preocuparse por el almacenamiento, procesamiento y búsqueda de la información relevante en un momento determinado. Esta búsqueda de información relevante está dirigida por los intereses particulares de los usuarios que pueden depender de su rol al interior de la organización.

Para enfrentar este problema se pueden utilizar técnicas reconocidas en el campo de la inteligencia computacional tales como extracción de información, redes semánticas y demás que faciliten la representación y procesamiento del conocimiento tácito almacenado en los documentos, utilizando dichas técnicas, se deben encontrar mecanismos que permitan administrar este conocimiento de modo que sea útil al interior de las organizaciones; los pasos necesarios en este proceso son:

1. Extracción del conocimiento a partir de diversas fuentes
2. Representación y Almacenamiento del conocimiento
3. Transformación del conocimiento en otras formas que sean útiles a los usuarios
4. Distribución del conocimiento obtenido y generado a las personas interesadas en él

5. Objetivo General y Objetivos Específicos

5.1. Objetivo General

Formular un modelo de adquisición de conocimiento basado en sistemas inmunes artificiales que permita, a partir de un conjunto de documentos relacionados, extraer los conceptos relevantes y establecer las relaciones entre ellos.

5.2. Objetivos Específicos

1. Plantear un modelo de red inmune para la extracción de conceptos a partir de documentos de texto

2. Modelar las relaciones entre conceptos adquiridos mediante el modelo de red inmune desarrollado
3. Establecer un modelo para la representación del conocimiento adquirido por la red inmune, esta representación se puede basar en ontologías o redes semánticas.

6. Metodología

El desarrollo de este proyecto se hará en 4 fases bien definidas; inicialmente se llevará a cabo una revisión bibliográfica de los temas pertinentes para el modelo que se desea plantear, estos temas parten desde las nociones generales de inmunología, pasando luego al campo de los sistemas inmunes artificiales y posteriormente se basará en la búsqueda de información pertinente al área de extracción de información semántica y técnicas de representación de conocimiento. Esta revisión bibliográfica se basará en artículos disponibles gratuitamente en internet o en las bases de datos a las cuales se tiene acceso en la universidad, enfocándose principalmente en el área biológica y de ciencias de la computación. En este punto se definirá si alguno de los modelos existentes de sistema inmune artificial es adecuado para la tarea de adquisición de conocimiento que se quiere lograr, o si por el contrario se debe definir un modelo nuevo que contemple otros componentes del sistema inmune que no han sido modelados aun.

A partir de la información obtenida en la primera fase se procederá al planteamiento e implementación del modelo de sistema inmune para la tarea de adquisición de conocimiento, esta implementación ira acompañada de experimentos con conjuntos de datos conocidos en el área de extracción de información semántica y clasificación de texto. Una vez se hayan realizado experimentos con estos conjuntos de datos, se procederá a la experimentación en un dominio particular para el cual se realizará un análisis de los resultados obtenidos con la ayuda de un experto en el área que ayudará a determinar si dichos resultados son adecuados.

7. Actividades a Desarrollar

1. Fase de revisión bibliográfica
 - a) Recopilación de información bibliográfica sobre inmunología
 - b) Recopilación de información bibliográfica sobre sistemas y redes inmunes artificiales
 - c) Recopilación de información sobre técnicas de extracción de información semántica en documentos de texto
 - d) Identificación de modelos de extracción y representación de conocimiento
2. Fase de planteamiento del modelo

- a) Planteamiento del modelo de extracción de conocimiento basado en sistemas inmunes
 - b) Implementación del modelo
 - c) Realización de experimentos preliminares con conjuntos de datos conocidos
 - d) Ajustes al modelo de acuerdo al comportamiento obtenido
- 3. Fase de experimentación en un dominio dado
 - a) Identificación de un dominio adecuado para realizar experimentación con el modelo desarrollado
 - b) Adquisición de documentos del dominio establecido para la experimentación
 - c) Realización de experimentos con los datos obtenidos
- 4. Fase de análisis semántico
 - a) Definición de la representación del conocimiento obtenido a partir de la experimentación
 - b) Búsqueda de representaciones similares usadas en el dominio dado (ontologías, redes semánticas, tesauros)
 - c) Comparación de la representación generada con los modelos encontrados
 - d) Análisis del conocimiento representado con ayuda de un experto en el área
- 5. Terminación del documento de tesis

8. Cronograma

9. Bibliografía Preliminar

10. Recursos Físicos

11. Costos del Trabajo y Fuentes de Financiación

12. Comentario con Visto Bueno del Director

Referencias

- [1] J. F. Allen and A. M. Frisch. What's in a semantic network? In *ACL Proceedings, 20th Annual Meeting*, pages 19–27, 1982.

- [2] V. R. Benjamins, D. Fensel, and A. Gómez-Pérez. Knowledge management through ontologies. In *PAKM*, 1998.
- [3] S. Chakrabarti. *Mining the Web. Discovering Knowledge from Hypertext Data*. Morgan Kaufmann Publishers, 2002.
- [4] R. Davis, H. Shrobe, and P. Szolovits. What is a knowledge representation. *AI Magazine*, 14(1):17–33, 1993.
- [5] L.Ñ. de Castro and J. Timmis. Artificial Immune Systems: A Novel Approach to Pattern Recognition. In L. A. J. Corchado and C. Fyfe, editors, *Artificial Neural Networks in Pattern Recognition*, pages 67–84. University of Paisley, Jan. 2002.
- [6] D. Elliman. Automatic derivation of on-line document ontologies, July 26 2001.
- [7] E. Gabrilovich and S. Markovitch. Feature generation for text categorization using world knowledge. In *IJCAI*, pages 1048–1053, 2005.
- [8] J. Greensmith and S. Cayzer. An artificial immune system approach to semantic document classification. Technical Report HPL-2003-141, Hewlett Packard Laboratories, July 16 2003.
- [9] T. L. Griths and M. Steyvers. A probabilistic approach to semantic representation, Apr. 29 2002.
- [10] R. Holtappels. Dominating immune response, immunodominance and its significance in immunity. *B.I.F. FUTURA*, 20(3), 2005.
- [11] D. Izhaky and I. Pecht. What else can the immune system recognize? In *Proceedings of the National Academy of Sciences of the United States of America*, volume 95, pages 11509–11510, September 1998.
- [12] S. I. Nishimura. A study of spatial formation of immune cells. *Genome Informatics*, 12:302–303, 2001.
- [13] A. Scime. *Web Mining: applications and techniques*. Idea Group, 2005.
- [14] J. Twycross. An immune system approach to document classification. Technical Report HPL-2002-288, Hewlett Packard Laboratories, Oct. 23 2002.