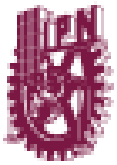
INSTITUTO POLITÉCNICO NACIONAL



CENTRO DE INVESTIGACIÓN EN COMPUTACIÓN
Laboratorio de Procesamiento de Lenguaje Natural

RESPUESTA AUTOMÁTICA A PREGUNTAS SOBRE DOCUMENTOS LEGALES EN ESPAÑOL

TESIS

QUE PARA OBTENER EL GRADO DE
MAESTRO EN CIENCIAS DE LA COMPUTACIÓN

PRESENTA
ING. ALFREDO LÓPEZ MONROY

DIRECTORES DE TESIS:

DR. ALEXANDER GELBUKH

DR. FRANCISCO HIRAM CALVO CASTRO

México, D.F. Octubre 2008

SIP-14

INSTITUTO POLITECNICO NACIONAL

SECRETARIA DE INVESTIGACIÓN Y POSGRADO

ACTA DE REVISIÓN DE TESIS

En la Ciudad de México, D.F. siendo las 11:00 horas del día 13 del mes de Junio de 2008 se reunieron los miembros de la Comisión Revisora de Tesis designada por el Colegio de Profesores de Estudios de Posgrado e Investigación del:

Centro de Investigación en Computación

para examinar la tesis de grado titulada:

"RESPUESTA AUTOMÁTICA A PREGUNTAS SOBRE DOCUMENTOS LEGALES EN ESPAÑOL"

LÓPEZ	**MONROY**	**ALFREDO**
Apellido paterno	materno	nombre(s)

Con registro:

B	0	6	1	0	5	0

aspirante al grado de: ***MAESTRO EN CIENCIAS DE LA COMPUTACIÓN***

Después de intercambiar opiniones los miembros de la Comisión manifestaron ***SU APROBACIÓN DE LA TESIS***, en virtud de que satisface los requisitos señalados por las disposiciones reglamentarias vigentes.

LA COMISIÓN REVISORA

Presidente
Dr. José Luis Oropeza Rodríguez

Secretario
Dr. Salvador Godoy Calderón

Primer vocal
(Director de Tesis)
Dr. Alexander Gelbukh Kahn

Segundo vocal
(Director de Tesis)
Dr. Francisco Hiram Calvo Castro

Tercer vocal
Dr. Miguel Jesús Torres Ruiz

EL PRESIDENTE DEL COLEGIO
Dr. Jaime Álvarez Gallegos

INSTITUTO POLITECNICO NACIONAL
CENTRO DE INVESTIGACION
EN COMPUTACION
DIRECCION

INSTITUTO POLITECNICO NACIONAL

SECRETARÍA DE INVESTIGACIÓN Y POSGRADO

CARTA CESION DE DERECHOS

En la Ciudad de México el día 12 del mes Octubre del año 2008, el (la) que suscribe Alfredo López Monroy alumno (a) del Programa de Maestría en Ciencias de la Comp con número de registro B061050, adscrito a Centro de Investigación en Computación, manifiesta que es autor (a) intelectual del presente trabajo de Tesis bajo la dirección de Alexander Gelbukh e Hiram Calvo y cede los derechos del trabajo intitulado Respuesta automática a preguntas sobre documentos legales, al Instituto Politécnico Nacional para su difusión, con fines académicos y de investigación.

Los usuarios de la información no deben reproducir el contenido textual, gráficas o datos del trabajo sin el permiso expreso del autor y/o director del trabajo. Este puede ser obtenido escribiendo a la siguiente dirección alopezm301@ipn.mx. Si el permiso se otorga, el usuario deberá dar el agradecimiento correspondiente y citar la fuente del mismo.

Alfredo López Monroy

Nombre y firma

Resumen

En un sistema de búsqueda de respuestas (*SBR*) el usuario introduce preguntas en lenguaje natural y obtiene respuestas que pueden ser ya sea, construidas por el sistema a partir de un conjunto de documentos, o bien simplemente extractos de texto que contienen la respuesta. Dependiendo de los temas a los que se refiere la colección de documentos, los *SBR* se dividen en *SBR* de dominio abierto (incluyen documentación médica, legal, bibliográfica, periódicos, etc.) y *SBR* de dominio restringido (se refieren a un solo tópico, p. ej. diagnósticos médicos). En el presente trabajo se describe una arquitectura para un sistema de búsqueda de respuestas restringido a documentos legales en lenguaje español, específicamente a textos normativos del Instituto Politécnico Nacional (IPN) y enfocado a preguntas realizadas al Abogado General del IPN. La arquitectura propuesta se basa en la representación de la colección de documentos en forma de un grafo ponderado en el que cada nodo representa los artículos de los textos normativos y las aristas el valor inverso de la medida de semejanza coseno. Dada una pregunta, el sistema la divide en dos partes (Q1 y Q2) que posteriormente se agregan al grafo como dos nuevos nodos. Mediante el algoritmo de Dijkstra se encuentran las rutas de peso mínimo de Q1 a Q2. Las rutas así encontradas contienen artículos altamente relacionados entre ellos y con las partes que integran la pregunta de tal forma que, en la mayoría de los casos, a partir del conjunto de artículos devueltos por el sistema, es posible inferir la respuesta adecuada. Adicionalmente, se emplearon técnicas de procesamiento de lenguaje natural con la finalidad de evaluar su impacto en el sistema, para lo cual se propuso un esquema de evaluación y bajo el mismo se compararon las respuestas obtenidas con el *SBR* propuesto y un sistema basado en el modelo tradicional de espacio vectorial para la misma tarea. Bajo los criterios de evaluación empleados, el sistema propuesto muestra notablemente un mejor desempeño con respecto al sistema basado en el modelo de espacio vectorial. Los resultados obtenidos muestran que el enfoque adoptado es adecuado, sin embargo aún existe posibilidad de mejora para la tarea de búsqueda de respuestas.

Abstract

In a QA system, the users poses natural language questions and gets answers that can be either built by the system from a document collection or simply text fragments that contain the answer. These systems are classified with regard to the collection of documents which they access. If the collection is about several subjects, the system is called Open Domain QAS; whereas for a only subject, it is called Restricted Domain QAS. This work describes a Question Answering System (QAS) architecture restricted to legal documents in Spanish language, focused to National Polytechnic Institute regulation texts. The proposed architecture is based on a weighted graph in which each node represents an article from the documents and the value associated to the edges correspond to degree of similarity given by the standard cosine similarity measure. Given a question, the system splits it in two parts. Each of them is added as part of the graph as two new nodes. Then, several paths are constructed from part A of the question to part B, so that the shortest path contains the relevant articles to the question. Additionally, the system employs natural language processing techniques to verify their impact on the system. We propose several evaluation criteria and compare the results obtained for our system with the answers given by a traditional information retrieval system adjusted for article retrieval instead of document retrieval. The results show that our system performs at least twice as better with regard to the traditional Information Retrieval model for Question Answering.

Contenido

Índice de figuras

Índice de tablas

Abreviaciones

MEV: Modelo de espacio vectorial

LLA: Lista lematizada de artículos

LBA: Lista básica de artículos

LBA_R: Lista básica de artículos. Removiendo palabras sin contenido

LLA_R: Lista lematizada de artículos. Removiendo palabras sin contenido

GBP: Grafo construido a partir de la lista básica de artículos

GLE: Grafo construido a partir de la lista lematizada de artículos

GBP_R: Grafo construido a partir de la lista básica de artículos. Removiendo palabras sin contenido

GLE_R: Grafo construido a partir de la lista lematizada de artículos. Removiendo palabras sin contenido

***SBR*:** Sistema de búsqueda de respuestas

Capítulo I

1 Introducción

Actualmente, el intenso uso de computadoras y de las redes que las interconectan, aunado a las nuevas tecnologías de almacenamiento, ha conducido a una cantidad sin precedente de información de todo tipo: texto, imágenes, audio y video. El acceso a tal información, se lleva a cabo a través de bases de datos, herramientas muy útiles para almacenar la información estructurada o semi-estructurada. No obstante, el mayor porcentaje de esta información se encuentra en forma de texto, el cual no posee una estructura explícita, lo cual impide acceder a tal información usando directamente las mismas técnicas que se emplean para información estructurada. Los modernos motores de búsqueda permiten a través de un conjunto de palabras clave recuperar un conjunto de documentos; sin embargo, es necesario que los usuarios realicen un esfuerzo adicional para filtrarlos y localizar la información específica que necesitan [1].

En los últimos años, se ha realizado una extensa investigación sobre mecanismos alternos que permitan acceder a la información almacenada en medios electrónicos de una forma más fácil que la proporcionada por los actuales buscadores. Parte de esta investigación se ha enfocado en los Sistemas de Búsqueda de Respuestas (*SBR*). En un *SBR* el usuario introduce preguntas en lenguaje natural y el sistema, a diferencia de los buscadores que devuelven un conjunto de documentos, devuelve respuestas. Las respuestas son encontradas a partir de la consulta a una base de documentos. La información entregada al usuario debe ser capaz de satisfacer la pregunta realizada.

1.1 Sistemas de búsqueda de respuestas

En general, los *SBR* se clasifican de acuerdo con la colección de documentos a la que tienen acceso, si la colección trata sobre diversos temas se le llama de *dominio de abierto* y si trata de un sólo tópico se le conoce como de *dominio restringido*. Entre éstos, los que más atención han recibido son los de dominio abierto gracias a su promoción en foros internacionales, los cuales se han enfocado a *SBR* capaces de dar respuesta a preguntas simples y concretas hechas por usuarios casuales. Los usuarios casuales constituyen el nivel más elemental de usuarios y se espera que los sistemas manejen el tipo de preguntas que plantean [2]. La mayoría de los *SBR* de dominio abierto son orientados exclusivamente al idioma inglés, sin embargo también se encuentran trabajos para otros idiomas, incluido el español, e incluso sistemas multilingües.

No obstante los esfuerzos realizados en los *SBR*, las evaluaciones realizadas a los *SBR* de dominio abierto a preguntas concretas aún no han alcanzado el 100% de precisión. Tal resultado muestra la complejidad del problema y la necesidad de otros enfoques, por lo que parte de la investigación sobre *SBR* se ha dirigido además al desarrollo de sistemas de dominio restringido. A este respecto, se han desarrollado *SBR*, a los que se han incorporado incluso interfaces de voz en lenguaje natural [4], desarrollados para distintas áreas como la médica [5, 6, 7], doméstica [4], de la construcción [8], legal [9, 10], entre otras. Particularmente en el área legal, en 2003 se llevó a cabo el taller *Question Answering for interrogating legal documents* dentro del foro JURIX (The foundation for Legal Knowledge Based Systems).

1.2 Definición del problema

A pesar del extenso trabajo realizado en la tarea de búsqueda de respuestas aún existen amplias áreas por investigar, en particular en lo que respecta a la búsqueda de respuestas sobre documentación normativa en español, tópico hasta ahora poco explorado; sin embargo de gran interés.

Los actuales mecanismos electrónicos de acceso a documentación legislativa son aún deficientes para dar respuesta rápida a preguntas planteadas por personas interesadas en la normatividad. Por ejemplo considérese la siguiente pregunta:

Pregunta:

¿Cómo se lleva a cabo el procedimiento de elección de representantes alumnos ante el Consejo Técnico Consultivo Escolar?

Respuesta dada por el Abogado General del Instituto Politécnico Nacional:

El proceso de elección de representantes alumnos ante los consejos técnicos consultivos de las escuelas, centros y unidades de enseñanza media superior está regulado en los ***artículos 28 fracción IV de la Ley Orgánica, 206, 207, 209 y 213 fracción I del Reglamento Interno****, dispositivos retomados en las bases de la convocatoria autorizada mediante oficio número 014 SG/634/01, inscrita en la oficina del Abogado General con el número 582 a fojas 134 del libro de Registro de Actos Jurídicos Diversos. El artículo 213 Fracción I del Reglamento Interno preceptúa que los únicos requisitos para ser candidato a consejero alumno consisten en tener situación escolar regular y estar cursando estudios entre el tercero y penúltimo semestres. El artículo 209 del Reglamento Interno establece que comités conformados por alumnos con situación escolar regular son quienes deben elegir, en forma directa y por mayoría de votos, a los representantes educandos entre el Consejo Técnico Escolar.*

Actualmente, los usuarios que tienen alguna necesidad de satisfacer sus dudas en el campo normativo pueden utilizar los motores de búsqueda de Internet para localizar los artículos que proporcionen la respuesta que requieren; no obstante, buscadores como Google no brindan una respuesta directa, ya que muchas de las páginas principales están ligadas a páginas secundarias en donde posiblemente esté vinculada parte de la información buscada. Por otra parte, existen páginas gubernamentales (Figura 1) que contienen listados por título de documentos normativos. Sin embargo, tales reglamentos solamente están relacionados con la institución a la que pertenece la página, y solamente es posible visualizar el contenido completo del documento en el mismo navegador o descargarlo en algún formato comercial (Word, Acrobat, Power Point, etc.).

Una vez encontrados los documentos más adecuados, la tarea adicional para el usuario consiste en revisar, ocasionalmente incluso, el texto completo de los diversos documentos encontrados para obtener la respuesta deseada. En la mayoría de los casos, la revi-

sión de los documentos es engorrosa debido a que hay reglamentos que contienen más de 200 artículos, por lo que el usuario se ve en la necesidad de utilizar herramientas tales como los buscadores de ocurrencia de palabras de los editores o visualizadores de texto, que en muchas ocasiones resultan de poca utilidad, principalmente porque la respuesta no se halla en un solo reglamento, y porque muchas veces la forma de realizar la búsqueda textual no coincide con la forma en la que están escritos los términos en el reglamento.

Figura 1. Acceso que actualmente brinda el IPN a su documentación normativa[1]

A este respecto, el Instituto de Investigaciones Jurídicas de la Universidad Nacional Autónoma de México (UNAM), ofrece una alternativa, en su página[2] se puede encontrar un concentrado de textos legislativos, no solo pertenecientes a la UNAM, sino también a legislación federal, estatal e incluso internacional que ofrece para su consulta, la que consiste básicamente en:

[1] http://www.abogadogeneral.ipn.mx/reglamentos.html
[2] http://info4.juridicas.unam.mx/unijus/frames/unv.htm

a) Visualización de documentos completos listados por título, o

b) Búsqueda de documentos completos o artículos, basada en el modelo booleano, con las características de incorporar búsquedas de términos sin importar mayúsculas/minúsculas, no considerar acentos, además de incorporar la búsqueda de términos con la misma raíz morfológica p. ej. *constituci*, para abarcar términos como *constitució*n, *constitucio*nes, *constitucio*nal, etc.

Se ofrece la ayuda de que los términos de búsqueda aparecen resaltados en los textos retornados como respuesta (los que pueden ser hasta 100, Figura 2) además que se puede utilizar la herramienta de búsqueda de palabras del navegador que se utilice, sin embargo aún estas herramientas son insuficientes para una búsqueda de información fácil y rápida sobre documentos normativos.

Figura 2. Pantalla de búsqueda del sistema de documentos de la UNAM

Bajo este escenario un *SBR* orientado a documentación normativa abriría el camino para brindar mecanismos de acceso a la información de una forma mucho más sencilla que la proporcionada por los medios antes descritos.

1.3 Propuesta

La propuesta del presente trabajo, consiste en el desarrollo de un *SBR* que sea capaz de proporcionar un conjunto de artículos, a partir de una colección de leyes y reglamentos, mediante los cuales sea posible inferir respuestas a preguntas introducidas en lenguaje natural, basado en representar una colección de documentos como un grafo ponderado y la extracción de respuestas a partir de éste (Figura 3); se considera además la aplicación de técnicas de procesamiento de lenguaje natural para evaluar el impacto en el sistema. Como medio de evaluación del alcance del sistema se proponen criterios basados en el conjunto de preguntas/respuestas encontrado en la página del Abogado General del Instituto Politécnico Nacional.

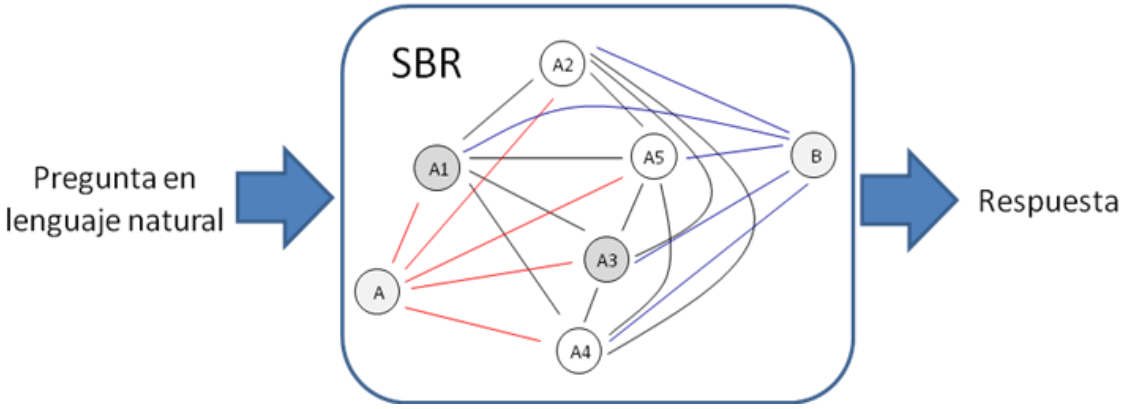


Figura 3. Sistema de búsqueda de respuestas propuesto

1.4 Objetivos

A continuación, se resumen los objetivos alcanzados en el desarrollo del presente trabajo de tesis.

1.4.1 Objetivo general

- El diseño e implementación de un Sistema de búsqueda de Respuestas enfocado al área legal, particularmente documentos normativos, enfocado al conjunto de preguntas/respuestas dadas en la página del Abogado General del Instituto Politécnico Nacional (IPN) para su evaluación, empleando el conjunto de documentos referidos a la normatividad del IPN.

1.4.2 Objetivos particulares

- Proponer una arquitectura de *SBR* adecuada para el problema que trata el presente trabajo, utilizando técnicas de procesamiento de lenguaje natural y técnicas del área de recuperación de información.
- Proponer un esquema de evaluación del sistema adecuado que permita cuantificar su precisión en la tarea y el impacto de las distintas técnicas empleadas.
- Verificar que el *SBR* propuesto se desempeña mejor que uno basado en solamente técnicas de procesamiento de recuperación de información.

1.5 Aportaciones

Esta tesis es resultado de la investigación realizada en el desarrollo de un sistema de búsqueda de respuestas restringido al dominio legal. El trabajo realizado ofrece las siguientes contribuciones a la mejora de los actuales mecanismos de acceso a la información normativa en lenguaje español:

- Desarrollo e implementación de una arquitectura para un Sistema de Búsqueda de Respuestas restringido al dominio legal en idioma español, lo que es particularmente importante debido a que los principales trabajos se han enfocado a *SBR* de dominio abierto en idioma inglés y el terreno en dominio restringido ha sido poco explorado y aún más en lo que se refiere a la documentación legal en español.
- Desarrollo de un marco de trabajo para la representación de textos normativos, misma que refleja la estructura de tales documentos así como un método para extracción de fragmentos de texto que permiten inferir respuestas a preguntas introducidas en lenguaje natural. El marco de trabajo propuesto es muy flexible de tal forma que permite subsecuentes mejoras y como se muestra en el presente trabajo, es adecuado para el problema de búsqueda de respuestas sobre documentación referida a leyes y reglamentos; además que podría extenderse a otros textos legales como pueden ser ordenanzas, códigos y estatutos.

- Enfoque de solución propuesto, considerando tanto el conjunto de preguntas/respuestas obtenido de la página del Abogado General del Instituto Politécnico Nacional (IPN), como la colección de documentos (26 textos normativos acerca de la normatividad del IPN).
- Desarrollo de un esquema de evaluación que permite verificar la precisión del sistema propuesto así como compararlo con otros enfocados al mismo problema. En este caso no se encuentran en la literatura trabajos similares, sin embargo en el presente trabajo se implementó también un sencillo sistema para la misma tarea basado en el modelo de espacio vectorial. Bajo los criterios de evaluación propuestos se encuentra que el enfoque propuesto es adecuado y tiene un mejor desempeño que el basado solamente en técnicas del área de recuperación de información.

1.6 Organización del trabajo

El resto del documento se encuentra organizado como sigue:

Para brindar al lector un panorama general de la búsqueda de respuestas, en el capítulo 2 se proporciona un breve repaso de los *SBR* desde sus orígenes hasta los enfoques actuales. En el capítulo 3 se proporciona un breve repaso de teoría de grafos, el modelo de espacio vectorial; así como métodos de preparación de texto y recursos del área de procesamiento de lenguaje natural. Si el lector conoce de estos tópicos puede omitir este capítulo e ir directamente al capítulo 4 en el que se muestra en que consiste la propuesta del presente trabajo y la forma en la que se emplean las bases teóricas del capítulo 3 en la arquitectura general del *SBR* desarrollado. Posteriormente, en el capítulo 5 se describen los experimentos llevados a cabo, los resultados obtenidos y la forma de evaluación propuesta. Finalmente, en el capítulo 6 se encuentran las conclusiones del desarrollo del trabajo; así como también el trabajo futuro que la presente investigación sugiere.

Capítulo II

2 Estado del arte

La idea de brindar mejores mecanismos de acceso a la información almacenada en medios electrónicos a través de la incorporación del lenguaje natural no es nueva, ésta se remonta a la década de los 50's. En la presente sección se proporciona un breve repaso de sistemas orientados a la tarea de la búsqueda de respuestas a través de los años y algunos de los enfoques actuales.

2.1 Breve historia de los *SBR*

La investigación y el desarrollo de sistemas capaces de responder a preguntas en lenguaje natural data de 1950, con la famosa prueba de Turing para solucionar el problema de si una máquina piensa. A. Turing propuso una prueba llamada "el juego de la imitación" en el cual, un humano se comunica con una máquina a través de un teletipo preguntándole cosas. Turing consideraba que una máquina sería inteligente si un humano no se daba cuenta si quién contestaba era una máquina u otro humano.

Sin embargo, no sólo la búsqueda de respuestas tiene una importancia teórica sino también práctica, ya que también surgió el interés en el desarrollo de interfaces en lenguaje natural que permitieran un acceso más sencillo a la información almacenada en medios electrónicos. Debido a lo anterior, se mencionan en la literatura distintos trabajos referidos a la tarea de búsqueda de respuestas a partir de preguntas realizadas en lengua-

je natural, desarrollados ya desde la década de los 50's, sin embargo los primeros sistemas conocidos aparecen hasta la década de los 60's y 70's.

Entre los trabajos más citados en la literatura se encuentran BASEBALL (1963) y LUNAR (1971). Estos primeros sistemas tuvieron como características en común que se enfocaron a la extracción de respuestas a preguntas introducidas en lenguaje natural a partir de bases de datos estructuradas de dominios restringidos. El sistema BASEBALL se enfocaba a responder preguntas acerca de los juegos llevados a cabo en una temporada de la Liga de Béisbol de los Estados Unidos de Norteamérica, por ejemplo:

¿Cuántos partidos jugaron los Yankees en julio?

Este era el tipo de preguntas que el sistema era capaz de contestar, para lo cual realizaba tanto un análisis sintáctico y semántico de las preguntas.

Por otra parte, LUNAR fue diseñado con la finalidad de responder a preguntas sobre datos acerca de análisis químicos de muestras de rocas y suelo traídas de la misión lunar Apollo 11. Un ejemplo de pregunta que el sistema LUNAR manejaba es:

¿Cuál es la concentración promedio de aluminio en las rocas que tienen gran contenido de álcali?

La evaluación del sistema se llevó a cabo en la Convención de Ciencia Lunar en 1971. En ésta, LUNAR fue capaz de responder el 90% de preguntas realizadas por geólogos lunares en activo. Cabe mencionar que a los investigadores participantes no se les proporcionaron instrucciones previas acerca del formato de preguntas a introducir.

Durante los años posteriores continuaron las investigaciones en el campo de búsqueda de respuestas, resultado de los logros obtenidos con los sistemas antes mencionados. Sin embargo, la principal limitante encontrada fue la dificultad en la construcción de bases de datos para otros dominios, lo que provocó por algún tiempo el abandono de la tarea. Existen, no obstante, algunas investigaciones realizadas sobre el problema de búsqueda de respuestas en general enfocadas a comprensión de historias (1977), sistemas de dialogo interactivos (1972) y en áreas como la extracción y recuperación de información [1].

En el año de 1993 aparece START[3] el primer sistema de búsqueda de respuestas en línea desarrollado en el Instituto Tecnológico de Massachusetts, el que aún hoy se encuentra funcionando. Este sistema se basa en una anotación en lenguaje natural de información tal como texto, imágenes, video y audio. La anotación consiste en colecciones de frases que describen el contenido de distintos fragmentos de información [22]. Es importante decir que este sistema sólo funciona para el idioma inglés.

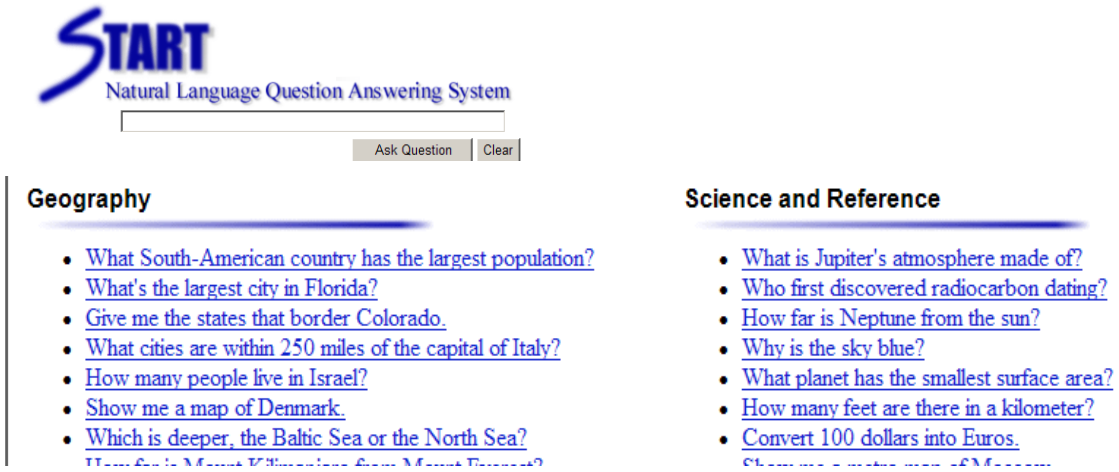


Figura 4. Pantalla principal de START, *SBR* implementado por el *MIT*

Es hasta el año de 1999 que el interés en los *SBR* resurge entre la comunidad de la Inteligencia Artificial y la de Procesamiento de Lenguaje Natural, con la incorporación del *"Question Answering Track"* en el marco de la conferencia anual *Text REtrieval Conference*, *TREC*, la que a su vez ha motivado la investigación en muchas otras direcciones.

2.2 Panorama actual de los *SBR*

Hoy en día la investigación sobre la tarea de búsqueda de respuestas es muy extensa; se ha abordado desde distintas perspectivas, utilizando las más variadas herramientas y se ha aplicado a diversas áreas. Debido a esto, actualmente existen distintas clasificaciones para los *SBR*, la más general y la que empleamos en el presente trabajo es la basada en el alcance del sistema:

SBR de dominio abierto o

SBR de dominio restringido

[3] http://start.csail.mit.edu/

Como se mencionó en el capítulo anterior, la diferencia entre ellos es básicamente el tipo de preguntas/respuestas que el sistema acepta y el tipo de documentación a la que tienen acceso. En el caso de los *SBR* de dominio abierto, éstos se han enfocado a preguntas referidas a distintos temas; se pueden encontrar preguntas sobre geografía, espectáculos, deportes, medicina, etc. Para contestar tales tipos de preguntas dichos sistemas emplean documentación muy variada: periódicos, libros, enciclopedias e incluso Internet. En lo que respecta a los *SBR* de dominio restringido, éstos se orientan más bien a una aplicación muy concreta como por ejemplo del área médica; un ejemplo más específico podría ser la búsqueda de enfermedades sobre una colección de diagnósticos médicos a partir de preguntas en lenguaje natural que incluyen solamente síntomas.

Ejemplos de *SBR* de dominio abierto son los sistemas presentados en el *Text REtrieval Conference*[4] (*TREC*) y en el *Cross Lingual Evaluation Forum*[5] (*CLEF*). Considerando los trabajos presentados en estos foros los *SBR* de dominio abierto se pueden clasificar también por monolingües o multilingües. En el caso del *TREC* solamente se presentan sistemas orientados al idioma inglés, mientras que en el *CLEF* abarcan los principales idiomas europeos (por supuesto incluido el español), en sistemas no sólo monolingües sino también capaces de manejar documentación y preguntas/respuestas en distintos idiomas.

Por otra parte la investigación sobre *SBR* de dominio restringido es considerablemente menor a la de los de dominio abierto, sin embargo existe también un gran interés en este tipo de sistemas ya que a pesar de que no existen foros similares a los del *TREC*/*CLEF* para *SBR* de dominio restringido, en los últimos años se han llevado a cabo distintos talleres en los que se han presentado trabajos orientados para distintas áreas entre las que más destacan es la médica y la legal. La mayoría de los trabajos de dominio restringido son orientados al idioma inglés pero se encuentran también trabajos en otros idiomas tales como el francés, español y portugués. Debido a que la característica principal de los *SBR* de dominio restringido es que las preguntas que manejan se pueden responder a partir de solamente cierta documentación específica no se abordan usualmente problemas multilingües en esta área de búsqueda de respuestas.

[4] http://trec.nist.gov/
[5] http://www.clef-campaign.org/

2.2.1 *SBR* de dominio abierto

Como ya se ha mencionado, actualmente los *SBR* de dominio abierto son los que mayor atención han ganado debido a los foros *Text REtrieval Conference* (*TREC*) mismo que se celebra en Estados Unidos de Norteamérica y el *Cross Lingual Evaluation Forum* (*CLEF*) el que se lleva a cabo en Europa. En la presente sección, se describe de forma general en que consisten las pruebas en estos foros, la más importante es la prueba sobre búsqueda de respuestas *"Question answering Track"* del *TREC*, a continuación se describen las características principales de esta prueba que inicia en 1999.

En el *TREC* se comenzó manejando preguntas sobre hechos concretos (llamadas en inglés *factoid questions)* referidas a una cantidad, nombre, fecha, lugar, etc., p. ej. ¿Cuántas calorías tiene una BigMac? ¿Dónde se encuentra el Taj Mahal? ¿Cuándo Nixon visitó China?, lo que se mantuvo hasta el año 2001 [30]. A partir de ese año se han ido agregando otros tipos de preguntas, sin embargo se continúo concentrándose principalmente en preguntas factuales, debido a esto se resaltan solamente los resultados alcanzados para tal tipo de preguntas.

En las ediciones del *TREC*-8 (1999) y el *TREC*-9 (2000) se solicitaba a los participantes que sus sistemas retornaran hasta 5 respuestas de 50 bytes o 250 bytes, cada respuesta contenía la respuesta de la longitud requerida y el número de identificación del documento de donde se extrajo la respuesta. La métrica de evaluación consistía en asignar de calificación a cada pregunta el valor inverso de la posición en la que se encontraba la respuesta correcta y 0 si no se encontraba dentro del conjunto de 5 respuestas [23]. Por ejemplo si la respuesta se encontraba en la posición 5 la calificación asignada era 0.2, de esta forma la evaluación global del sistema era calculada a través del promedio del conjunto de preguntas contestado correctamente, medida llamada en inglés *Mean Reciprocal Rank* (*MRR*) definida de la siguiente forma:

$$MRR = \sum_{i=1}^{q} \frac{1}{r_i} / q \qquad (1)$$

La diferencia entre el *TREC* – 8 y el *TREC* 9 fue que en este último el conjunto de preguntas era de 500 mientras que en el *TREC* – 8 solamente fueron 200. Para el caso de las preguntas sobre hechos concretos, considerando las respuestas de longitud 50 bytes, el mejor sistema fue capaz de responder aproximadamente el 70% de las preguntas en el

TREC – 8 y cerca del 65% en el *TREC* – 9, a pesar de la leve diminución en los resultados se consideró un buen resultado debido a que las preguntas empleadas en la primera edición fueron especialmente diseñadas para la prueba, mientras que las del *TREC* – 9 fueron recolectadas de usuarios reales [25]. En las primeras ediciones los participantes siguieron más o menos el mismo enfoque. El sistema primero trataba de clasificar la pregunta en alguna categoría, ¿quién? implicaba una persona o una organización y una pregunta que comenzará con ¿cuándo? implicaba que se requería una fecha como respuesta. Posteriormente, el sistema recuperaba una pequeña cantidad de documentos empleando tecnología estándar de recuperación de documentos. El sistema realizaba un procesamiento, usualmente con técnicas de procesamiento de lenguaje natural, para detectar entidades del mismo tipo que la pregunta. Si se encontraba la entidad deseada lo suficientemente cerca de las palabras que contenía la pregunta, el sistema retornaba tal entidad como respuesta, en caso de no encontrar alguna respuesta, el sistema en ocasiones podía intentar realizar otra búsqueda empleando diferentes criterios [24].

En el *TREC* – 10 (2001) se continúo con preguntas sobre hechos concretos, con la diferencia que ahora solamente se requería que la longitud de las respuestas fuera de solamente 50 bytes, además que no todas las preguntas tenían respuesta. También se agregaron preguntas que requerían respuestas que se encontraban en más de un documento (llamadas en inglés *list questions*), en esta prueba las preguntas siempre contenían el número de respuestas que requerían, por ejemplo, "Nombre 4 ciudades de EU que tengan un teatro Shubert". La métrica de evaluación para este tipo de preguntas se enfocó en la precisión, la que se definió como el número de respuestas correctas encontradas entre el número de respuestas indicado en la pregunta.

En el 2002 (*TREC* – 11) la prueba siguió con las preguntas sobre hechos concretos y respuestas que requieren más de una respuesta con el cambio principal que las respuestas dadas por los sistemas fueran precisas, no fragmentos de texto que las contuvieran, además los sistemas debían de retornar solamente una respuesta (al contrario de las 5 de las ediciones anteriores), otro cambio es que el sistema debía agregar el nivel de confianza que tenía sobre la respuesta devuelta [26], así la medida de evaluación empleada fue la *medida de confianza ponderada* (denominada en inglés *confidence weighted score, cws*), de forma más precisa:

$$cws = \frac{1}{q}\sum_{i=1}^{q}\frac{c_i}{i} \quad (2)$$

Para el 2003 (*TREC* – 12) se incorporaron preguntas consideradas de mayor dificultad que las preguntas sobre hechos concretos, como son preguntas sobre definiciones (en inglés *definiton questions*) y preguntas que requieren más de una respuesta (*list questions*) pero sin agregar el número de respuestas que se desean como en el *TREC* – 10 y 11, por ejemplo, "Enlista los nombres de gomas de mascar". Por su parte, las preguntas sobre definiciones requerían de respuesta información interesante acerca de una cosa o persona en particular, por ejemplo, ¿Quién es Vlad el Empalador?. Para la prueba del *TREC* – 12 sobre preguntas factuales se emplearon 413 preguntas, el mejor resultado alcanzado en la evaluación en este tipo de preguntas fue del 70% de precisión, definida como la fracción de respuestas evaluadas como correctas:

$$acc = \frac{1}{q}\sum_{i=1}^{q} acc_i \quad (3)$$

Después de 5 años de iniciada la prueba el enfoque general para encontrar respuestas permaneció esencialmente sin cambio. Los sistemas generalmente determinaban la posible respuesta a través de clasificar la pregunta en alguna categoría, recuperaban documentos o pasajes que podrían contener la respuesta y posteriormente, empleando los términos de la pregunta, intentaban hallar fragmentos de texto que proporcionaran la respuesta [27]. No obstante de que el enfoque general permaneció sin cambio, las técnicas individuales empleadas se refinaron, por ejemplo, en el proceso de clasificar las preguntas, en las primeras ediciones la clasificación empleaba reglas simples hechas manualmente, para el 2003 se empleaban en su lugar clasificadores automáticos como máquinas de soporte vectorial [31].

A partir del 2004 (*TREC* – 13) se agrupan las preguntas sobre hechos concretos y las preguntas de varias respuestas en dos series donde cada una se asocia con un objeto y las preguntas en las series se refieren a dicho objeto, además se agrega una pregunta final de cada serie, misma que denominaron pregunta tipo "otro" que se interpreta como "dime otras cosas interesantes acerca de este objeto, no sé lo suficiente para preguntarlo yo mismo". Este último tipo de preguntas es más o menos equivalente a las preguntas sobre definiciones del *TREC* 2003, a continuación se muestra un ejemplo:

Objeto: Cometa Hale-Bopp

Pregunta factual: ¿Cuándo fue descubierto el cometa?

Pregunta factual: ¿Cada cuándo se aproxima a la Tierra?

Pregunta de respuesta múltiple: ¿En que países fue visible en su última visita?

Pregunta Otro:

En el *TREC* – 13, para las preguntas factuales, se obtuvo el 77% de precisión en 230 preguntas [3].

En el 2004 los posibles objetos para cada serie podían ser una persona, una organización o una cosa. En el 2005, la categoría Evento se agregó como posible objeto, para lo cual se solicitó que las respuestas devueltas fueran temporalmente correctas con respecto a la época definida por las series de preguntas. El esquema de evaluación permaneció sin cambio desde el 2002, calificando cada respuesta como correcta, incorrecta o no soportada, esta última indicando que la respuesta es devuelta por el sistema pero el documento de la que se extrajo no contiene la información correcta. La evaluación arrojó como mejor resultado para las preguntas factuales el 71.3% de precisión sobre 362 preguntas [28].

En el *TREC* 2006 el requisito de sensibilidad a la dependencia temporal se hizo más explícito entre en la distinción entre respuestas correctas a nivel local y global, de tal forma que a preguntas dadas en el tiempo presente debían no solamente ser soportadas por documentos localmente correctos sino también debían ser la respuesta más actual en la colección de documentos (globalmente correcta). Para reflejar tal situación, se agregaron dos criterios de evaluación considerando que una respuesta podía ser localmente correcta, indicando con esto que la respuesta es correcta y es soportada por el documento que la acompaña sin embargo un documento reciente contradecía la respuesta o globalmente correcta para el caso contrario. La evaluación de las preguntas factuales continuó siendo la misma: la precisión del sistema definida como la fracción de respuestas consideradas correctas, el mejor resultado alcanzado bajo este criterio fue de 57.8% de 403 preguntas [29].

En la prueba *TREC* del 2007 se repitió el formato de preguntas dadas en series, pero con un cambio radical en el género de los documentos. En vez de solamente ser noticias, la colección de documentos se compuso además de *blogs*. En el *TREC* 2006 y 2007 se

agregó una prueba, la búsqueda de respuestas interactiva (*complex, interactive Question Answering, ciQA*). La finalidad de esta prueba es pasar de responder preguntas sobre hechos concretos a preguntas más complejas, así como enfocarse en sistemas que permitan interacción con los usuarios al contrario del modelo implícito impuesto en las pruebas anteriores del *TREC* basado en una sola interacción [30].

En el foro *Cross Lingual Evaluation Forum* las pruebas y preguntas son similares a las diseñadas en el *TREC* la principal diferencia es que se enfocan a los lenguajes europeos; por lo tanto se encuentran sistemas no solo monolingües para el francés, español, alemán, etc., sino también sistemas multilingües los cuales son desarrollados para realizar búsquedas en documentación en varios lenguajes, lo que impone retos aún mayores.

2.2.2 *SBR* de dominio restringido

El desarrollo de sistemas de búsqueda prácticos como se describió en secciones anteriores comienza precisamente con la construcción de sistemas de búsqueda de dominio restringido, sin embargo hoy en día la metodología general ha pasado de emplear bases de datos codificadas a mano referidas a dominios simples al empleo de colecciones de texto como fuente principal de conocimiento para dominios más complejos, como se mostró en la sección anterior. Sin embargo, aún existen muchos terrenos por explorar en la tarea de búsqueda de respuestas.

Una característica principal de los *SBR* de dominio restringido es la integración de información del dominio específico bajo el cual se trabaja, información que puede ser desarrollada especialmente para la tarea de búsqueda de respuestas o para propósitos distintos.

Se ha encontrado que los enfoques actuales y técnicas típicas empleadas en sistemas de búsqueda de dominio abierto no son adecuadas para aplicaciones reales [34] en las cuales las preguntas son de una dificultad mucho mayor que las abordadas en los foros *TREC*/*CLEF*. Ejemplos de tales aplicaciones son:

Interfaces a manuales digitales. Muchas aplicaciones de software son muy complejas y vienen acompañadas de una extensa documentación. Un sistema de búsqueda de respuestas que sea capaz de hallar respuestas específicas a preguntas de los usuarios sobre tal documentación podría ser de mucha utilidad.

Interfaces a fuentes de conocimiento. Muchas actividades humanas, así como cada disciplina poseen fuentes de conocimiento específico propias. Un ejemplo es el dominio médico, el cual posee una enorme cantidad de información técnica y fuentes de conocimiento que podrían ser de gran utilidad en el desarrollo de sistemas de búsqueda de respuestas.

Sistemas de ayuda en grandes organizaciones. El personal de asistencia en grandes organizaciones requiere dar respuestas rápidas a los clientes. No obstante de que la mayor parte de esta información se puede encontrar en repositorios de preguntas-frecuentes disponibles para el personal de asistencia, habrá siempre solicitudes de los usuarios que sean únicas y requieran que el personal posea mecanismos rápidos de acceso a la información.

Podría argumentarse que enfocar la investigación en dominios restringidos es limitante debido a que los resultados son muy específicos y no son susceptibles a generalizar. En [34] muestran que tal podría ser el caso de los primeros sistemas de búsqueda de respuestas, los que se enfocaron a dominios restringidos simplemente por las limitaciones en el poder computacional y el interés teórico. Pero este no es el caso en nuestros días. La disponibilidad de extensas fuentes de conocimiento confiables en dominios complejos permite el desarrollo de investigación provechosa que puede expandirse a dominios abiertos.

Actualmente, la investigación en sistemas de búsqueda de dominio restringido se ha concentrado a problemas relacionados con la incorporación de información del dominio en específico en el que se trabaja, esto con la esperanza de alcanzar un desempeño aceptable para aplicaciones reales. Sin embargo, se espera que la investigación en SBR en dominio restringido lleve a la convergencia de tanto la búsqueda de respuestas basada en fuentes de conocimiento estructurado como la basada en colecciones de documentos.

En especial, la investigación e incorporación de técnicas de Inteligencia Artificial en el área legal tiene ya varios años, sin embargo la mayor parte de esta investigación se ha llevado a cabo en el contexto de desarrollo de sistemas expertos, los cuales emplean principalmente fuentes de conocimiento tales como ontologías además de incorporar reglas que permiten inferir relaciones y extraer información de las mismas.

Recientemente en un taller especial del foro JURIX (2003[6]) se abordó el tema de la búsqueda de respuestas en el dominio legal, se presentaron distintos trabajos tratando principalmente la problemática, retos y perspectivas en el desarrollo de sistemas restringidos al área legal. Durante años posteriores en el foro JURIX, a pesar de no organizar una sesión especial para SBR limitados al dominio legal, han continuado recibiendo trabajos enfocados a dicha área, sin embargo no se han publicado trabajos enfocados específicamente a documentos normativos. Debido a que tal foro se lleva a cabo en Europa existen trabajos en distintos idiomas sin embargo no se encuentran trabajos para el idioma español. En [9] en JURIX (2005) se describe un SBR restringido a documentos jurídicos para el lenguaje portugués, éste se basa como detallan sus autores en el uso de técnicas de procesamiento de lenguaje natural (análisis sintáctico y semántico) y una interpretación semántica/pragmática utilizando ontologías e inferencia lógica. A la fecha de publicación del trabajo no reportan una evaluación de su sistema.

Debido a la menor investigación en SBR de dominio restringido y aún más en el dominio legal, no se incluye más literatura, sin embargo esto también indica que la tarea es un problema abierto interesante de investigar.

[6] http://www.cs.uu.nl/jurix03/

Capítulo III

3 Marco teórico

Nuestra propuesta para el Sistema de Búsqueda de Respuestas se basa principalmente en una estructura matemática ampliamente utilizada y sumamente importante en Ciencias de la Computación: los grafos. Además de dicha estructura se emplean herramientas, conceptos y definiciones de distintos campos; particularmente, técnicas tradicionalmente utilizadas en Procesamiento de Lenguaje Natural (*PLN*) con técnicas del área de Recuperación de Información (*RI*).

En las siguientes secciones del presente capítulo se proporciona un breve repaso de los conceptos de teoría de grafos, así como de las herramientas de *PLN* y *RI*, que son empleados en el presente trabajo.

3.1 Grafos

Los grafos son estructuras muy útiles para representar una amplia diversidad de situaciones debido a lo cual han sido utilizados para resolver una gran cantidad de problemas en áreas muy diferentes que van desde teoría de circuitos hasta procesamiento de lenguaje natural. A continuación, se proporcionan algunos conceptos básicos de la teoría de grafos comenzando con la definición de grafo: [17]:

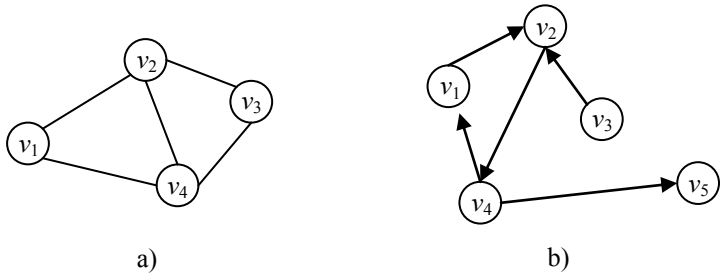
Definición.
Un **grafo** G consiste en un conjunto de vértices V (o nodos) y un conjunto de aristas (E) tal que cada arista $e \in E$ se asocia con un par no ordenado de vértices.

Existen en general dos tipos de grafos:

Definición.
Grafo no dirigido: Si existe una arista única e asociada con lo vértices v y w se escribe $e = (v, w)$ o $e = (w, v)$. En este contexto, (v, w) denota una arista entre v y w en un grafo no dirigido y no es un par ordenado.

Definición.
Grafo dirigido: Si hay una arista única e asociada con el par ordenado (v, w) de vértices, se escribe $e = (v, w)$, que denota una arista de v a w.

En la Figura 5, se muestran ejemplos de grafo dirigido y no dirigido. Se dice que una arista e en un grafo (no dirigido o dirigido) que se asocia con el par de vértices v y w es incidente sobre v y w, y se dice que v y w son incidentes sobre e y son vértices adyacentes. Si G es un grafo (no dirigido o dirigido) con vértices V y aristas E, se escribe $G = (V, E)$. A menos que se especifique lo contrario, los conjuntos V y E son finitos y V es no vacío.



a) b)

Figura 5. a) Grafo no dirigido, b) Grafo dirigido

La definición de grafo, no dirigido o dirigido, permite que diferentes aristas se asocien con el mismo par de vértices. Por ejemplo, en la Figura 6, las aristas e_1 y e_2 se asocian ambas con el par de vértices $\{v_1$ y $v_2\}$. Estas aristas se llaman *aristas paralelas*. Una arista incidente en un mismo vértice se llama *lazo*. Por ejemplo, en la Figura 6, la arista $e_3 = (v_2, v_2)$ es un lazo. Un vértice como v_4 en Figura 6, que no incide en ninguna arista, se llama *vértice aislado*. Una gráfica sin lazos ni aristas paralelas se llama *grafo simple*.

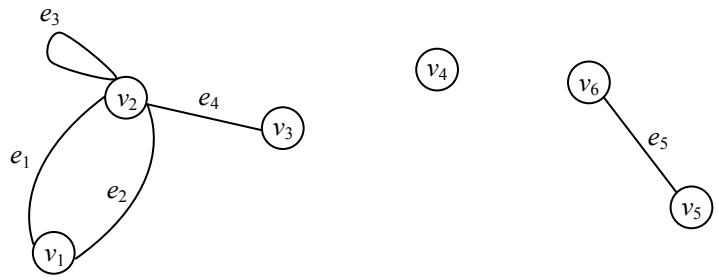


Figura 6. Grafo con aristas paralelas y un lazo

Existe un tipo de grafo que asocia valores a las aristas de vértices adyacentes, denominado grafo ponderado, de forma más precisa:

Definición.
Grafo ponderado: Un grafo con números en las aristas, se llama grafo ponderado. Si la arista e_j se etiqueta k_j, se dice que el peso de la arista es k.

En un grafo ponderado, Figura 7, la longitud de una ruta es la suma de los pesos de las aristas en la ruta.

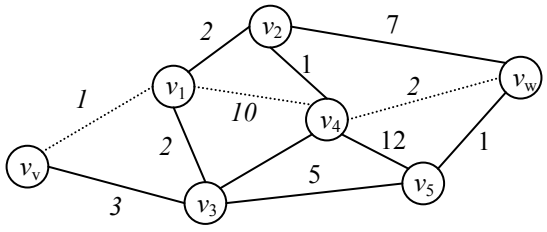


Figura 7. Grafo ponderado. La línea punteada indica una ruta entre el vértice *v* y *w*.

Para definir de una forma más precisa *ruta sobre un grafo* es necesario definir primero *trayectoria sobre un grafo*:

Definición.
Sean v_o y v_n vértices en un grafo. Una *trayectoria* de v_o a v_n de longitud n es una sucesión alternantes de $n + 1$ vértices y n aristas que comienza en el vértice v_o y termina en el vértice v_n. $(v_o, e_1, v_1, e_2, v_2, \ldots, v_{n-1}, e_n, v_n)$ Donde la arista e_i es incidente sobre los vértices v_{i-1} y v_i para $i = 1, \ldots, n$.

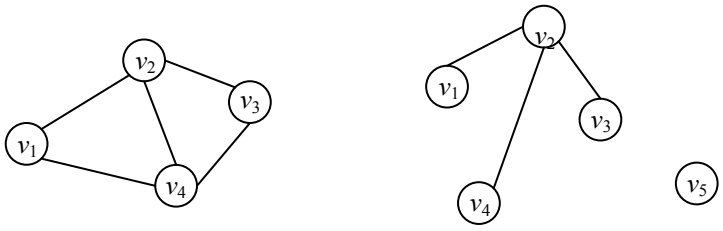
En términos del concepto trayectoria, un grafo puede ser conexo o no-conexo. De forma más precisa:

Un grafo G es *conexo* si dados cualesquiera dos vértices *v* y *w* en *G*, existe una trayectoria de *v* a *w*, en caso contrario es no conexo, en la Figura 8 se muestran ejemplos de ambos tipos de grafos.

Nótese que la definición de *trayectoria* permite repeticiones de vértices o aristas o ambos. Se pueden obtener otras clases de trayectorias imponiendo restricciones sobre los nodos y/o vértices. En particular nos interesa:

Sean *v* y *w* vértices en un grafo *G*.

Una *trayectoria simple* de *v* a *w* es una *ruta* de *v* a *w* sin vértices repetidos.



a) b)

Figura 8. a) Grafo conexo, b) Grafo no conexo

Existen algunos problemas para los cuales es útil determinar la ruta más corta entre un par de vértices en un grafo ponderado. A continuación se muestra un algoritmo que encuentra la ruta más corta entre dos vértices de un grafo ponderado conexo.

Algoritmo de la ruta más corta de Dijkstra: Este algoritmo encuentra la longitud de una ruta más corta del vértice *a* al vértice *z* en un grafo ponderado conexo. El peso de la arista (i, j) es $w(i, j) > 0$, y la etiqueta del vértice *x* es $L(x)$. Al terminar, $L(z)$ es la longitud de la ruta más corta de *a* a *z*.

El algoritmo de Dijkstra consiste en asignar etiquetas temporales a los nodos. Sea $L(v)$ la etiqueta del vértice v. En cualquier paso del algoritmo, algunos vértices poseen etiquetas temporales y el resto permanentes. Al inicio, todos los vértices tienen etiquetas temporales. En cada iteración del algoritmo el estado de una de las etiquetas temporales cambia a permanente, así el algoritmo termina cuando z recibe una etiqueta permanente. En este punto $L(v)$ proporciona la longitud de la ruta más corta de a a z.

Algoritmo de Dijkstra

Entrada: Un grafo G ponderado conexo en el que todos los pesos son positivos, conjunto de vértices de a a z.

Salida: $L(z)$, la longitud de la ruta más corta de a a z.

```
1.  Fdijkstra(w, a, z, L) {
2.     L(a) = 0
3.     Para todos los vértices x ≠ a
4.     L(x) = ∞
5.     T = conjunto de todos los vértices
6.     //T es el conjunto de todos los vértices cuyas distancias más cortas desde a
7.     //no se han encontrado
8.       While (z ∈ T) {
9.          Seleccionar v ∈ T con L(v) mínimo
10.         T = T – {v}
11.         Para cada x ∈ T adyacente a v
12.         L(x) = min { L(x) , L(v) + w(v, x) }
13.      }
14. }
```

3.2 Modelo de espacio vectorial

El modelo de espacio vectorial (MEV) es uno de los más ampliamente utilizados en el área de recuperación de información, debido principalmente a su simplicidad. El modelo ha sido utilizado para la recuperación de documentos por medio de solicitudes formadas de términos clave que expresan la necesidad de información de los usuarios; generalmente la recuperación se realiza sobre colecciones de documentos que pueden contener desde algunos cientos hasta miles de ellos.

En el MEV, cada documento de la colección es representado por un vector de términos extraídos de los documentos, con pesos asociados representando la importancia de las palabras en el documento y en la colección de documentos, de forma similar, dada una solicitud, ésta se modela como una lista de términos con pesos asociados que representan su importancia en la solicitud.

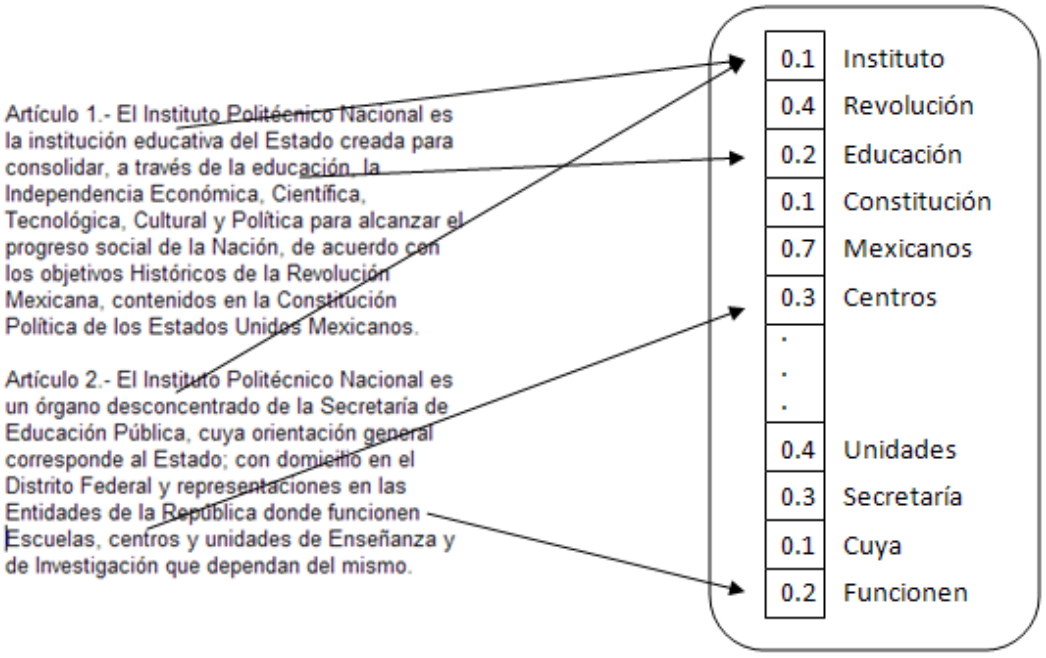


Figura 9. Representación documentos en el MEV

El valor de ponderación para cada término en un vector-documento puede determinarse de muchas formas. Un enfoque común emplea el método llamado *tf·idf* (*term frequency inverse document frequency*), en el cual el peso de cada término se determina a través de dos factores:

1) El número de ocurrencias del término *j* en el documento *i* (la frecuencia del término, *term frequency* $tf_{i,j}$), y

2) El número de ocurrencias en la colección completa de documentos (frecuencia en la colección de documentos, *document frequency* df_j). De forma precisa, el peso del término *j-ésimo* en el *i-ésimo* documento es:

$$w_{i,j} = tf_{i,j} * idf_j = tf_{i,j} * log\frac{N}{df_j} \quad \textbf{(4)}$$

Donde *N* es el número de documento en la colección de documentos.

Existen actualmente variantes del modelo tal como fue propuesto [15, 16] las que básicamente consisten en la elección de los términos que se eligen para representar tanto documentos como solicitudes, el modelo básico considera como términos las palabras sin excluir ninguna. El MEV básico se ilustra mediante la Figura 10 y Figura 11.

Solicitud, Q: "oro plata camión"
D_1: "cargamento de oro dañado en un fuego"
D_2: "entrega de plata arribó en un camión plata"
D_3: "cargamento de oro arribó en un camión"
$D = 3$; $idf = \log (D/df_i)$

Términos	Frecuencia tf_i							Peso, $w_i = tf_i * idf_i$			
	Q	D_1	D_2	D_3	df_i	D/df_i	idf_i	Q	D_1	D_2	D_3
a	0	1	1	1	3	3/3=1	0	0	0	0	0
arribó	0	0	1	1	2	3/2=1.5	0.1761	0	0	0.1761	0.1761
dañado	0	1	0	0	1	3/1=3	0.4771	0	0.4771	0	0
entrega	0	0	1	0	1	3/1=3	0.4771	0	0	0.4771	0
fuego	0	1	0	0	1	3/1=3	0.4771	0	0.4771	0	0
oro	1	1	0	1	2	3/2=1.5	0.1761	0.1761	0.1761	0	0.1761
en	0	1	1	1	3	3/3=1	0	0	0	0	0
de	0	1	1	1	3	3/3=1	0	0	0	0	0
plata	1	0	2	0	1	3/1=3	0.4771	0.4771	0	0.9542	0
cargamento	0	1	0	1	2	3/2=1.5	0.1761	0	0.1761	0	0.1761
camión	1	0	1	1	2	3/2=1.5	0.1761	0.1761	0	0.1761	0.1761

Figura 10. Ejemplo del modelo de espacio vectorial básico basado en *tf·idf*[7]

Como puede observarse de la anterior figura, el método de ponderación de términos asigna pesos altos a términos que aparecen frecuentemente en un número pequeño de documentos dentro de la colección y al contrario, los términos que aparecen frecuentemente dentro de la colección de documentos reciben valores de ponderación muy bajos e incluso para el ejemplo obsérvese cómo para las palabras en, de, a, los valores asignados por *tf·idf* son iguales a cero. Esto es muy importante ya que el método en sí puede, en ocasiones, discriminar por sí mismo términos que podrían no ser de importancia en la colección de documentos, reduciendo con ello el espacio de términos, lo que a veces puede ser de mucha utilidad.

Una vez que se determinan los pesos de cada término, es necesaria una función que permita medir el grado de semejanza entre la solicitud y los vectores documento. Una medida común de semejanza, conocida como medida de semejanza coseno, determina el

[7] Ejemplo adaptado al español de *http://www.miislita.com/term-vector/term-vector-3.html*

ángulo entre los vectores documento y el vector correspondiente a la solicitud, considerando un espacio euclidiano V – dimensional, donde V es el tamaño del vocabulario, como se ilustra en la Figura 11.

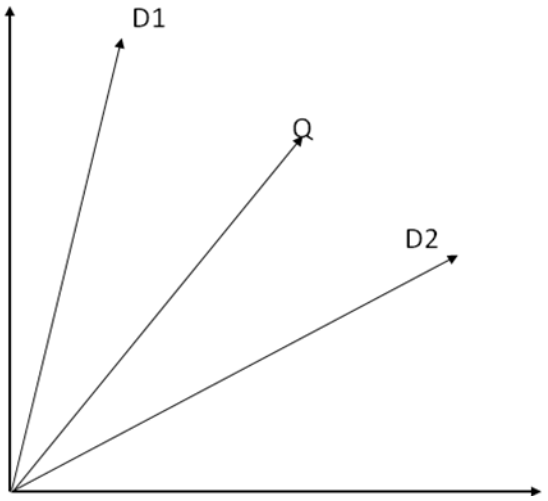


Figura 11. Modelo de espacio vectorial

La semejanza entre un documento D_i y una solicitud dada Q se define como:

$$sem(Q, D_i) = \frac{\sum_{j=1}^{V} w_{Q,j} * w_{i,j}}{\sqrt{\sum_{j=1}^{V} w_{Q,j}^2 * \sum_{j=1}^{V} w_{i,j}^2}} \qquad sem(q, Di) \in [0,1] \qquad (5)$$

Donde $w_{Q,j}$ es el peso del término j-ésimo en la solicitud, y se define de forma similar a $w_{i,j}$ (esto es, $tf_{Q,j} * idf_j$). El denominador en esta ecuación, llamado factor de normalización, no toma en cuenta el efecto de la longitud de los documentos. Entonces, un documento que contenga {x,y,z} tendrá exactamente la misma puntuación que otro documento que contenga {x, x, y, y, z, z} debido a que ambos se representarán por el mismo vector. Se puede debatir si esto es razonable o no, sin embargo cuando la longitud de los documentos varíe considerablemente, tendrá sentido tomarlo en cuenta.

La forma tradicional de implementar el modelo de espacio vectorial en un sistema de recuperación de documentos, consiste básicamente en crear el espacio euclidiano de vectores-documento, una vez que se tiene éste, la tarea del sistema de recuperación consistirá solamente en convertir las solicitudes introducidas al sistema en un vector-solicitud y posteriormente compararlo con todos los vectores-documento, de esta forma la salida

del sistema será una lista ordenada de documentos comenzando con el de mayor semejanza con la solicitud.

La forma básica de representar los documentos en el MEV es tal como se mostró en la presente sección, tomando solamente las palabras sin excluir ninguna. Sin embargo, dependiendo de la tarea, se pueden realizar modificaciones típicas de los sistemas de recuperación de información como las que a continuación se describen.

En algunos sistemas de recuperación de información no todas las palabras son utilizadas para representar los documentos, la lista de palabras no utilizadas son llamadas palabras sin contenido. Una lista de palabras sin contenido (en inglés llamadas *stopwords list*) enlista palabras consideradas de poca utilidad para tareas de búsqueda [14]. Palabras sin contenido comunes son, el, la, para. Estas palabras tienen un significado importante en español, sin embargo, su contribución en la recuperación de documentos puede ser en algunos casos poco significativa. Debido a que la lista de palabras sin contenido usualmente se compone de palabras que ocurren frecuentemente en los documentos, su eliminación puede reducir significativamente la norma de los vectores-documento y facilitar así su manipulación. Esto, sin embargo, tiene el efecto de que una vez que las palabras sin contenido son removidas es imposible realizar la búsqueda de una frase que las contenga; es por esto que muchos sistemas de recuperación de información no remueven las palabras sin contenido al representar los documentos pero sí en la solicitud introducida por los usuarios. Google por ejemplo, para búsquedas en inglés, no toma en cuenta palabras comunes como, el (the), y (and), como (how), donde (where), así como ciertos dígitos y letras, sin embargo da la opción de poder incluirlos en la búsqueda[8].

Para mejorar el modelo de espacio vectorial comúnmente se recurre a técnicas tradicionales del área de Procesamiento de Lenguaje Natural, mismas que se aplican principalmente (aunque no se restringen), a la colección de documentos. En la siguiente sección se muestran algunas de tales técnicas.

3.3 Métodos y recursos de Procesamiento de Lenguaje Natural

En el área de Procesamiento de Lenguaje Natural se manejan principalmente colecciones de documentos (llamadas corpus) a las que usualmente se realiza un pre-

[8] http://www.google.com/support/bin/answer.py?answer=981

procesamiento antes de ser utilizadas en las diversas tareas del área. Tal pre-procesamiento consiste básicamente en lo siguiente:

- Lematizar corpus
- Remover símbolos, caracteres especiales, etc.

Existen dos tipos de lematización:

En el área de recuperación de información, lematizar se le conoce al proceso de dada una palabra encontrar su "raíz" llamada en inglés *stem*. Debido a lo anterior, a este proceso se le conoce en inglés como *stemming* y solamente consiste en "cortar" la palabra, por ejemplo, dadas las palabras constitución, constitucional, constituciones, la raíz de ellas sería constituci-. Los principales algoritmos de *stemming* han sido desarrollados para el idioma inglés, el más conocido es el algoritmo de *stemming* de Porter [14]

Por otra parte, en el área de Procesamiento de Lenguaje Natural lematización se refiere al proceso de dada una palabra encontrar los principales fenómenos morfológicos en ella y "eliminarlos", llevando la palabra a su "forma básica". Por ejemplo, se reconoce el proceso de inflexión que indica, número y tiempo. La palabra de la izquierda es la proporcionada al lematizador y la de la derecha es la palabra que proporciona este mismo.

Estudiantes → estudiante

Trabaje→trabajar

Los lematizadores se basan en analizadores morfológicos para determinar la categoría gramatical de cada palabra y así determinar la raíz correcta.

También es común que entre la preparación de las colecciones de documentos se eliminen de los textos ciertos símbolos que son de poca utilidad, así como caracteres especiales, lo que depende del formato en que se encuentre la colección y de la aplicación que se dará a la colección. Por ejemplo, si se tratase de páginas html, es común eliminar todas las sentencias que forman parte del lenguaje, así como signos de puntuación. En algunos casos, nuevamente dependiendo de la tarea, la colección de documentos se convierte a minúsculas o mayúsculas y se eliminan dígitos y letras individuales; así como palabras sin contenido.

Por otra parte, existen ciertos recursos de PLN útiles en otras tareas. En el presente trabajo solamente se describen los tesauros como ejemplo de tales recursos.

Un tesauro en su forma más general se puede considerar como una lista de palabras y asociadas a ellas otras relacionadas. A continuación se muestra un ejemplo[9] para la palabra coche.

Sinónimos: auto, automóvil, máquina, motor, vehículo.

Palabras relacionadas: autobús, carruaje, micro-bus, mini-bus, buggy, sedan, compacto, convertible, cupé, jeep, limousine, todo-terreno, deportivo, camioneta, pipa, subcompacto, van.

Como se puede observar del ejemplo anterior, un tesauro da como resultado un conjunto de palabras que pueden ser desde los mismos sinónimos de la palabra considerada hasta un conjunto de palabras muy relacionadas. Éstas últimas dependen de la construcción del tesauro y/o del alcance. Considerando la forma en que construye un tesauro se pueden encontrar en general dos tipos:

- Construidos de forma manual — Éstos son elaborados por humanos, quiénes se encargan de recolectar el conjunto de palabras relacionadas que consideran pueden asociarse a una palabra determinada.
- Construidos de forma automática (tesauros distribucionales) — Un tesauro distribucional se construye de forma automática por medio de un algoritmo computacional a partir de un corpus determinado. Su característica principal es que obtiene las palabras relacionadas dependiendo del contexto en que se encuentren. Por ejemplo, si se encuentra la palabra *nube* y *lluvia*, el tesauro puede incluir estas relaciones; sin embargo, si encuentra también *nube de electrones*, agregará nube y electrón también como palabras relacionadas.

Se pueden distinguir además otros dos tipos de tesauros, por el tipo de aplicación que se le dará:

- Tesauros de dominio específico — Son tesauros (manuales o automáticos) construidos a partir de corpus referidos a un solo tema p. ej. medicina.
- Tesauros de dominio general — Abarcan palabras de distintas áreas y tienen un vocabulario general.

[9] Ejemplo basado en el tesauro en línea Merriam Webster (http://www.merriam-webster.com)/

Entre una de las aplicaciones que se le han dado a los tesauros se encuentran la expansión de términos para tareas como búsqueda de respuestas, recuperación de documentos, clasificación, etc.

Las aplicaciones típicas de los tesauros en sistemas de recuperación de información son principalmente para expandir los términos de las solicitudes de los usuarios, mediante lo cual se intenta ampliar la búsqueda. De esta manera se pretende recuperar documentos que quizá expresen la información deseada en una forma distinta.

Capítulo IV

4 Descripción del *SBR*

En el presente capítulo, se describe la construcción del Sistema de Búsqueda de Respuestas para el dominio legal, en función del tipo de pregunta-respuesta que se desea maneje el sistema. Se destacan las características más importantes de la arquitectura y su implementación.

4.1 Delimitación del conjunto de preguntas–respuestas

El punto de partida es la descripción del conjunto de preguntas-respuestas que se utilizó para el desarrollo del trabajo, debido a que es la base de éste. En lo que respecta al tipo de preguntas, se utilizó un conjunto de preguntas hechas por la comunidad del Instituto Politécnico Nacional (IPN) y contestadas por el Abogado General del IPN a través de su página electrónica. El conjunto de preguntas original está constituido por 42 sentencias interrogativas, de las cuales se eligieron solamente 21 de ellas, en función de la complejidad de la misma pregunta y la respuesta otorgada por el Abogado General. A continuación se muestran algunos ejemplos representativos, tanto de las preguntas seleccionadas como de las no elegidas. El conjunto completo de preguntas seleccionadas se encuentra en el Apéndice C.

Tipo de preguntas elegidas

- ¿Cómo se lleva a cabo el procedimiento de elección de representantes alumnos ante el Consejo Técnico Consultivo Escolar?
- ¿Cuáles son los consejeros que deben abstenerse de participar en procesos de elección de terna para la designación de Director?
- ¿Es procedente que un alumno solicite mención honorífica a nivel licenciatura si escogió su Titulación por la opción de Escolaridad?

Tipo de preguntas no elegidas

- ¿Cuál es la interpretación de las fracciones II y VIII del artículo 107 del Reglamento Interno, en relación a aquellos alumnos que dentro de las instalaciones de la unidad, juegan baraja, domino, dados, ruleta etc., o que en sus conversaciones utilicen lenguaje soez, altisonante o albur?
- ¿Cómo se debe hacer la interpretación jurídica del artículo 174, fracción IV -... tener una antigüedad mínima de cinco años realizando actividades académicas en el Instituto-?

Como puede observarse, en este último tipo de preguntas se solicitaba al Abogado la interpretación de determinado artículo de la normatividad del Instituto. Por esta razón, posiblemente este tipo de preguntas quedarían fuera del alcance de cualquier sistema hasta ahora concebido, ya que la interpretación requiere de cierto raciocinio que no se tiene con la simple lectura del artículo, sino sólo con cierta experiencia adquirida. A partir de que en este trabajo se está proponiendo un nuevo *SBR*, puede considerarse por qué es deseable partir de preguntas relativamente sencillas que permitan explorar las bondades y las limitantes del sistema propuesto.

A continuación se muestran las respuestas a la preguntas mostradas anteriormente, con la finalidad de analizar sus características y verificar su grado de complejidad.

Respuestas a preguntas elegidas

¿Cómo se lleva a cabo el procedimiento de elección de representantes alumnos ante el Consejo Técnico Consultivo Escolar?

El proceso de elección de representantes alumnos ante los consejos técnicos consultivos de las escuelas, centros y unidades de enseñanza media superior está regulado en

los **artículos 28 fracción IV de la Ley Orgánica, 206, 207, 209 y 213 fracción I del Reglamento Interno**, dispositivos retomados en las bases de la convocatoria autorizada mediante oficio número 014 SG/634/01, inscrita en la oficina del Abogado General con el número 582 a fojas 134 del libro de Registro de Actos Jurídicos Diversos.

El artículo 213 fracción I del Reglamento Interno preceptúa que los únicos requisitos para ser candidato a consejero alumno consiste en tener situación escolar regular y estar cursando estudios entre el tercero y penúltimo semestres. El artículo 209 del Reglamento Interno establece que comités conformados por alumnos con situación escolar regular son quienes deben elegir, en forma directa y por mayoría de votos, a los representantes educandos entre el Consejo Técnico Escolar.

¿Cuáles son los consejeros que deben abstenerse de participar en procesos de elección de terna para la designación de Director?

El **artículo 181 del Reglamento Interno** del Instituto Politécnico Nacional dispone que los directores y subdirectores de las escuelas, centros y unidades de enseñanza son los únicos integrantes de los Consejos Técnicos Escolares que deben abstenerse de participar en las sesiones de dichos órganos colegiados en las que se trate de elegir las ternas para los cargos que ellos mismos ocupan.

¿Es procedente que un alumno solicite mención honorífica a nivel licenciatura si escogió su Titulación por la opción de Escolaridad?

Del análisis integral del **Reglamento de Titulación Profesional** del Instituto Politécnico Nacional se desprende lo siguiente: La opción de titulación por escolaridad resulta procedente si el promedio general (del alumno) es superior a nueve y todas las materias fueron aprobadas en forma ordinaria (Capítulo II. "De las Opciones de Titulación" **Artículo 13**). El pasante solo puede aspirar a recibir mención honorífica, si además de cubrir otros requisitos previstos en este reglamento, presenta examen profesional (Capítulo VII "Del Examen Profesional", **artículo 43**). La titulación por escolaridad no requiere de la presentación de examen profesional, por lo tanto, no se puede enmarcar en el artículo 43 fracción II del Reglamento en comento; al elegir esta opción de titulación, el pasante no puede obtener mención honorífica.

Respuestas a preguntas no elegidas

¿Cómo se debe interpretar "tener la calidad de personal académico de tiempo completo, con nombramiento definitivo, categoría dictaminada de titular", señalado en el artículo 169, fracción III, del Reglamento Interno?

Debe entenderse como todo personal académico de carrera (quien tiene a su cargo la responsabilidad de las actividades que lleven a la realización y superación académica e investigación dentro del Instituto) que labore con un nombramiento de 40 horas de trabajo por semana.

¿Cuál es la interpretación de las fracciones II y VIII del artículo 107 del Reglamento Interno, en relación a aquellos alumnos que dentro de las instalaciones de la unidad, juegan baraja, domino, dados, ruleta etc., o que en sus conversaciones utilicen lenguaje soez, altisonante o albur?

En efecto, el hecho que alumnos dentro de las instalaciones del Instituto dediquen su tiempo a la práctica de juegos de azar, cualesquiera que éstos sean, desvirtúa la misión de educar que tiene nuestra Casa de Estudios (artículo 1 de la Ley Orgánica), con o cual se demerita su prestigio académico; por lo que quienes incurran en las conductas descritas incumplen el precepto normativo establecido en la fracción II del artículo 107 del Reglamento Interno. Por otra parte, por lo que se refiere a la fracción VIII del artículo 107 del Reglamento Interno, en su hipótesis normativa prescribe como una obligación de los alumnos el guardar respeto a los miembros de la comunidad politécnica y a los visitantes del Instituto, a lo que sí tomamos en cuenta que el Diccionario de la Lengua Española de la Real Academia Española, define en una de sus acepciones la palabra respeto como "Miramiento, consideración, atención, causa o motivo particular", y consideración está definida como urbanidad. En este tenor si algún alumno se conduce para con otro con lenguaje soez o altisonante, entendidos éstos, como lo define el diccionario en cita, como "bajo, grosero, indigno, vil", luego entonces, está faltando a la obligación de respeto que le está prescrito en dicho dispositivo normativo. La actualización de los supuestos anteriores de conformidad con la fracción I del artículo 108 del Reglamento Interno es causa de responsabilidad, lo que propiciaría el tener que aplicar alguna de las sanciones previstas

en el artículo 110, tomando en cuenta los criterios establecidos en el artículo 111 del Reglamento Interno.

De los ejemplos anteriores, se pueden observar que las principales características de las respuestas de las preguntas seleccionadas son:

- Se responden directamente a partir de la información contenida en los artículos referidos a la normatividad del IPN.
- La respuesta no sólo se encuentra en más de un artículo del mismo documento sino incluso en diferentes documentos.
- Parte de los artículos de ciertas respuestas se relacionan directamente con las preguntas a través de solamente palabras en común que comparten, sin embargo el resto de artículos que componen la respuesta no contienen palabras que expresa la pregunta pero sí comparten información con los artículos que directamente se relacionan con la pregunta y además son parte esencial de la respuesta (véase el primer ejemplo de pregunta-respuesta seleccionada).
- A pesar de que en ocasiones (tercer ejemplo), es necesario un proceso de razonamiento mínimo para obtener una respuesta final, como por ejemplo, un sí o un no, la respuesta aún así se encuentra básicamente en los artículos dados como respuesta.

En contraste, el grado de dificultad de las respuestas de las preguntas no elegidas se debe a:

- En general, la respuesta no sólo requiere de interpretar cierta información normativa del IPN, sino también de otras fuentes e incluso de los propios conocimientos del Abogado debidos a su formación académica.

Considerando el conjunto de preguntas-respuestas seleccionado el sistema de búsqueda de respuestas se desarrolló para aceptar en principio las preguntas tal como fueron enviadas al Abogado, sin embargo en lo que respecta a las respuestas solamente se limitan a los artículos que responden la pregunta, tal como se muestra en la Figura 12.

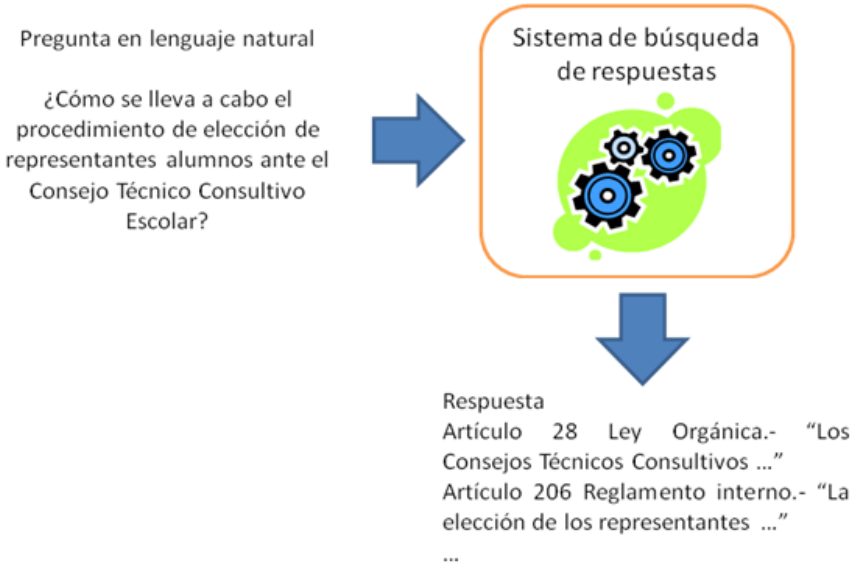


Figura 12. Entrada/Salida del *SBR* propuesto

A partir del conjunto de preguntas/respuestas seleccionadas también se definieron los textos que conformaron la colección de documentos a utilizar en el sistema, la cual, por supuesto, incluye la normatividad del IPN, misma que se encuentra en la página del Abogado General. Se enlistan a continuación algunos de los 26 documentos que el sistema utiliza, en el apéndice D se enlista el conjunto total de documentos:

Ley Orgánica

Reglamento orgánico

Reglamento interno

Reglamento de las condiciones generales de trabajo del personal no docente

Reglamento de las condiciones interiores de trabajo del personal académico

Reglamento del consejo general consultivo

Reglamento de titulación profesional

Las características que se consideran más importantes en estos documentos, son:

- Los documentos normativos tienen una estructura definida, se dividen en artículos, mismos que se agrupan en capítulos, los que pueden dividirse en secciones y éstas a su vez, en ocasiones, constituyen los capítulos que se agrupan en títulos; por ejemplo, la Ley Orgánica del IPN consta de 34 artículos divididos en seis capítulos, mientras que el Reglamento del Archivo Histórico se compone de 72

artículos divididos en 10 capítulos y el capítulo décimo dividido a su vez en diez secciones.

- Además de la estructura que poseen, en los textos normativos se encuentra que, a pesar de las divisiones mencionadas de los documentos existe una cierta relación entre artículos de un mismo documento e incluso de distintos documentos normativos, lo que se puede apreciar a través de las múltiples referencias que se encuentran en los documentos, por ejemplo:

  Reglamento del Archivo Histórico

  Artículo 5°. " … no cuenten con el espacio físico a que se *refiere el artículo anterior* para el resguardo de la documentación …"

  Artículo 21. " … dispuesto en los artículos 17, 18, 19 y 20 del presente reglamento …"

  Ley Orgánica

  Artículo 15.- "El Secretario General será nombrado por … y *deberá reunir los requisitos señalados en el artículo 13 de esta Ley*"

  Reglamento orgánico

  Artículo 7. "El Abogado General del Instituto será designado en los términos de lo dispuesto por los a*rtículos 23 de la Ley Orgánica y 144 del Reglamento Interno* y ...".

A continuación, se describe detalladamente la forma en que las características más importantes, descritas en la presente sección, tanto del conjunto de preguntas-respuestas seleccionado, así como de la colección de documentos, se reflejan en el mecanismo propuesto para la búsqueda y extracción de respuestas.

4.2 Esquema propuesto para la búsqueda de respuestas

Una de las formas más utilizadas para la representación de documentos ha sido la del modelo de espacio vectorial, no solamente para la recuperación de los mismos sino también para su clasificación; sin embargo en el contexto de un sistema de recuperación de documentos, el MEV, solo captura la semejanza que existe entre la solicitud expresada

por el usuario y los documentos, y no considera la semejanza existente entre los documentos mismos, lo que podría ser de mucha utilidad ya que de esta forma se podrían extraer documentos que quizá no estén relacionados de forma explícita con la solicitud del usuario, pero si se encuentren relacionados fuertemente con algún documento que tiene relación directa con la solicitud, de tal forma que quizá el documento así recuperado arroje un mejor resultado.

En el presente trabajo, se utiliza el modelo de espacio vectorial solamente en lo que respecta a la transformación de la colección de documentos a vectores, ya que para capturar la estructura de los textos normativos, se decidió tomar cada artículo de éstos como un documento. Con la finalidad de reflejar las relaciones existentes entre los artículos así como la estructura de los textos normativos, en lugar de representar cada vector-artículo en un espacio euclidiano como en el MEV, se escogió la representación en forma de un grafo ponderado en el que los nodos representan los artículos de los textos normativos y los valores asociados a las aristas el inverso de la semejanza entre los artículos, valor dado por la comparación de todos los vectores-artículo entre sí empleando la medida de semejanza coseno, tal como se ilustra en la Figura 13.

En caso que la semejanza entre un par de nodos sea igual a cero no existe enlace entre ellos.

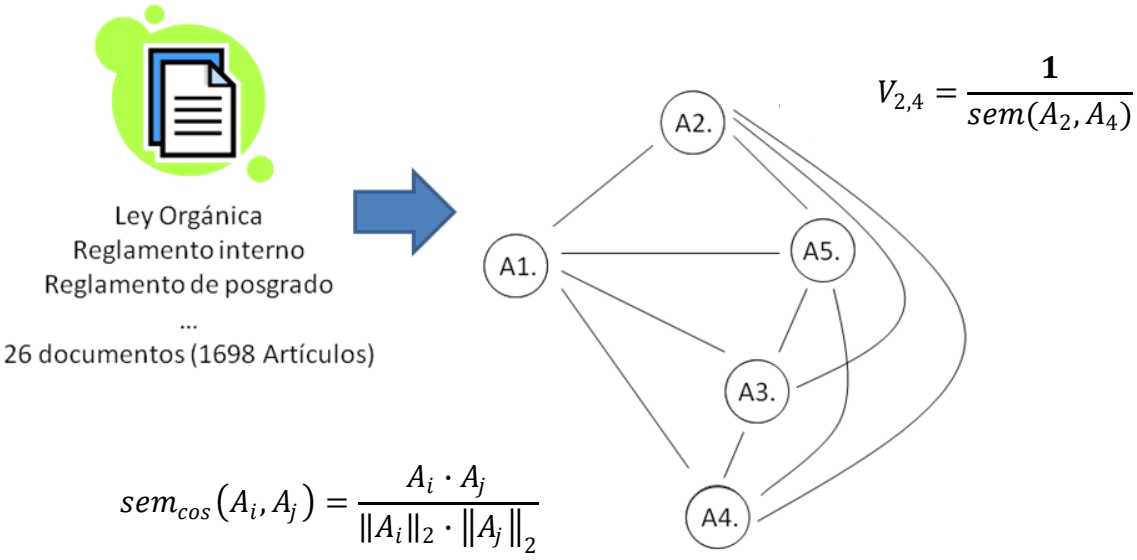


Figura 13. Representación de los documentos normativos

De forma similar que en el MEV, donde dada una solicitud ésta se convierte a un vector-solicitud, el cual se agrega al espacio vectorial euclidiano y después es comparado con los vectores-documento; en nuestro esquema propuesto, dada una pregunta ésta se divide en dos partes, A y B, mismas que son agregadas al grafo ponderado como si se tratara de dos nuevos vectores-artículo. En contraste con el MEV, en donde se eligen como salida los vectores-documento de mayor semejanza con la solicitud; en el presente trabajo se propuso un mecanismo diferente, basado en el mismo grafo de artículos, en el cual se determinaron las rutas de peso mínimos entre el nodo A y el nodo B. Recordando que el peso de las aristas en el grafo viene dado por el valor inverso de la semejanza, mediante este método se espera que no sólo se recuperen aquellos artículos que en principio se encuentren altamente relacionados con las partes de la pregunta agregadas al grafo, sino también aquellos que quizá no se encuentren directamente relacionados con la pregunta pero sí con los artículos que lo están de forma inmediata, tal como se ilustra en la Figura 14, en esta figura se muestra una posible ruta, en la cual el nodo A5 que no tiene relación directa con los vértices A y B sería devuelto por su relación con los nodos A4 y A2 que si están vinculados directamente con ellos.

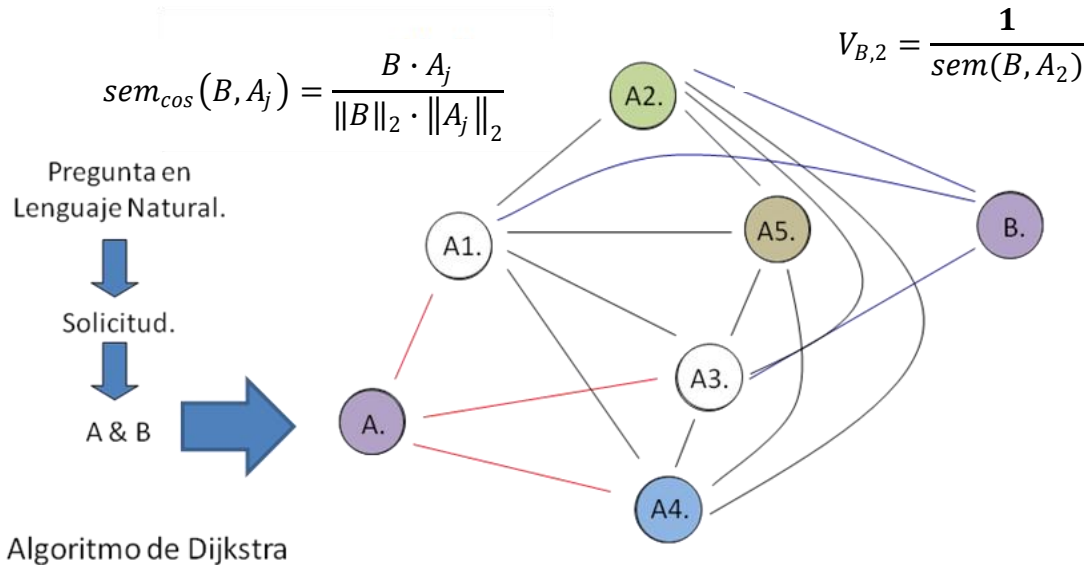


Figura 14. Esquema general propuesto para la búsqueda de respuestas.

4.3 Arquitectura

En función del esquema de búsqueda descrito en la sección precedente, se derivó la arquitectura del *SBR* (Figura 15), la que esencialmente consiste de dos módulos: un módulo está encargado de dividir la pregunta en los nodos A y B y el otro módulo está encargado de agregarlos al grafo de artículos y de encontrar las rutas de peso mínimo, lo que se realiza empleando el algoritmo de Dijkstra. A continuación se describe de forma general el funcionamiento de este par de módulos y de la arquitectura en general para la tarea de búsqueda de respuestas.

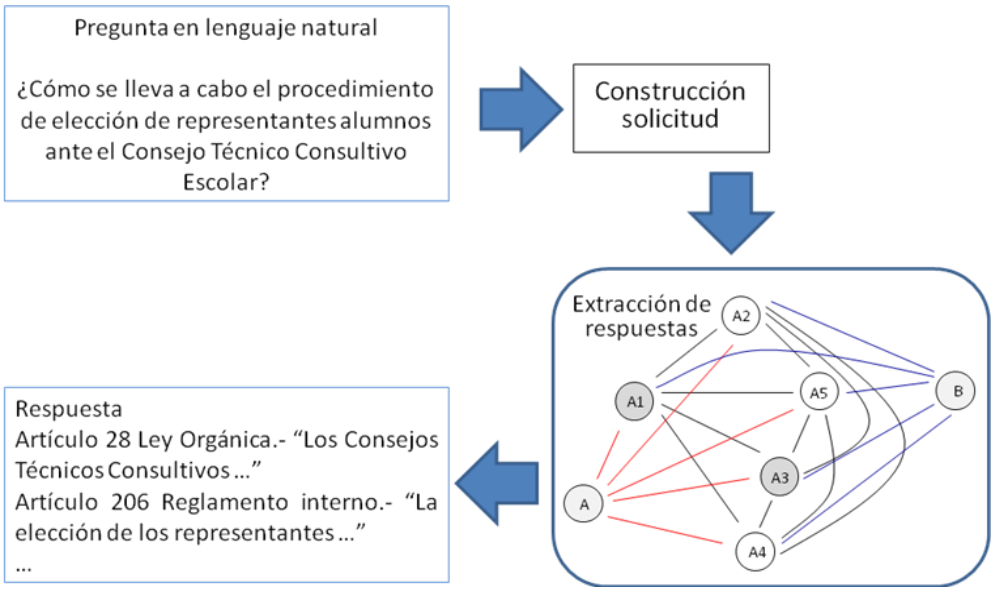


Figura 15. Arquitectura *SBR*

4.3.1 Descripción módulo construcción solicitud

A este módulo le corresponde la tarea de dividir la pregunta introducida al sistema en dos nodos, A y B. Como es de esperarse, el tipo de división realizado tendrá un efecto en el proceso de búsqueda extracción de respuestas. Por tal motivo, se realizaron una serie de experimentos con la finalidad de determinar una división óptima. Los tipos de divisiones y resultados obtenidos se detallan en el capítulo siguiente.

4.3.2 Descripción módulo extracción respuestas

Las partes de la pregunta obtenidas del módulo anterior son agregadas al grafo ponderado de artículo, y posteriormente se lleva a cabo lo que llamamos extracción de respues-

tas. La extracción de respuestas consiste en encontrar rutas de peso mínimo entre los nodos A y B que representan la pregunta de la siguiente manera:
Se ejecuta el algoritmo de Dijkstra, una vez que se encuentra una ruta se eliminan del grafo los nodos que la componen, el proceso se repite hasta el punto, en el que los vértices que representan la pregunta quedan aislados, es decir el grafo se vuelve no-conexo. Las *rutas respuesta* se ordenan de menor a mayor peso y se retornan al usuario con el texto de cada artículo.

4.4 Implementación

La implementación de la arquitectura del *SBR* se hizo utilizando el lenguaje de programación PERL versión 5.8.8.822. Esto debido a las herramientas que proporciona para el manejo y procesamiento de archivos de texto, así como a la amplia información disponible. Se escribieron dos *scripts* principales:

Uno encargado de construir el grafo ponderado de artículos y almacenarlo para su posterior uso y el script correspondiente al *SBR*. Los scripts escritos se incluyen en el anexo D.

A continuación se explican de forma general los scripts antes mencionados[10]:

a) Script encargado de construir el grafo

En éste se obtienen todos los artículos de los documentos normativos, se les aplica un sencillo pre-procesamiento y a partir de ellos se crean los vectores de términos. Una vez hecho esto, se crea un *índice invertido*, mismo que es de gran utilidad para el cálculo de los valores de ponderación de los términos extraídos de los documentos. Terminado el proceso de ponderación de términos se comparan los vectores entre sí y a partir de los

[10] Como primer enfoque se había decidido implementar el SBR por módulos; sin embargo el proceso de ejecutar Dijkstra, eliminar nodos y ejecutar de nuevo el algoritmo resultó lento (aproximadamente 4 minutos por pregunta). Debido a esto el SBR se implementó en un solo script sin emplear incluso sub-funciones ya que las variables que se pasan a las sub-funciones se realiza por valor y no por referencia, de tal forma que cada vez que se pasaba la matriz de adyacencia que representa el grafo ponderado de artículos el script consumía un tiempo considerable (80 minutos tomaba procesar 21 preguntas). Después de modificar el script tal como se encuentra en el anexo, se logró disminuir el tiempo a 35 minutos al procesar las 21 preguntas; las pruebas se realizaron en un equipo con 2 GB de memoria y procesador Intel Duo a 2.2 GHz

Finalmente, cabe aclarar que el tiempo que le toma al SBR contestar una pregunta puede ser optimizado, puesto que el lenguaje de implementación Perl es interpretado y por tal razón se puede consider más lento en comparación de uno compilado, como C.

valores obtenidos de la comparación se crea el grafo ponderado de artículos (mismo que se almacena en su forma típica de *matriz de adyacencia*).

b) Script del *SBR*

La parte fundamental del sistema de búsqueda de respuestas es el grafo ponderado de artículos, la tarea del script del *SBR* es primeramente realizar una copia del grafo ponderado de artículos, una vez realizado esto y cada vez que el sistema recibe una pregunta, ésta es pre-procesada y posteriormente agregada al grafo. Después de hecho esto, se ejecuta el algoritmo de Dijkstra y se obtiene la primera ruta de peso mínimo. Los artículos encontrados se devuelven al usuario y posteriormente se "eliminan" los vértices de la copia del grafo ponderado de artículos y se ejecuta nuevamente el algoritmo y así sucesivamente hasta el punto en que los vértices que representan la pregunta quedan aislados.

A continuación se muestran los pseudocódigos de los módulos del SBR.

```
//Pseudocódigo construcción del grafo
//Se transfieren artículos de los documentos a un arreglo.
k = 0  //Contador de artículos
Para j = 1 hasta 26 (documentos en la colección) hacer:
   //Se abre archivo de texto Lj
   Para i = 0 hasta n (número de líneas en Lj) hacer:
      Eliminar de Lj,i: espacios en blanco al inicio
      Si Lj,i inicia con "Artículo n." hacer:
         c = 0
         Artsk,c = "Artículo n."
         Incrementar k en 1
         Eliminar de Lj,i  "Artículo n."
         Convertir el resto de Lj,i a minúsculas y eliminar signos de
         puntuación,números y símbolos.
         c = 1
         Artsk,c = Lj,i
      En caso contrario hacer:
         Si c == 1 hacer:
            Convertir Lj,i a minúsculas
            Eliminar de Lj,i signos de puntuación, números y símbolos
            Artsk,c = Artsk,c concatenado con Lj,i
         Fin si
      Fin si
   Fin de para
Fin de para

//Se utiliza un arreglo hash (terms) para almacenar los términos de
//cada artículo y en cuantos artículos aparece cada uno
//se emplea un hash auxiliar (ht)

Separar el contenido de Arts1,1 por espacios en blanco
Para w = 0 hasta l (número de términos en el artículo) hacer:
   terms{término w} = 1
```

```
Fin de para

Para u = 2 hasta k (número total de artículos) hacer:
   Separar el contenido de Arts_{u,1} por espacios en blanco
   Para w = 0 hasta l (número de términos en el artículo) hacer:
      terms{término w} = terms{término w} + 1
      ht{término w} = ht{término w} + 1
   Fin de para
   Para w = 0 hasta l hacer:
      terms{término w} = terms{término w} + 1 – ht{término w}
   Fin de para
Fin de para

//Se calcula tf idf y se forman vectores-artículo mismos que se almace-
nan
//en el hash vec_art{Arts_{u,0}}
//Se almacena la ocurrencia de términos de cada artículo en el hash
//auxiliar ht con esta información junto termsart_k y terms{término w}
se
//obtiene el valor de ponderación de cada término

Para u = 1 hasta k (número total de artículos) hacer:
   Dividir el contenido de Arts_{u,1} por espacios en blanco
   Para w = 0 hasta l (número de términos en el artículo) hacer:
      ht{término w} = ht{término w} + 1
   Fin de para
   Se eliminan o no de ht{término w} lista de palabras sin contenido
   nt = 0  //Variable para almacenar número de términos en ht{término
w}
   Para cada elemento del hash ht{término w} hacer:
      nt = nt + 1
   Fin de para
   termsart_k = nt  //Número de términos en el artículo k
   Para w = 0 hasta termsart_k hacer:
      tfidf = (ht{término w}/termsart_k)*log(1698/ terms{término w})
      aux = aux concatenado con término w concatenado con | concatenado
      con tfidf concatenado con #
   Fin de para
   vec_art{Arts_{u,0}} = aux

   //Se comparan vectores-artículo y se genera matriz de adyacencia
   Para n = 1 hasta k – 1 hacer:
      magnv1 = 0
      magnv2 = 0
      Se separan los términos de vec_art{Arts_{n,1}} el resultado se alma-
cena
      en el arreglo tmp
      Para cada elemento de tmp hacer:
         Separar cada elemento de tmp por carácter |. Se supone se
         almacena resultado en arreglo tmp1
         magnv1 = magv1 + tmp_{1,1}
         vart1{tmp_{1,0}} = tmp_{1,1}   //Se almacena en hash vart cada
                                        //término del vector-artículo
      Fin de para
      ma{Arts_{n,1}#Arts_{n,1}} = 1e15
      Para t = n + 1 hasta k hacer:
         Se separa vec_art{Arts_{n,1}} por el carácter #
```

```
         Para cada elemento de tmp hacer:
            Se separa cada elemento de tmp por el carácter |
            magnv2 = magv2 + tmp1,1
            vart2{tmp1,0} = tmp1,1
         Fin de para
         pp = 0   //Variable que almacena producto punto
         Para cada elemento (llave) de vart1 hacer:
            pp = pp + vart1{llave}* vart2{llave}
         Fin de para
         sem = pp / ( raíz(magnv1) * raíz(magnv2) )
         Si sem es diferente de 0 hacer:
            sem_1 = 1 / ( 100 * sem )
            ma{Artst,1#Artsn,1} = sem_1
            ma{Artsn,1#Artst,1} = sem_1
         En caso contrario hacer:
            ma{Artst,1#Artsn,1} = 1e15
            ma{Artsn,1#Artst,1} = 1e15
         Fin si
      Fin de para
   Fin de para
   Finalmente se transfiere el contenido del hash ma a un archivo o BD,
así como también el conjunto de vectores-artículo


//Pseudocódigo SBR.
//Se emplea un procedimiento similar a la construcción del grafo
//debido a lo cual algunos pasos solo se mencionan
Se obtiene la matriz de adyacencia y se almacena en mai,j
Se obtiene el conjunto de vectores-artículo y se almacena en vak
Dada una pregunta se almacena en qi
Se lematiza qi y se expanden términos
Se convierte qi a minúsculas y se eliminan signos de puntuación
Se eliminan de qi términos que no existen en los reglamentos
Se eliminan o conservan palabras sin contenido
Se divide qi en dos partes (Q1 y Q2) dependiendo del esquema utilizado
Se asigna peso a los términos de Q1 y Q2
Se compara Q1 con el conjunto de vectores artículo considerando Q1 como
el artículo 1699. Se realiza lo mismo para Q2 (artículo 1700). Los re-
sultados de la comparación se agregan a la matriz de adyacencia
Se ejecuta algoritmo de dijkstra sobre la matriz de adyacencia. Entrada
algoritmo. Nodo inicial: 1699. Nodo final 1700. Número de nodos 1700.
//Se utiliza el algoritmo de Dijkstra mostrado en la sección 3.1
//La salida del mismo es:
//ruta (r)=Nodo inicial-> nodo 1 -> nodo 2 -> … -> nodo n -> Nodo final
//Peso de la ruta (p).En caso que el algoritmo devuelva de peso de la
//ruta un número mayor a 1e15 no hay relación entre los vértices
//1699 y 1698
Nodo 1, nodo 2, … nodo n se mapean a sus respectivos artículos y se al-
macenan en el mismo orden que da el algoritmo de ir del nodo 1699 al
1700.
//Se eliminan los nodos respuesta de la matriz de adyacencia
Para i = 1 hasta n (número de nodos respuesta) hacer:
   Para j = 1 hasta 1698 hacer:
      mai,j = 1e15
   Fin de para
Fin de para
Mientras p < 1e15 hacer:
```

```
 Ejecutar algoritmo de Dijkstra
 Si p < 1e15 hacer:
    Nuevamente:
    Nodo 1, nodo 2, … nodo n se mapean a sus respectivos
    artículos y se almacenan en el mismo orden
    Se eliminan los nodos respuesta de la matriz de adyacencia
 Fin si
Fin mientras
```

.

Capítulo V

5 Metodología Experimental

Existen parámetros importantes de los que depende el desempeño general del *SBR* propuesto. En las siguientes secciones se describen los experimentos que se llevaron a cabo para investigar el impacto de tales parámetros, así como los criterios propuestos para la evaluación de los experimentos y al final se discuten y reportan los resultados obtenidos.

5.1 Experimentos

En general, los experimentos consistieron en realizar variaciones en los siguientes parámetros del sistema:

- En la *Colección de documentos*: Se aplicaron técnicas de procesamiento de lenguaje natural con el objetivo de verificar el impacto de su aplicación en el sistema propuesto. Las técnicas fueron específicamente:
    a) Lematización de la colección de documentos
    b) Eliminación de palabras sin contenido
- En la *Pregunta*: Se evaluaron las respuestas en función de la variación de los siguientes parámetros:
    a) División

b) Expansión de pregunta a través del uso de tesauros (diccionarios de ideas afines)

c) Eliminación de las palabras sin contenido

5.1.1 Colección de documentos: Construcción del grafo

Primeramente se comenzó con la preparación de los documentos, que consistió en el siguiente pre-procesamiento:

- La colección de documentos se convirtió a minúsculas.
- Se eliminaron los signos de puntuación, paréntesis, numeraciones romanas, números, guiones, letras de la *a* la *z*.
- Se eliminaron los artículos transitorios de los documentos. Se consideró que los artículos transitorios son útiles solamente un breve periodo de tiempo, debido a lo cual no se incluyeron.

A partir de la colección de documentos pre-procesada se formó una lista denominada, *Lista Básica de Artículo* (*LBA*), compuesta por todos los artículos de la colección de textos normativos, mismos que se enumeraron comenzando desde 1, hasta el número total de artículos que fueron 1698.

El contenido de la LBA se lematizó de acuerdo con el esquema utilizado por Calvo *et. al.* [18], de esta manera se generó una segunda lista llamada *Lista Lematizada de Artículos (LLA).*

Las listas LBA y LLA, se representaron como vectores, y fue obtenido su valor de *Term Frequency Inverse Document Frequency* (*tf·idf*) a partir de las ecuaciones (**6**), (**7**) y (**8**). Cabe recordar que en tales ecuaciones se considera a los artículos como documentos.

$$tfidf = tf_{t,j} * idf_i \qquad (\mathbf{6})$$

$$tf_{t,j} = \frac{n_{i,j}}{\sum n_{k,j}} \qquad (\mathbf{7})$$

Donde: $n_{i,j}$ indica el número de ocurrencias del término considerado en el artículo a_j y

El denominador representa la ocurrencia de todos los términos en el artículo a_j.

$$idf_i = \log\left(\frac{|A|}{|\{a_j : t_i \in a_j\}|}\right) \quad \textbf{(8)}$$

Donde: |A| es el número total de artículos en la colección de documentos y

$|\{a_j : t_i \in a_j\}|$ es el número de artículos donde aparece el término t_i.

En la tabla 1, se presentan ejemplos de los valores de *tfidf*, *tf* y *idf* de algunos los vectores obtenidos a partir de las listas *LBA* y *LLA*.

Finalmente, se construyó un grafo por cada lista de artículos (en el que como se recordará, cada nodo representa un artículo de los textos normativos) y los valores asociados a las aristas representan el valor inverso de la medida de semejanza estándar coseno. El grafo correspondiente a *LBA* se denomino *GBP*, mientras que el grafo correspondiente a *LLA* se llamó *GLE*.

También se investigó el efecto de eliminar palabras sin contenido de la colección de documentos, para lo cual se crearon dos nuevas listas de artículos, resultado de eliminar la lista de palabras sin contenido, que se encuentra en el apéndice B, de las listas *LBA* y *LLA*. Las nuevas listas creadas a partir de tanto *LBA* como *LLA* se denominaron *LBA_R* y *LLA_R*, respectivamente, para indicar que se *removieron* las palabras sin contenido.

La lista de palabras sin contenido fue propuesta basándose en la literatura [14] y se comprobó de forma manual que los términos elegidos tuvieran valores de ponderación de valor muy bajo (<0.01, dado que el valor máximo que un término puede obtener es ~3.22). Algunas de las palabras que se encuentran en dicha lista son: *el*, *la*, *los*, *de*, mismas que, como se observa en la Tabla 1, tienen asociados valores de ponderación mucho menores al valor considerado. El problema de determinar la lista de forma automática estriba en que los valores de ponderación de los términos varían de artículo en artículo, y en algunos artículos los valores asociados a los términos de la lista son mayores al valor de 0.01; sin embargo, en general para la mayoría de los artículos los términos elegidos resultan con valores menores al del umbral considerado.

De igual forma, como se procedió con la lista básica de artículos (LBA) y la lista lematizada de artículos (LLA) se construyeron un par de grafos de artículos a partir de las listas LBA_R y LLA_R. En la Tabla 2 se muestran ejemplos de los vectores a partir de los cuales fueron elaborados tales grafos, éstos últimos denominados GBP_R y GLE_R.

Tabla 1. Ejemplos de los vectores obtenidos a partir de las listas LBA y LLA

Vectores obtenidos a partir de LBA

Artículo 1 # términos: 39

Término	*tfi*	*dfi*	***tfidf***
impartan	2	16	***0.104***
conjunta	1	7	***0.061***
extranjero	1	17	***0.051***
país	1	31	***0.045***
aquellos	1	39	***0.042***
educativas	1	42	***0.041***
…			
con	1	726	***0.009***
por	1	815	***0.008***
del	2	1214	***0.007***
las	1	1043	***0.005***
el	2	1448	***0.004***
en	1	1259	***0.003***
los	1	1292	***0.003***
de	4	1624	***0.002***
la	1	1367	***0.002***

Artículo 2 # términos: 33

Término	*tfi*	*dfi*	***tfidf***
aplicación	2	89	***0.078***
vigilancia	1	10	***0.068***
obligatoria	1	11	***0.066***
interpretación	1	17	***0.061***
emitirá	1	20	***0.058***
observancia	1	22	***0.057***
…			
reglamento	1	443	***0.018***
su	1	652	***0.013***
las	2	1043	***0.013***
del	3	1214	***0.013***
para	1	765	***0.010***
la	2	1367	***0.006***
en	1	1259	***0.004***
el	1	1448	***0.002***
de	2	1624	***0.001***

Vectores obtenidos a partir de LLA

Artículo 1 # términos: 30

Término	*tfi*	*dfi*	***tfidf***
impartir	2	72	***0.092***
conjunta	1	7	***0.079***
extranjero	1	17	***0.067***
país	1	31	***0.058***
educativas	1	42	***0.054***
instituciones	1	56	***0.049***
…			
estudio	1	354	***0.023***
como	1	414	***0.02***
reglamento	1	457	***0.019***
instituto	1	548	***0.016***
para	1	765	***0.012***
con	1	726	***0.012***
por	1	815	***0.011***
en	1	1259	***0.004***
de	6	1669	***0.001***

Artículo 2 # términos: 28

Término	*tfi*	*dfi*	***tfidf***
aplicación	2	89	***0.091***
vigilancia	1	10	***0.08***
obligatoria	1	11	***0.078***
interpretación	1	17	***0.071***
observancia	1	22	***0.067***
necesarios	1	45	***0.056***
…			
nacional	1	251	***0.03***
presente	1	286	***0.028***
investigación	1	321	***0.026***
el	1	313	***0.026***
reglamento	1	457	***0.02***
ser	1	736	***0.013***
para	1	765	***0.012***
en	1	1259	***0.005***
de	5	1669	***0.001***

Tabla 2. Ejemplos de los vectores obtenidos a partir de las listas LBA_R y LLA_R

Vectores obtenidos a partir de LBA_R

Artículo 1 # términos: 25

Término	*tfi*	*dfi*	***tfidf***
impartan	2	16	***0.162***
conjunta	1	7	***0.095***
extranjero	1	17	***0.080***
país	1	31	***0.070***
educativas	1	42	***0.064***
establecer	1	50	***0.061***
…	1		
coordinación	1	169	***0.040***
politécnico	1	184	***0.039***
desarrollo	1	219	***0.036***
nacional	1	251	***0.033***
así	1	262	***0.032***
estudios	1	284	***0.031***
presente	1	294	***0.030***
reglamento	1	443	***0.023***
instituto	1	548	***0.020***

Artículo 2 # términos: 25

Término	*tfi*	*dfi*	***tfidf***
aplicación	2	89	***0.102***
vigilancia	1	10	***0.089***
obligatoria	1	11	***0.088***
interpretación	1	17	***0.080***
emitirá	1	20	***0.077***
observancia	1	22	***0.076***
…			
secretaría	1	140	***0.043***
cumplimiento	1	168	***0.04***
instituto	2	548	***0.039***
politécnico	1	184	***0.039***
programas	1	249	***0.033***
nacional	1	251	***0.033***
presente	1	294	***0.030***
investigación	1	321	***0.029***
reglamento	1	443	***0.023***

Vectores obtenidos a partir de LLA_R

Artículo 1 # términos: 23

Término	*tfi*	*dfi*	***tfidf***
impartir	2	72	***0.119***
conjunta	1	7	***0.104***
extranjero	1	17	***0.087***
país	1	31	***0.076***
educativas	1	42	***0.07***
instituciones	1	56	***0.064***
…			
establecer	1	175	***0.043***
politécnico	1	183	***0.042***
desarrollo	1	221	***0.039***
nacional	1	251	***0.036***
así	1	261	***0.035***
presente	1	286	***0.034***
estudio	1	354	***0.03***
reglamento	1	457	***0.025***
instituto	1	548	***0.021***

Artículo 2 # términos: 23

Término	*tfi*	*dfi*	***tfidf***
aplicación	2	89	***0.111***
vigilancia	1	10	***0.097***
obligatoria	1	11	***0.095***
interpretación	1	17	***0.087***
observancia	1	22	***0.082***
necesarios	1	45	***0.069***
…			
cumplimiento	1	168	***0.044***
instituto	2	548	***0.043***
politécnico	1	183	***0.042***
corresponder	1	183	***0.042***
programas	1	249	***0.036***
nacional	1	251	***0.036***
presente	1	286	***0.034***
investigación	1	321	***0.031***
reglamento	1	457	***0.025***

Resumiendo, para los experimentos sobre la colección de documentos, se crearon cuatro grafos, GBP, GLE, GBP_R y GLE_R, sobre los cuales se llevaron a cabo los experimentos realizados sobre el formato de la pregunta.

5.1.2 Pregunta

El procedimiento que se realizó en el módulo construcción solicitud, es similar al que se empleó en la construcción del grafo, es decir, recibida una pregunta, ésta se sometió al proceso descrito en los puntos 1 a 4, para construir la "pre –solicitud". El paso 5, está relacionado con la forma en la cual se realiza la división, a partir de la cual se forman los nodos A y B.

1. *El texto de la pregunta se convierte a minúsculas.*
2. *Se eliminan de la pregunta, signos de puntuación (entre ellos los signos de interrogación de las preguntas), paréntesis, numeraciones romanas, números, guiones, letras de la "a" a la "z", si los contuviere.*
3. *Eliminación de palabras sin contenido*. El objetivo de este paso, fue investigar el efecto en el proceso de búsqueda de respuestas, al eliminar palabras sin contenido sólo en la solicitud. Así, este paso se aplicó a la solicitud, en función de las características del grafo usado, es decir, la pregunta se lematizó o no y se conservaron o eliminaron las palabras sin contenido. De acuerdo con lo anterior, la eliminación se realizó en los *SBR* basados en los grafos GLE y GBP.
4. *Expansión de términos*. La finalidad de este paso fue la de expandir el alcance de la búsqueda, (que en un principio estaría acotada a las palabras expresadas en la pregunta) al incluir términos que no fueron introducidos por el usuario. La expansión consiste en agregar palabras similares y/o relacionadas con las palabras de la pre- – solicitud. Los experimentos fueron llevados a cabo solamente en los *SBR* basados en los grafos que utilizan la *Lista Lematizada de Artículos*, SGLE y SGLE_R.

   Para los experimentos de expansión se utilizaron dos herramientas:

   a. Tesauro distribucional construido a partir de la enciclopedia Encarta 2004 [19, 20].

b. Tesauro basado en el diccionario Anaya [21] emplea solamente los sinónimos de las palabras que contiene el diccionario.

5. *División de la pre-solicitud*. La división es uno de los parámetros más significativos, ya que de la forma en que fue dividida la pregunta depende la recuperación de los artículos, por esta razón se investigaron 2 tipos de división para saber cuál fue la más óptima en nuestros experimentos.

   a. División a la mitad (DM): La pre-solicitud se divide a la mitad y con cada parte se forman los vectores-solicitud, representados en el grafo de la Figura 14 como los vértices A y B. Si el número de términos de la solicitud es impar, el vector correspondiente al vértice A contendrá una palabra más.

   b. División mezcla de términos (DMT): ésta consistió en formar el vector-solicitud correspondiente al vértice A, a partir de los términos impares de la pre-solicitud y el correspondiente al vértice B con los términos pares. Con el objetivo de investigar el efecto en la extracción de artículos, de ir del nodo B al nodo A, se realizaron otro par de experimentos que consistieron en intercambiar el contenido de los nodos A y B de las divisiones anteriores, debido a lo cual se denominaron DM' y DMT'.

Para ilustrar el proceso anterior, a continuación se muestra un ejemplo en el que se considera el *SBR* basado en el grafo GLE (formado a partir de la lista de artículos lematizada y conservando palabras sin contenido), por lo que la solicitud se lematizó y se conservaron las palabras sin contenido.

Considérese la siguiente pregunta:

¿Cómo se lleva a cabo el procedimiento de elección de representantes alumnos ante el consejo técnico consultivo escolar?

Después de ejecutar los pasos 1-3, se obtiene:

llevar a cabo procedimiento de elección de representantes alumnos ante consejo técnico consultivo escolar

Aplicando la expansión de términos usando el tesauro del diccionario Anaya, paso 4b, se tiene:

llevar transportar acarrear dirigir guiar manejar usar ponerse gastar exceder adelantar sobrepasar cuidar aguantar sobrellevar tolerar sufrir arrancar cercenar conseguir obtener a cabo punta terminación fin soga cordel cordón extremo procedimiento de elección votación sufragio de representantes alumnos ante consejo técnico consultivo escolar

Finalmente, aplicando los dos tipos de división explicadas en el paso 5, se obtienen el contenido de los nodos A y B, de la siguiente manera:

5a

Nodo A. llevar dirigir manejar usar exceder adelantar cuidar sufrir conseguir obtener cabo punta terminación

Nodo B. fin procedimiento de elección votación de representantes alumnos ante consejo técnico consultivo escolar

5b

Nodo A. llevar manejar exceder cuidar conseguir cabo terminación procedimiento elección de alumnos consejo consultivo

Nodo B. dirigir usar adelantar sufrir obtener punta fin de votación representantes ante técnico escolar

5.2 Metodología de evaluación del *SBR*

La metodología de evaluación estuvo constituida por dos etapas; la primera etapa consistió en hacer una comparación de las calificaciones obtenidas en todos los experimentos realizados, para conocer cuál fue la trascendencia o el impacto de cada uno de los parámetros variados, para conseguir la mayor semejanza al conjunto de las 21 respuestas proporcionadas por el Abogado General del Instituto Politécnico Nacional (IPN). Por otro lado, también fue deseable comparar nuestro método con algún otro; sin embargo, en la literatura no se encontró un trabajo con el cual pudiéramos compararnos. Es por ello que se decidió implementar un sistema para la misma tarea basado solamente en el modelo de espacio vectorial, el cual ha sido empleado en *SBR* [32, 33]. Esto último debido a que nuestro método se basa parcialmente en el modelo de espacio vectorial (MEV) y además una de las hipótesis del trabajo es que empleando el grafo (parte fun-

damental de nuestra propuesta), en lugar de un espacio vectorial podría ser posible obtener mejores resultados, al menos para la presente tarea.

5.2.1 Evaluación de los experimentos

Para verificar tanto el desempeño del *SBR* propuesto en la tarea de encontrar los artículos respuesta, como para comparar los resultados de los experimentos llevados a cabo, se propuso el siguiente criterio de evaluación, basado en las respuestas del conjunto de prueba. El criterio de evaluación propuesto es el siguiente:

> *Relevancia: La salida del SBR debe de dar respuesta a la preguntas, es decir, es deseable que el sistema devuelva, por lo menos aquellos artículos que el abogado general proporciona como respuesta a la misma pregunta.*

Como *SBR* prototipo se decidió que como primera aproximación se regresara un conjunto de artículos y evaluar que entre éstos se encontrara la respuesta a la pregunta introducida, considerando también la posición en la que son retornados. De esta forma, el procedimiento se diseñó para que en caso que los artículos respuesta se encuentren en los primeros lugares se asigne la máxima calificación, no importando el orden en que el sistema los devuelva. Por ejemplo, la pregunta 1 tiene como respuesta 5 artículos, si el sistema devuelve los cinco artículos en las primeras 5 posiciones a la respuesta se le asigna una calificación de 1 no importando el orden en que aparezcan. En caso contrario, la calificación disminuye si se encuentran en las siguientes cinco posiciones. El criterio de evaluación descrito se refleja en el siguiente procedimiento:

Se decidió que el conjunto de resultados devuelto por el *SBR* se limitará a 100 artículos (aproximadamente el 5.8% del total de artículos que el sistema maneja). El grupo de artículos regresados en la respuesta se dividió en paquetes de 5 (el número máximo de artículos que se pueden encontrar de respuesta), y a cada paquete se le asignó el valor, V_i, de acuerdo con la siguiente expresión:

$$V_i = 1 - \frac{(n-1)}{N} \qquad (9)$$

Donde: n es el número de paquete de artículos respuesta

N es el máximo número de paquetes que se pueden encontrar (20)

Finalmente, a cada respuesta de una pregunta determinada se le califica a través de la siguiente expresión:

$$Calificación \quad CR_i = \frac{\sum_{i=1}^{n} nV_i}{n_{AR}} \tag{10}$$

Donde: CR_i es la calificación asignada a la respuesta de la i-ésima pregunta,

n es el número de paquete donde se encuentra el artículo respuesta,

V_i es el valor definido anteriormente y

n_{AR} es el número de artículo respuesta encontrados.

Debido a que cada experimento se realizó a partir de las 21 preguntas seleccionadas, la calificación dada a cada uno resulta del promedio de la calificación, CR_i, es decir:

$$Calificación \quad CS = \frac{\sum_{i=1}^{np} CR_i}{np} \tag{11}$$

Donde: np es el número total de preguntas del sistema (21)

Para ilustrar la metodología anterior, considérese el siguiente ejemplo:

Se describe de manera detallada la evaluación del SBR empleando el grafo construido a partir de la lista básica de artículos (sin lematizar), conservando palabras sin contenido en la solicitud y realizando el tipo de división a la mitad. Considere nuevamente la siguiente pregunta:

¿Cómo se lleva a cabo el procedimiento de elección de representantes alumnos ante el Consejo Técnico Consultivo Escolar?

Pre-solicitud construida.

se lleva cabo el procedimiento de elección de representantes alumnos ante el consejo técnico consultivo escolar

Después de realizar la división sobre la pre-solicitud se obtiene:

Nodo A: se lleva cabo el procedimiento de elección de

Nodo B: representantes alumnos ante el consejo técnico consultivo escolar

A continuación se muestra los resultados de la evaluación para la pregunta de ejemplo:

Respuesta dada por el *SBR*.

Resultados de la búsqueda: Artículos encontrados 4 de 5

No. de Artículo	Documento	Posición en la cual el *SBR* retorna el artículo
A209	R Interno	1
A206	R Interno	5
A207	R Interno	10
A28	Ley Orgánica	30

Número de artículos encontrados en cada paquete:

Artículos en paquete 1: 2

Artículos en paquete 2: 1

Artículos en paquete 6: 1

Calificación de la pregunta, aplicando la ecuación (8), CS = 0.74

Como se observa del ejemplo anterior, el esquema de evaluación propuesto toma en cuenta no sólo los artículos encontrados por el sistema sino también la posición en la que son recuperados. Al ser un primer prototipo de un *SBR* se consideró que un sistema que fuera capaz de al menos retornar los artículos que se considera responden la pregunta sería de utilidad, ya que trabajos posteriores se enfocarían solamente a lograr que los resultados se acerquen a las primeras posiciones. Por otra parte, por razones de espacio no se incluyen el texto de todos los artículos retornados por el sistema; sin embargo, es importante mencionar que los artículos devueltos en las primeras posiciones no son totalmente ajenos a la solicitud — también agregan información que podría ser de utilidad, aunque el Abogado General no los considera. Por lo regular, él se enfoca a dar respuestas directas y resumidas, y en algunos casos es conveniente contar con mayor información. En la sección 5.3 se ofrece un ejemplo de los resultados devueltos por el sistema a una pregunta específica.

De acuerdo con las variables establecidas, se realizaron 48 experimentos en el *SBR* propuesto, así que con el propósito de organizar los tipos de experimentos, se construyó la Tabla 3, en donde se encuentran los nombres y la descripción de cada experimento.

Tabla 3. Nombre y descripción de los experimentos realizados para el *SBR* propuesto.

Lista de palabras		Solicitud		Tipo de división		Cambio en el contenido de los nodos A y B	
Palabras sin contenido		Palabras sin contenido					
Conservando	Removiendo	Conservando	Removiendo	DM	DMT	DM´	DMT´
LBA		✓		✓	✓	✓	✓
			✓	✓	✓	✓	✓
	LBA_R		✓	✓	✓	✓	✓
LLA		✓		✓	✓	✓	✓
			✓	✓	✓	✓	✓
	LLA_R		✓	✓	✓	✓	✓
Aplicando el tesauro distribucional							
LLA		✓		✓	✓	✓	✓
			✓	✓	✓	✓	✓
	LLA_R		✓	✓	✓	✓	✓
Aplicando el tesauro basado en el diccionario Anaya							
LLA		✓		✓	✓	✓	✓
			✓	✓	✓	✓	✓
	LLA_R		✓	✓	✓	✓	✓

Total de experimentos realizados: 48

5.2.2 Experimentos realizados en el sistema basado en el MEV

En la sección 3.2 se mencionó que en el MEV cada documento de la colección es representado por un vector de términos extraídos de los documentos. En este caso se consideró el mismo conjunto de vectores generados a partir de las listas descritas en la sección 5.1.1. Dada una pregunta, ésta se modela nuevamente como una lista de términos con pesos asociados que representan su importancia en la solicitud. La solicitud fue procesada de acuerdo con los pasos 1 a 3 de la sección 5.2.

La extracción de respuestas en el MEV está dada a partir de la medida de semejanza entre los vectores documento y el vector solicitud (sección 3.2, ecuación 8). De esta forma, sólo se regresan los artículos más relevantes de la solicitud, mismos que al igual que el sistema de búsqueda de respuestas se limitaron a cien artículos. Los experimentos usando el sistema basado en el modelo de espacio vectorial, están organizados en la Tabla 4, en donde se pueden observar el nombre y las características de cada uno de ellos.

Tabla 4. Nombre y descripción de los experimentos basados en el MEV.

Lista de palabras		Solicitud	
Palabras sin contenido		Palabras sin contenido	
Conservando	Removiendo	Conservando	Removiendo
LBA		✓	
	LBA_R		✓
LLA		✓	
	LLA_R		✓
Aplicando el tesauro distribucional			
LLA		✓	
	LLA_R		✓
Aplicando el tesauro Anaya			
LLA		✓	
	LLA_R		✓

Total de experimentos realizados: 8

Como se observa de la anterior tabla, los experimentos realizados con el sistema basado en el MEV son similares a los realizados con el *SBR* con la excepción de que no se aplica la división (debido a que ésta es solo característica del grafo).

5.3 Resultados

De acuerdo con lo establecido en las secciones 5.1 y 5.2 el número total de experimentos realizados fue de 56, así que los principales valores de los resultados de estos experimentos, se encuentran registrados en Tabla 5 a Tabla 8.

En la tabla 5, están colectados los resultados sin lematizar la lista de artículos y los resultados obtenidos de los experimentos al agregar herramientas de procesamiento de lenguaje natural al sistema básico propuesto. En la Tabla 6 se reportan los resultados obtenidos al aplicar lematización a la lista de artículos básica y al conjunto de preguntas, finalmente en la Tabla 7 y Tabla 8 se muestran los resultados al aplicar el tesauro distribucional y el basado en el diccionario Anaya, respectivamente.

Las columnas relacionadas con las respuestas encontradas indican el grado de alcance que tiene el sistema al contestar las preguntas introducidas. Así, si el sistema retornó todos los artículos que el Abogado General consideró para su respuesta, entonces se considera que la respuesta del sistema está dentro del rubro "completo". Si la respuesta sólo contuvo algunos de los artículos de la respuesta del Abogado General, entonces queda dentro clasificada "Parcial", y finalmente, si el sistema no devolvió ningún artículo, entonces se considera que la respuesta no fue contestada.

Tabla 5. Resultados de los experimentos correspondientes a las listas de artículos sin lematizar.

Lista de artículos		Solicitud		Tipo de división	Calificación promedio	Respuestas encontradas		
Palabras sin contenido								
Conserv.	Remov.	Conserv.	Remov.			Completo	Parcial	Sin resp.
SBR propuesto								
LBA		✓		DM	0.5342	12	5	4
				DMT	0.4988	13	3	5
				DM´	**0.5351**	**12**	**5**	**4**
				DMT´	0.5023	13	3	5
			✓	DM	0.4851	10	5	6
				DMT	0.4865	10	6	5
				DM´	0.4858	9	6	6
				DMT´	0.4889	10	6	5
	LBA_R		✓	DM	0.4603	9	6	6
				DMT	0.4723	10	5	6
				DM´	0.4683	10	5	6
				DMT´	0.4716	10	5	6
Sistema basado en el MEV								
LBA		✓		-	**0.2253**	**3**	**5**	**13**
	LBA_R		✓	-	0.1892	4	6	11

Tabla 6. Resultados de los experimentos correspondientes a las listas de artículos lematizadas.

Lista de artículos		Solicitud		Tipo de división	Calificación promedio	Respuestas encontradas		
Palabras sin contenido								
Conserv.	Remov.	Conserv.	Remov.			Completo	Parcial	Sin resp.
SBR propuesto								
LLA		✓		DM	0.5170	11	8	2
				DMT	0.5536	12	8	1
				DM´	0.5175	12	7	2
				DMT´	0.5548	12	8	1
			✓	DM	0.5187	12	7	2
				DMT	0.6151	12	8	1
				DM´	0.5175	12	7	2
				DMT´	0.5548	12	8	1
	LLA_R		✓	DM	0.5026	13	6	2
				DMT	0.6103	13	7	1
				DM´	0.5115	13	6	2
				DMT´	**0.6230**	**13**	**7**	**1**
Sistema basado en el MEV								
LLA		✓		-	0.1976	2	5	14
	LLA_R		✓	**-**	**0.3055**	**5**	**7**	**9**

Tabla 7. Resultados de los experimentos correspondientes a las listas de artículos lematizadas, a las que se les aplicó el tesauro Anaya.

Lista de artículos		Solicitud		Tipo de división	Calificación promedio	Respuestas encontradas		
Palabras sin contenido								
Conserv.	Remov.	Conserv.	Remov.			Completo	Parcial	Sin resp.
SBR propuesto								
LLA		✓		DM	0.4689	12	6	3
				DMT	**0.5461**	**13**	**6**	**2**
				DM´	0.4719	11	7	3
				DMT´	0.5456	13	6	2
			✓	DM	0.4883	13	6	2
				DMT	0.5257	11	6	4
				DM´	0.4856	12	7	2
				DMT´	0.5276	11	6	4
	LLA_R		✓	DM	0.4736	12	7	2
				DMT	0.5325	12	6	3
				DM´	0.4783	12	7	2
				DMT´	0.5336	12	6	3
Sistema basado en el MEV								
LLA		✓		-	0.2428	4	4	13
	LLA_R		✓	-	**0.2815**	**4**	**7**	**10**

Tabla 8. Resultados de los experimentos correspondientes a las listas de artículos lematizadas, a las que les se aplicó el tesauro distribucional.

Lista de artículos		Solicitud		Tipo de división	Calificación promedio	Respuestas encontradas		
Palabras sin contenido								
Conserv.	Remov.	Conserv.	Remov.			Completo	Parcial	Sin resp.
SBR propuesto								
LLA		✓		DM	0.5228	12	6	3
				DMT	0.4933	11	6	4
				DM´	0.5253	12	6	3
				DMT´	0.4905	11	6	4
			✓	DM	0.5250	12	6	3
				DMT	0.5138	12	5	4
				DM´	**0.5273**	**12**	**6**	**3**
				DMT´	0.5126	12	5	4
	LLA_R		✓	DM	0.5167	12	5	4
				DMT	0.5150	12	6	3
				DM´	0.5159	12	5	4
				DMT´	0.5205	12	6	3
Sistema basado en el MEV								
LLA		✓		-	0.1869	3	3	15
	LLA_R		✓	-	**0.2107**	**4**	**4**	**13**

Como se mencionó anteriormente, el número de experimentos realizados fue de 56, cuyos resultados finales están mostrados en las tablas anteriores. Con la finalidad de mostrar en detalle los resultados más representativos, en la Tabla 9 se muestran los registros de las 21 preguntas, para la mejor y la peor calificación observada en nuestros experimentos, los cuales corresponden a la lista de artículos lematizada, removiendo palabras sin contenido, tanto en la lista de artículos como en la solicitud, con división DTM' sin aplicar algún tesauro y a la lista lematizada de artículos conservando palabras sin contenido en la lista de artículos y removiéndolas de la solicitud, con un tipo de división DM', con la aplicación del tesauro distribucional, respectivamente. En la Tabla 10, se muestran también el detalle de la mejor y la peor calificación empleando el sistema basado en el MEV. A continuación se muestran las 21 preguntas empleadas para la evaluación así como las respuestas compuestas solamente por los artículos dados por el Abogado General del IPN.

Conjunto de preguntas/respuestas empleadas en la evaluación.

1. ¿Cómo se lleva a cabo el procedimiento de elección de representantes alumnos ante el Consejo Técnico Consultivo Escolar?
Artículo 28 Ley Orgánica
Artículos 206, 207, 209 y 213 Reglamento Interno

2. ¿Cuáles son los consejeros que deben abstenerse de participar en procesos de elección de terna para la designación de Director?
Artículo 181 del Reglamento Interno

3. ¿Es procedente que un alumno solicite mención honorífica a nivel licenciatura si escogió su Titulación por la opción de Escolaridad?
Artículo 13 y 43 del Reglamento de Titulación Profesional

4. ¿Están facultados para actuar como asesores y/o sinodales en exámenes profesionales de licenciatura los profesores de nivel superior que han sido asesores de tesis de maestría y doctorado a pesar de que no poseen titulo profesional de licenciatura, pero cuentan con documentación probatoria de haber cursado estudios de posgrado en instituciones públicas nacionales como el CINVESTAV, o bien que estudiaron a nivel licenciatura en el extranjero y sus títulos no están registrados ante la Dirección General de Profesiones.
Artículos 24 y 33 Reglamento de Titulación Profesional

5. ¿La esposa del actual director puede ser candidata a dicho cargo?
Artículo 168 reglamento Interno del IPN
Artículo 14, fracción XVII, de la Ley Orgánica
Artículo 21 del mismo ordenamiento.

6. ¿La Oficina del Abogado General o cualquier otra autoridad del IPN puede pronunciarse sobre la constitución, estructuración, organización y funcionamiento de las sociedades, asociaciones o agrupaciones de alumnos y/o egresados?
Artículo 4º. Ley Orgánica del IPN

7. ¿Los Colegios de Profesores o los Consejos Técnicos Consultivos Escolares están facultados para calificar sí los aspirantes cumplen o no los requisitos para ser director de una escuela, centro o unidad de enseñanza y de investigación?
Artículo 14 y 29 Ley orgánica
Artículo 264 Reglamento Interno

8. ¿Procede integrar consejo técnico consultivo escolar en un centro de investigación a efecto de que lleve a cabo proceso de elección de terna para designación de subdirectores a pesar de que el centro sólo sea de investigación, no así de enseñanza?
Artículos 27 y 28 de la Ley Orgánica
Artículos 202 y 264 del Reglamento Interno

9. ¿Procede que un mismo aspirante sea propuesto para integrar más de una terna en el proceso de elección de ternas para la designación de subdirectores técnico, administrativo y académico?
Artículos 174 y 175 del Reglamento Interno.

10. ¿Procede que una persona que en dos ocasiones ha sido subdirector participe en el proceso de elección de ternas para la designación de subdirectores técnico, administrativo y académico y en su caso ejerza otra vez ese cargo.
Artículos 21 de la Ley Orgánica

Artículo 182 Reglamento Interno

11. ¿Puede un mismo individuo por segundo año consecutivo ser electo consejero representante del personal académico?
Artículo 215 Reglamento interno

12. ¿Pueden prestar el servicio social alumnos que hayan acreditado el 70% de créditos pero adeuden materias?
Artículo 2 Reglamento servicio social

13. ¿Pueden votar los profesores del colegio que se encuentran en -año sabático-?
Artículo 80 Reglamento de las Condiciones Interiores de Trabajo del Personal Académico del IPN

14. ¿Qué debe entenderse por tener estudios de posgrado?
Artículo 4 del Reglamento de Estudios de Posgrado:

15. ¿Respecto del perfil de Subdirector de Investigación Aplicada, procede incluir en la convocatoria para ocupar el cargo la mención ?poseer estudios de posgrado? como requisito indispensable, por analogía con las subdirecciones Académica y Científica?
Artículos 174 y 175 Reglamento Interno

16. ¿Tienen derecho de votar o no los profesores interinos adscritos a alguna escuela?
Artículo 208 del Reglamento Interno del IPN
Artículo 6 del Reglamento de las Condiciones Interiores de Trabajo del Personal Académico

17. ¿Un alumno que adeuda cinco materias puede seguir fungiendo como consejero representante de los alumnos?
Artículo 81 Reglamento Interno
Artículo 44 del Reglamento de Estudios Escolarizados para los Niveles Medio Superior y Superior

18. ¿Un consejero que ha faltado a varias sesiones del Consejo Técnico Consultivo Escolar puede participar en el proceso de elección de terna?
Artículo 200 del Reglamento Interno

19. ¿Un Maestro Decano en período de prejubilación puede presidir el Consejo Técnico Consultivo Escolar o el Colegio de Profesores para dirigir el proceso de elección de ternas para la designación de director?
Artículo 20 y 23 Reglamento del Decanato

20. ¿Una persona puede ser director por tercera ocasión?
Artículo 21 de la Ley Orgánica del IPN

21. ¿Cuáles son los requisitos que debe cumplir un director de escuela, centro o unidad de enseñanza y de investigación, para designar a los jefes de división o de departamento a su cargo?
Artículo 173 del Reglamento Interno

Tabla 9. Calificaciones por pregunta de 2 variaciones representativas del *SBR*

	LLA_R ♦. Lista lematizada de artículos, removiendo palabras sin contenido. Removiendo palabras con contenido en solicitud. División mezclada DMT'. Ruta de B a A. Sin emplear tesauros.		LLA ♣. Lista lematizada de artículos, conservando palabras sin contenido. Removiendo palabras sin contenido en solicitud División a la mitad DM'. Ruta de B a A. Utilizando Tesauro distribucional	
#	Artículos encontrados	Calif	Artículos encontrados	Calif
1	Artículos encontrados 4 de 5 A209 R Interno Posición 1 A206 R Interno Posición 2 A207 R Interno Posición 36 A28 Ley Orgánica Posición 44 Artículos en paquete 1: 2 Artículos en paquete 8: 1 Artículos en paquete 9: 1	0.65	Artículos encontrados 3 de 5 A206 R Interno Posición 1 A209 R Interno Posición 5 A207 R Interno Posición 8 Artículos en paquete 1: 2 Artículos en paquete 2: 1	0.59
2	Artículos encontrados 1 de 1 A181 R Interno Posición 5 Artículos en paquete 1: 1	1	Artículos encontrados 1 de 1 A181 R Interno Posición 3 Artículos en paquete 1: 1	1
3	Artículos encontrados 2 de 2 A43 Rtitulación Posición 1 A13 Rtitulación Posición 5 Artículos en paquete 1: 2	1	Artículos encontrados 2 de 2 A13 Rtitulación Posición 1 A43 Rtitulación Posición 2 Artículos en paquete 1: 2	1
4	Artículos encontrados 1 de 2 A33 Rtitulación Posición 6 Artículos en paquete 2: 1	0.475	Artículos encontrados 1 de 2 A33 Rtitulación Posición 33 Artículos en paquete 7: 1	0.35
5	Artículos encontrados 2 de 3 A168 R Interno Posición 16 A21 Ley Orgánica Posición 54 Artículos en paquete 4: 1 Artículos en paquete 11: 1	0.45	Artículos encontrados 1 de 3 A168 R Interno Posición 96 Artículos en paquete 20: 1	0.016
6	Artículos encontrados 1 de 1 A4 Ley Orgánica Posición 22 Artículos en paquete 5: 1	0.8	Artículos encontrados 1 de 1 A4 Ley Orgánica Posición 53 Artículos en paquete 11: 1	0.5
7	Artículos encontrados 2 de 3 A264 R Interno Posición 8 A29 Ley Orgánica Posición 9 Artículos en paquete 2: 2	0.633	Artículos encontrados 2 de 3 A29 Ley Orgánica Posición 13 A14 Ley Orgánica Posición 82 Artículos en paquete 3: 1 Artículos en paquete 17: 1	0.366
8	Artículos encontrados 2 de 4 A264 R Interno Posición 10 A28 Ley Orgánica Posición 56 Artículos en paquete 2: 1 Artículos en paquete 12: 1	0.35	Artículos encontrados 1 de 4 A264 R Interno Posición 24 Artículos en paquete 5: 1	0.2
9	Artículos encontrados 2 de 2 A175 R Interno Posición 12 A174 R Interno Posición 38 Artículos en paquete 3: 1 Artículos en paquete 8: 1	0.775	Artículos encontrados 2 de 2 A175 R Interno Posición 4 A174 R Interno Posición 58 Artículos en paquete 1: 1 Artículos en paquete 12: 1	0.725
10	Artículos encontrados 1 de 2 A182 R Interno Posición 78 Artículos en paquete 16: 1	0.125	Artículos encontrados 0 de 2	0
11	Artículos encontrados 1 de 1 A215 R Interno Posición 35 Artículos en paquete 7: 1	0.7	Artículos encontrados 1 de 1 A215 R Interno Posición 30 Artículos en paquete 6: 1	0.75

12	Artículos encontrados 1 de 1 A2 RServ.Soc. Posición 49 Artículos en paquete 10: 1	0.55	Artículos encontrados 1 de 1 A2 RServ.Soc. Posición 54 Artículos en paquete 11: 1	0.5
13	Artículos encontrados 1 de 1 A80 RCITPA Posición 67 Artículos en paquete 14: 1	0.35	Artículos encontrados 0 de 1	0
14	Artículos encontrados 1 de 1 A4 R.Posg. Posición 2 Artículos en paquete 1: 1	1	Artículos encontrados 1 de 1 A4 R.Posg. Posición 14 Artículos en paquete 3: 1	0.9
15	Artículos encontrados 2 de 2 A175 R Interno Posición 4 A174 R Interno Posición 69 Artículos en paquete 1: 1 Artículos en paquete 14: 1	0.675	Artículos encontrados 2 de 2 A175 R Interno Posición 5 A174 R Interno Posición 40 Artículos en paquete 1: 1 Artículos en paquete 8: 1	0.825
16	Artículos encontrados 1 de 2 A6 RCITPA Posición 9 Artículos en paquete 2: 1	0.475	Artículos encontrados 1 de 2 A6 RCITPA Posición 29 Artículos en paquete 6: 1	0.375
17	Artículos encontrados 2 de 2 A44 REst.Esc. Posición 2 A81 R Interno Posición 34 Artículos en paquete 1: 1 Artículos en paquete 7: 1	0.85	Artículos encontrados 2 de 2 A44 REst.Esc. Posición 40 A81 R Interno Posición 52 Artículos en paquete 8: 1 Artículos en paquete 11: 1	0.575
18	Artículos encontrados 1 de 1 A200 R Interno Posición 2 Artículos en paquete 1: 1	1	Artículos encontrados 1 de 1 A200 R Interno Posición 6 Artículos en paquete 2: 1	0.95
19	Artículos encontrados 2 de 2 A23 RDecanato Posición 42 A20 RDecanato Posición 70 Artículos en paquete 9: 1 Artículos en paquete 14: 1	0.475	Artículos encontrados 2 de 2 A20 RDecanato Posición 58 A23 RDecanato Posición 59 Artículos en paquete 12: 2	0.45
20	***Artículos encontrados 0 de 1***	***0***	***Artículos encontrados 0 de 1***	***0***
21	Artículos encontrados 1 de 1 A173 R Interno Posición 30 Artículos en paquete 6: 1	0.75	Artículos encontrados 1 de 1 A173 R Interno Posición 3 Artículos en paquete 1: 1	1
	EVALUACION FINAL: ***0.6230***		EVALUACION FINAL: ***0.5273***	

De la tabla anterior, se observa de manera general, que los valores de la calificación fueron ligeramente más altos para el experimento en donde no se aplicó el tesauro, sin embargo, la aplicación del tesauro distribucional implicó que los artículos respuesta fueran retornados en mejores posiciones.

A continuación se muestran los primeros cinco resultados que devuelve el *SBR* de mayor calificación a la pregunta 1.

1. ¿Cómo se lleva a cabo el procedimiento de elección de representantes alumnos ante el Consejo Técnico Consultivo Escolar?

Artículo 28 Ley Orgánica

Artículos 206, 207, 209 y 213 Reglamento Interno

Artículo 209. La elección de los representantes alumnos ante los consejos técnicos consultivos escolares de las escuelas, centros y unidades de enseñanza se sujetará a las siguientes bases:

I. Se integrará un comité de elección por cada programa académico de educación media superior, superior y de posgrado;

II. Se elegirá un representante alumno regular por cada grupo escolar para integrar los comités de elección previstos en los términos de la fracción anterior, y

III. Los miembros del comité elegirán, en forma directa y por mayoría de votos de entre ellos, a los representantes alumnos ante el Consejo Técnico Consultivo Escolar correspondiente, de conformidad con lo dispuesto por los artículos 27, fracciones VII y VIII, y 28, fracción IV, de la Ley Orgánica.

Artículo 187. La elección de los representantes tanto del personal académico como de los alumnos se llevará a cabo cada año en el mes inmediato posterior al inicio del ciclo escolar, en los términos de las respectivas convocatorias que el propio Consejo emita.

Artículo 189. La elección de los representantes alumnos se sujetará a las siguientes bases:

I. Se integrará un comité de elección con los consejeros alumnos de los consejos técnicos consultivos escolares de cada nivel educativo, en cada una de las siguientes ramas del conocimiento: Ciencias Sociales y Administrativas; Médico Biológicas y de Ingeniería y Ciencias Físico Matemáticas, y

II. Cada comité de elección elegirá, en forma directa y por mayoría de votos, a sus representantes.

Artículo 206. La elección de los representantes tanto del personal académico como de los alumnos ante los consejos técnicos consultivos escolares se llevará a cabo cada año, en el mes inmediato posterior al inicio del ciclo escolar, en los términos de las respectivas convocatorias que los propios consejos expidan.

Artículo 30.- El Director General, con acuerdo del Consejo General Consultivo establecerá las bases permanentes en el Reglamento Interno del Instituto Politécnico Nacional para la acreditación de los representantes profesores y representantes alumnos ante el propio Consejo General y los Consejos Técnicos Consultivos Escolares.

Tabla 10. Calificaciones por pregunta de 2 variaciones representativas del MEV

#	Artículos encontrados	Calif		Calif
1	Artículos encontrados 2 de 5 A206 R Interno Posición 1 A209 R Interno Posición 2 Artículos en paquete 1: 2	0.4	Artículos encontrados 2 de 5 A206 R Interno Posición 1 A209 R Interno Posición 2 Artículos en paquete 1: 2	0.4
2	Artículos encontrados 1 de 1 A181 R Interno Posición 4 Artículos en paquete 1: 1	1	Artículos encontrados 0 de 1	0
3	Artículos encontrados 1 de 2 A13 Rtitulación Posición 2 Artículos en paquete 1: 1	0.5	Artículos encontrados 1 de 2 A13 Rtitulación Posición 37 Artículos en paquete 8: 1	0.325
4	Artículos encontrados 0 de 2	0	Artículos encontrados 0 de 2	0
5	Artículos encontrados 1 de 3 A21 Ley Orgánica Posición 85 Artículos en paquete 17: 1	0.066	Artículos encontrados 0 de 3	0
6	Artículos encontrados 0 de 1	0	Artículos encontrados 0 de 1	0
7	Artículos encontrados 0 de 3	0	Artículos encontrados 0 de 3	0
8	Artículos encontrados 1 de 4 A264 R Interno Posición 5 Artículos en paquete 1: 1	0.25	Artículos encontrados 1 de 4 A264 R Interno Posición 3 Artículos en paquete 1: 1	0.25
9	Artículos encontrados 0 de 2	0	Artículos encontrados 0 de 2	0
10	Artículos encontrados 1 de 2 A21 Ley Orgánica Posición 12 Artículos en paquete 3: 1	0.45	Artículos encontrados 0 de 2	0
11	Artículos encontrados 1 de 1 A215 R Interno Posición 5 Artículos en paquete 1: 1	1	Artículos encontrados 0 de 1	0
12	Artículos encontrados 1 de 1 A2 RServ.Social Posición 98 Artículos en paquete 20: 1	0.05	Artículos encontrados 0 de 1	0
13	Artículos encontrados 0 de 1	0	Artículos encontrados 0 de 1	0
14	Artículos encontrados 0 de 1	0	Artículos encontrados 1 de 1 A4 R.Posgrado Posición 3 Artículos en paquete 1: 1	1
15	Artículos encontrados 1 de 2 A175 R Interno Posición 1 Artículos en paquete 1: 1	0.5	Artículos encontrados 0 de 2	0
16	Artículos encontrados 0 de 2	0	Artículos encontrados 0 de 2	0
17	Artículos encontrados 1 de 2 A44 REst.Escol. Posición 3 Artículos en paquete 1: 1	0.5	Artículos encontrados 0 de 2	0
18	Artículos encontrados 1 de 1 A200 R Interno Posición 1 Artículos en paquete 1: 1	1	Artículos encontrados 1 de 1 A200 R Interno Posición 1 Artículos en paquete 1: 1	1
19	Artículos encontrados 0 de 2	0	Artículos encontrados 0 de 2	0
20	Artículos encontrados 1 de 1 A21 Ley Orgánica Posición 32 Artículos en paquete 7: 1	0.7	Artículos encontrados 0 de 1	0
21	Artículos encontrados 0 de 1	0	Artículos encontrados 1 de 1 A173 R Interno Posición 6 Artículos en paquete 2: 1	0.95
EVALUACION FINAL: 0.3055			EVALUACION FINAL: 0.1869	

5.4 Descripción y análisis de resultados

En esta sección se presentan la descripción de los resultados de las tablas 5 a 8 y posteriormente, se realiza la discusión de resultados.

Tabla 11. Resumen de los mejores resultados de las tablas 5 - 8.

Lista de artículos		Solicitud		Tipo división	Calif. Promedio	Respuestas encontradas		
Palabras sin contenido								
Conserv.	Remov.	Conserv.	Remov.			Completa	Parcial	Sin. resp
SBR. Lista de artículos sin lematizar								
LBA		✓		DM´	0.5351	12	5	4
MEV								
LBA		✓		-	0.2253	3	5	13
SBR. Lista de artículos lematizada								
	LLA_R		✓	**DMT´**	**0.6230**	**13**	**7**	**1**
MEV								
	LLA_R		✓	**-**	**0.3055**	**5**	**7**	**9**
SBR. Lista de Artículos lematizada empleando diccionario Anaya								
LLA		✓		DMT	0.5461	13	6	2
MEV								
	LLA_R		✓	-	0.2815	4	7	10
SBR. Lista de Artículos lematizada empleando diccionario distribucional								
LLA			✓	DM´	0.5273	12	6	3
MEV								
	LLA_R		✓	-	0.2107	4	4	13

En la tabla 5 están registrados los resultados de los experimentos del *SBR*, cuya arquitectura está basada en el grafo, sin que se considere sin ningún medio que optimice la búsqueda, tal como el lematizador o el tesauro. También son mostrados los resultados de los experimentos obtenidos del sistema basado en el MEV, sin que tampoco se considere la inclusión de algún lematizador o tesauro.

De acuerdo con los valores de las calificaciones, el mejor resultado se encuentra para el experimento en donde la lista de artículos y la solicitud conservan las palabras sin contenido, para una división a la mitad, considerando que se camina sobre el grafo, del nodo B al nodo A. Este valor de calificación es de 0.5351. Se destaca que el valor de la calificación no cambia significativamente cuando se camina en sentido contrario sobre el grafo, es decir, del nodo A al nodo B. La calificación para este caso es de 0.5342. Esta observación puede ser generalizada para los demás experimentos. En el caso de la mejor respuesta, el *SBR* propuesto contestó 12 preguntas de manera completa, 5 preguntas de manera parcial y 4 preguntas no fueron contestadas, lo que en porciento equivale al 57, 23 y 20%, respectivamente. Para las mismas características de la lista de palabras y solicitud pero considerando una división de mezcla de términos, DMT, el *SBR* muestra una mejora, de una pregunta más, contestada completamente. Al respecto, debe considerarse, que el valor de la calificación está relacionado con la posición en la cual el *SBR* retorna los artículos, así entre más alta sea la calificación, implica que los artículos que dan respuesta a la pregunta aparecen en las primeras posiciones. Se destaca un compromiso entre la posición y el número de respuestas contestadas completamente. Los valores de las calificaciones, en los 8 experimentos, del *SBR* son muy próximos, hay una diferencia del 7% entre el valor más alto y el valor más bajo, presentados en la tabla. Lo que implicaría que al parecer el tipo de divisiones y la forma en que se camine en el grafo no son parámetros que impacten de manera significativa en la eficiencia para dar respuesta del *SBR*.

Por otro lado, las calificaciones del sistema basado en el MEV, están muy por debajo de las obtenidas en cualquier caso por el *SBR*, ya que para el mejor caso, la calificación más alta es 0.2253, y sólo contesta 3 preguntas completas, 5 preguntas son contestadas parcialmente y 13 quedan sin respuesta. Es decir, 14% de las preguntas son contestadas, 24 % son parcialmente contestadas y el 62% quedaron sin contestar. Estos valores muestran el caso opuesto a nuestro *SBR*.

Finalmente, para ambos sistemas se observa que las mejores calificaciones se obtuvieron cuando la lista de artículos y la solicitud conservaron las palabras sin contenido.

La tabla 6, muestra los resultados de los experimentos, en donde la lista de artículos fue lematizada. El mejor resultado se obtuvó para el experimento en donde a la lista de artí-

culos y a la solicitud se le removieron las palabras sin contenido. El valor de la calificación obtenida es de 0.6230. Para este caso, 13 preguntas fueron contestadas completamente, 7 preguntas fueron contestadas parcialmente y sólo 1 quedó sin respuesta. Lo que en porciento equivale al 62, 33 y 5%, respectivamente. De manera similar que los resultados de la tabla anterior, las calificaciones para un tipo de división dado son muy cercanas, lo cual indica que no importa si se camina del nodo A al nodo B o viceversa.

En contraste, con el mejor valor de calificación obtenido para el sistema basado en el MEV, es muy bajo, 0.3055, para el cual se contestaron 5 preguntas de manera completa, 7 preguntas de manera parcial y 14 preguntas quedaron sin respuesta.

Para ambos sistemas, los mejores resultados se obtuvieron a partir de la lista de artículos y solicitud, a las cuales se les removió las palabras sin contenido. Se destaca, de manera general, que la acción de lematizar la lista de artículos, produjo un incremento en el valor de la calificación, con respecto a las calificaciones de la tabla 5, donde no se incluyó esta herramienta, y por lo tanto, se incrementó de número de preguntas contestadas completamente, de 2 a 3.

Los resultados mostrados en la tabla 7, corresponden a los experimentos en donde, además de lematizar la lista de artículos se incluyó la acción de un tesauro distribucional. El mejor resultado observado es para el experimento en donde la lista de artículos y la solicitud conservan palabras sin contenido para una división de mezcla de términos, DMT, cuando se camina del nodo A al nodo B, El valor de la calificación es de 0.5461. Nuevamente, se aprecia que para este experimento sí, se invierte la forma de caminar, es decir del nodo B al nodo A, la calificación es 0.5456. Para ambos caso, las preguntas contestadas completamente fueron 13, 6 preguntas fueron contestadas parcialmente y sólo 3 quedaron sin respuesta. Los valores de las calificaciones están contenidas en un rango de 0.4689 a 0.5461, como puede observarse el rango es pequeño.

La mejor calificación arrojada del sistema basado en el MEV es de 0.2815 para el experimento en donde a la lista de artículos lematizada se le removieron las palabras sin contenido. En este caso las preguntas contestadas fueron 4 de forma completa, 7 preguntas fueron contestadas de manera parcial y 13 preguntas no fueron contestadas.

Cabe señalar, que para ambos sistemas el hecho de incluir la acción de un tesauro distribucional decrementó los valores de las calificaciones pero sin llegar a ser iguales o menores que los mostrados en la tabla 5.

Para terminar con esta sección, se describen las observaciones referentes a la tabla 8, la cual incluye los resultados de los experimentos correspondientes a la lista de artículos lematizada y a la aplicación del tesauro Anaya. De la inspección de los datos, la mejor calificación fue para el experimento en donde la lista de artículos lematizada conservó palabras sin contenido pero éstas fueron removidas de la solicitud, para el tipo de división a la mitad, DM, cuando se camina del nodo B al nodo A. La calificación para este caso fue de 0.5273. Para este mismo experimento, cuando se camina del nodo A al nodo B, la calificación es de 0.5250. Obsérvese que tal semejanza entre los valores, nuevamente implicaría que la forma de caminar sobre el grafo no es de relevancia en el proceso de búsqueda de respuesta. Adicionalmente, es importante señalar que los valores de esta tabla representan el rango más pequeño de todas la tablas, 0.4905 a 5273, lo cual implica un grado de homogeneidad alto, entre los valores. Lo que posiblemente indique que el tipo de división tampoco es un factor significativo para la búsqueda de respuestas. Ya que en casi todos los casos, fueron contestadas 12 preguntas de manera completa, de 5 a 6 de manera parcial y de 3 a 4 no fueron contestadas.

En cuando a la calificación derivada del sistema basado en el MEV, ésta fue de 0.2107 con que fueron contestadas sólo 3 preguntas de manera completa, 3 de forma parcial y 15 quedaron sin respuesta.

5.5 Discusión de los resultados

De acuerdo con la descripción de las tablas 5 a 8, se observa que para todos los casos los resultados del *SBR* propuesto fueron superiores a los resultados del sistema basado en el MEV. Por otro lado, los resultados más bajos obtenidos corresponden a los mostrados en la tabla 5.

De las dos observaciones anteriores se desprende, nuestro *SBR* pudo mejorar los resultados del sistema basado en el MEV, por el simple hecho de que está construido a través

del grafo, lo que implica que se está tomando en cuenta la relación que existe en los documentos mismos de la colección, además de su relación con la solicitud.

El uso del lematizador sobre la lista LBA mejora los resultados, en aproximadamente un 10%, sobre el valor de las calificaciones del *SBR*, mostradas en la tabla 5, lo cual necesariamente repercute en que aumente el número de preguntas contestadas, tanto completa como parcialmente de la tabla 6.

Los tesauros, al igual que el lematizador, tienen como finalidad mejorar de alguna manera la búsqueda de respuestas, sin embargo, esto no se ve reflejado en los resultados de la tabla 7 y 8, en las cuales se combinan estas dos herramientas. Los valores de las calificaciones de la tabla 6, que consideran meramente la aplicación del lematizador sobre la lista de artículos, son decrementados en las tablas 7 y 8, y adicionalmente se observa que número de preguntas sin contestar aumenta a 2 ó 3. Este hecho puede explicarse a partir de la forma de operar del *SBR*. El tesauro expande los términos de la solicitud, de tal manera, que por cada término presente, el tesauro puede agregar hasta 9 más. A partir, de las características del lenguaje legal -definido, preciso y técnico- podría ser que 9 términos sean demasiados, y que sólo hubiera bastado con 2 ó 3. Debido a que los tesauros están basados en documentos generales como La Enciclopedia Encarta 2004 y el Diccionario Anaya, es de esperarse que el tesauro complete la solicitud con términos que no necesariamente estén relacionados con el contexto en que se trabajó; así, no todos los términos de la solicitud estarían vinculados al contexto de la pregunta y ésta fuera la causa por la cual, la calificación disminuyó y las preguntas sin contestar aumentaron, con respecto a los resultados de la tabla 6.

Capítulo VI

6 Conclusiones

El *SBR* está basado en la construcción de un grafo, lo cual notablemente mejoró los resultados obtenidos con el sistema de búsqueda basado en el MEV; no obstante, las calificaciones mostradas en la tabla 5, muestran que el sistema aun tienen deficiencias, con respecto a las respuestas proporcionadas por el Abogado General del IPN para el conjunto de preguntas usadas.

Por otra parte, el uso del lematizador pudo mejorar las respuestas basadas puramente en el grafo en ~10%, con lo cual el número de preguntas contestadas, de forma completa y parcial, aumentó.

El uso de los tesauros usados, distribucional y el basado en el diccionario Anaya, no mostraron una mejora significativa sobre los resultados de los experimentos donde se empleó el lematizador. Se considera que esto fue debido a dos factores: las características del lenguaje legal y que los tesauros no están especializados en el marco legal, sino que son de uso general.

Los parámetros, tipo de división e inversión de los nodos A y B (donde está la solicitud) no mostraron tener un impacto relevante sobre la respuesta dada, en ningún caso considerado.

6.1 Recomendaciones y consideraciones

Es importante señalar que este tipo de *SBR* es propuesto por primera vez en su tipo, y surge de en una idea original, que basa su construcción en un grafo ponderado. Obviamente, como se ha visto, supera considerablemente al sistema de búsqueda de respuestas basado en el MEV, a pesar de que las preguntas contestadas completamente estén entre el 50 a 60%, de manera general. Este trabajo fue exploratorio para investigar su viabilidad en la búsqueda de respuestas en el dominio legal.

Del presente trabajo se derivan diferentes tópicos a considerar y mejorar:

- Tesauro distribucional.
- Experimentar con otras medidas de semejanza.
- Investigar con distintos tipos de ponderación de términos.
- Formas alternativas de extraer las rutas.

Referencias

1. Hirschman L, R. Gaizauskas. Natural Language Question Answering: The View From Here. *Natural Language Engineering* 7(4) 275 – 300, 2001.

2. Burger John, Cardie Claire, Chaudhri Vinay, Gaizauskas Robert, Harabagiu Sanda, Israel David, Jacquemin Christian, Lin Chin-Yew, Maiorano Steve, Miller George, Moldovan Dan, Ogden Bill, Prager John, Riloff Ellen, Singhal Amit, Shrihari Rohini, Strzalkowski Tomek, Voorhees Ellen, Weishedel Ralph. Issues, Tasks, and Program Structures to Roadmap Research in Question Answering (Q&A). In *Proceedings of Cross Language Evaluation Forum* (*CLEF*), Trondheim, Norway, 21-22 August, 2003.

3. Voorhees Ellen M. Overview of the *TREC* 2004 Question Answering Track. In *Proceedings of the 13th Text REtrieval Conference*, 2004.

4. Chung Hoojung, Young-In Song, Kyoung-Soo Han, Do-Sang Yoon, Joo-Young Lee, and Hae-Chang Rim. A Practical QA System in Restricted Domains. In *Workshop on Question Answering in Restricted Domains*. 42nd Annual Meeting of the Association for Computational Linguistics (ACL-2004), pages 39-45, Barcelona, Spain, 2004.

5. Tjong Erik, Kim Sang, Gosse Bouma, Maarten de Rijke. Developing Offline Strategies for Answering Medical Questions. In *Workshop on Question Answering in Restricted Domains*. 20th National Conference on Artificial Intelligence (AAAI-05), pages 41-45, Pittsburgh, PA., 2005.

6. Rinaldi Fabio, James Dowdall, and Gerold Schneider. Answering questions in the genomics domain. *In Proceedings of the ACL04 Workshop on Question Answering in Restricted Domains*, pages 46-53, Barcelona, Spain, 2004.

7. Yun Niu and Hirst Graeme. Analysis of Semantic Classes in Medical Text for Question Answering. In *Workshop on Question Answering in Restricted Domains*. 42nd Annual Meeting of the Association for Computational Linguistics (ACL-2004), pages 54-61, Barcelona, Spain, 2004.

8. Zhang Zhuo, Lyne Da Sylva, Colin Davidson, Gonzalo Lizarralde and Jian-Yun Nie, Domain-Specific QA for the Construction Sector. In *Workshop of IR4QA*: *Information Retrieval for Question Answering*, 27th ACM-SIGIR, Sheffield, July, 2004.

9. Quaresma Paulo and Irene Pimenta Rodrigues. A question-answering system for Portuguese juridical documents. In *Proceedings of the 10th international conference on Artificial intelligence and law*. International Conference on Artificial Intelligence and Law. 256 – 257. Bologna, Italy, 2005.

10. Quaresma Paulo and Irene Pimenta Rodrigues. A collaborative legal information retrieval system using dynamic logic programming. In *Proceedings of the 7th international conference on Artificial intelligence and law*. International Conference on Artificial Intelligence and Law. 190-191. Oslo, Norway, 1999.

11. Doan-Nguyen, Hai and Leila Kosseim. The problem of precision in restricted-domain question-answering. Some proposed methods of improvement. In *Workshop on Question Answering in Restricted Domains*. 42nd Annual Meeting of the Association for Computational Linguistics (ACL-2004), pages 8-15, Barcelona, Spain.

12. Diekema Anne R., Ozgur Yilmazel and Elizabeth D. Liddy. Evaluation of restricted domain question-answering systems. In *Workshop on Question Answering in Restricted Domains*. 42nd Annual Meeting of the Association for Computational Linguistics (ACL-2004), pages 2-7, Barcelona, Spain.

13. Mihalcea Rada, Random Walks on Text Structures. In *Proceedings of the Conference on Computational Linguistics and Intelligent Text Processing* (CICLing), Springer, Mexico City, February 2006.

14. Manning Christopher D. and Hinrich Schutze. *Foundations of Statistical Natural Language processing*. MIT Press, 1999.

15. Salton G., A. Wong and C. S. Yang. A vector Space Model for Automatic Indexing. *Information Retrieval and Language Processing*, 1975.

16. Salton G., Buckley C. Term – Weighting Approaches in Automatic Text Retrieval. *Information processing & management* Vol. 24. No. 5 pp. 513- 523. 1988.

17. Johnsonbaugh Richard. Matemáticas discretas. Sexta Edición. Pearson-Prentice-Hall. 2005

18. H. Calvo, A. Gelbukh. DILUCT: An Open-Source Spanish Dependency Parser Based on Rules, Heuristics, and Selectional Preferences. *In Lecture Notes in Computer Science* 3999, pp. 164-175

19. H. Calvo, A. Gelbukh, A. Kilgarriff, Distributional Thesaurus Versus WordNet: A Comparison of Backoff Techniques for Unsupervised PP Attachment. *In Computational Linguistics and Intelligent Text Processing* (CICLing 2005), Lecture Notes in Computer Science 3406, Springer 2005, pp. 177–188.

20. Biblioteca de Consulta Microsoft Encarta 2004, Microsoft Corporation, 1994–2004

21. Lázaro Carreter, F. (Ed.) Diccionario Anaya de la Lengua, Vox, 1991

22. Katz Boris, Borchardt Gary and Felshin Sue. Natural Language Annotations for Question Answering. FLAIRS Conference. 2006

23. Voorhees Ellen M. The *TREC* – 8 Question Answering Track Report. In *Proceedings of the Text 8th REtrieval Conference*, 1999.

24. Voorhees Ellen M. Overview of the *TREC* – 9 Question Answering Track. In *Proceedings of the Text 9th REtrieval Conference*, 2000

25. Voorhees Ellen M. Overview of the *TREC* 2001 Question Answering Track. In *Proceedings of the 10th Text REtrieval Conference*, 2001

26. Voorhees Ellen M. Overview of the *TREC* 2002 Question Answering Track. In *Proceedings of the 11th Text REtrieval Conference*, 2002

27. Voorhees Ellen M. Overview of the *TREC* 2003 Question Answering Track. In *Proceedings of the Text 12th REtrieval Conference*, 2003

28. Voorhees Ellen M. Dang Trang Hoa. Overview of the *TREC* 2005 Question Answering Track. In *Proceedings of the 14th Text REtrieval Conference*, 2005

29. Dan Trang Hoa, Lin Jimmy and, Kelly Diane. Overview of the *TREC* 2006 Question Answering Track. In *Proceedings of the 15th Text REtrieval Conference*, 2006

30. Dan Trang Hoa, Kelly Diane and, Lin Jimmy. Overview of the *TREC* 2007 Question Answering Track. In *Proceedings of the 16th Text REtrieval Conference*, 2007

31. Dell Zhang, Wee Sun Lee. Question Classification using Support Vector Machines. *In Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval.* 2003

32. Hedström Anna. Question Categorization for a Question Answering System using a Vector Space Model. Master´s thesis in Computational Linguistics.Uppsala University. Suiza. 2005

33. Singhal Amit, Abney Steve, Bacchiani Michiel , Collins Michael, Hindle Donald, Pereira Fernando. AT&T at TREC – 8. In *Proceedings of the Text 8th REtrieval Conference*, 1999.

34. Mollá Diego, Vicedo Jose Luis. Question Answering in Restricted domains: An overview.

Apéndice

A

Se encontró mediante experimentos que las palabras que se encuentran en esa lista tienen un valor muy bajo de relevancia con TFIDF. Esta lista fue establecida manualmente tomando, en cuenta este criterio.

el | la | los | las | un | uno | unas | unos | a | desde | para | de | con | contra | para | pero | bien | bueno | sin | so | sobre | tras | del | al | yo | mi | su | tu | y | que | por | se | en | una | o | ha | dos | sus | es | más | mas | como | cómo | muy | este | esta | éste | ésta | ese | esa | aquél | aquel | aquella | aquélla | aquellos | aquéllos | le | no | ser | sí | si | algo | algunos | alguno | algún | alguna | él | yo | tu | nos | nosotros | ustedes | ellos | ser | hacer | tener | se | me | te | los | lo | está | aquí | allí | ahí | sólo | solo

B

Conjunto de preguntas/respuestas obtenidas de la página del Abogado General del IPN (http://www.abogadogeneral.ipn.mx/faqdlc.html)

1. ¿Cómo se lleva a cabo el procedimiento de elección de representantes alumnos ante el Consejo Técnico Consultivo Escolar?

1 El proceso de elección de representantes alumnos ante los consejos técnicos consultivos de las escuelas, centros y unidades de enseñanza media superior está regulado en los **artículos 28 fracción IV de la Ley Orgánica, 206, 207, 209 y 213 fracción I del Reglamento Interno**, dispositivos retomados en las bases de la convocatoria autorizada mediante oficio número 014 SG/634/01, inscrita en la oficina del Abogado General con el número 582 a fojas 134 del libro de Registro de Actos Jurídicos Diversos.

2 El artículo 213 fracción I del Reglamento Interno preceptúa que los únicos requisitos para ser candidato a consejero alumno consiste en tener situación escolar regular y estar cursando estudios entre el tercero y penúltimo semestres

3 El artículo 209 del Reglamento Interno establece que comités conformados por alumnos con situación escolar regular son quienes deben elegir, en forma directa y por mayoría de votos, a los representantes educandos entre el Consejo Técnico Escolar.

2. ¿Cuáles son los consejeros que deben abstenerse de participar en procesos de elección de terna para la designación de Director?

El **artículo 181 del Reglamento Interno** del Instituto Politécnico Nacional dispone que los directores y subdirectores de las escuelas, centros y unidades de enseñanza son los únicos integrantes de los Consejos Técnicos Escolares que deben abstenerse de participar en las sesiones de dichos órganos colegiados en las que se trate de elegir las ternas para los cargos que ellos mismos ocupan.

3. ¿Es procedente que un alumno solicite mención honorífica a nivel licenciatura si escogió su Titulación por la opción de Escolaridad?

Del análisis integral del **Reglamento de Titulación Profesional** del Instituto Politécnico Nacional se desprende lo siguiente:

1 La opción de titulación por escolaridad resulta procedente si el alumno es superior a nueve y todas las materias fueron aprobadas en forma ordinaria (Capitulo II .De las Opciones de Titulación. **artículo 13**)

2 El pasante solo puede aspirar a recibir mención honorífica, si además de cubrir otros requisitos previstos en ese reglamento, presenta examen profesional (Capitulo VII ?Del Examen Profesional?, **artículo 43**)

3 La titulación por escolaridad no requiere de la presentación de examen profesional, por lo tanto, no se puede enmarcar en el artículo 43 fracción II del Reglamento en comento; al elegir esta opción de titulación, el pasante no puede obtener mención honorífica.

4. ¿Están facultados para actuar como asesores y/o sinodales en exámenes profesionales de licenciatura los profesores de nivel superior que han sido asesores de tesis de maestría y doctorado a pesar de que no poseen titulo profesional de licenciatura, pero cuentan con documentación probatoria de haber cursado estudios de posgrado en instituciones públicas nacionales como el CINVESTAV, o bien que estudiaron a nivel licenciatura en el extranjero y sus títulos no están registrados ante la Dirección General de Profesiones.

Los **artículos 24 y 33 fracción III del Reglamento de Titulación Profesional** del IPN establecen, respectivamente, que "El Asesor" en las opciones de titulación denominadas Proyecto de Investigación, Tesis, Memoria de experiencia profesional, Examen de conocimientos por áreas, Seminario de titulación y Practica Profesional debe ser titulado y los miembros del jurado para todas las opciones de titulación deberán reunir los siguientes requisitos III. Poseer título profesional debidamente registrado en la Dirección General de Profesiones. De lo anterior se desprende que los profesores que no reúnan los requisitos antes citados se encuentran material y jurídicamente impedidos para asesorar en cualquiera de las opciones de titulación y formar parte de los jurados correspondientes.

5. ¿La esposa del actual director puede ser candidata a dicho cargo?

El **Reglamento Interno del IPN en su artículo 168** establece que al frente de cada una de las escuelas, centros y unidades de enseñanza y de investigación habrá un director que será designado en los términos previstos por el **artículo 14, fracción XVII, de la Ley Orgánica**, y durará en su cargo tres años. Podrá ser designado por una sola vez, para otro periodo, de conformidad con lo dispuesto por el **artículo 21 del mismo ordenamiento**.

6. ¿La Oficina del Abogado General o cualquier otra autoridad del IPN puede pronunciarse sobre la constitución, estructuración, organización y funcionamiento de las sociedades, asociaciones o agrupaciones de alumnos y/o egresados?

El **artículo 4º, fracción X, de la Ley Orgánica del IPN**, establece que el Instituto puede participar en la constitución de asociaciones, sociedades y patronatos, que tengan por objeto impulsar el desarrollo de sus actividades y en la coordinación de las personas físicas o morales que contribuyan a la realización de las finalidades de esta Casa de Estudios. El Instituto se encuentra sujeto al régimen de facultades expresas, principio de derecho público a partir del cual, a través de sus órganos, instancias o servidores públicos, sólo puede realizar aquello que le está expresamente autorizado u ordenado en las disposiciones jurídicas vigentes y aplicables, por lo que resulta pertinente aclarar que el Instituto no puede -respaldar- a las mismas, ya que constituye un supuesto fáctico que no está previsto en la normatividad que rige al IPN.

7. ¿Los Colegios de Profesores o los Consejos Técnicos Consultivos Escolares están facultados para calificar sí los aspirantes cumplen o no los requisitos para ser director de una escuela, centro o unidad de enseñanza y de investigación?

Es facultad única y exclusiva del Director General designar a los directores de las escuelas, centros o unidades de enseñanza y de investigación, previa valoración del perfil que de cada uno de ellos realice, a efecto de determinar cual de los candidatos es el idóneo para representar como titular al Centro con todas las facultades y obligaciones que ello implica.

8. ¿Procede integrar consejo técnico consultivo escolar en un centro de investigación a efecto de que lleve a cabo proceso de elección de terna para designación de subdirectores a pesar de que el centro sólo sea de investigación, no así de enseñanza?
En términos de la normatividad vigente y aplicable en el IPN, los consejos técnicos consultivos escolares operan e las escuelas, centros y unidades de enseñanza; por su parte, corresponde a los colegios de profesores de los centros de investigación conocer y, en su caso, acordar sobre las ternas que se proponen a la Dirección General para la designación de subdirectores (**Artículos 27 y 28 de la ley Orgánica**, así como **202 y 264 fracción IV del Reglamento Interno**), por lo tanto, resulta improcedente integrar consejo técnico consultivo escolar en el caso de centros que solo se dediquen a la investigación, además de que es el colegio de profesores quien debe instrumentar el proceso de elección de terna para cargos de subdirector.

9. ¿Procede que un mismo aspirante sea propuesto para integrar más de una terna en el proceso de elección de ternas para la designación de subdirectores técnico, administrativo y académico?
Es procedente, debido a que la normatividad interna del IPN no la prohíbe, basta con que los participantes cumplan con los requisitos que, para ser subdirector, determinan los **artículos 174 y 175 del Reglamento Interno del Instituto**.

10. ¿Procede que una persona que en dos ocasiones ha sido subdirector participe en el proceso de elección de ternas para la designación de subdirectores técnico, administrativo y académico y en su caso ejerza otra vez ese cargo.
Resulta improcedente que el subdirector participe nuevamente en el proceso, de acuerdo a lo dispuesto por los **artículos 21 de la Ley Orgánica** del IPN y **182 de su Reglamento Interno**, en los cuales precisa el límite del ejercicio de los cargos directivos.

11. ¿Puede un mismo individuo por segundo año consecutivo ser electo consejero representante del personal académico?
El Reglamento Interno del Instituto Politécnico Nacional dispone en su **artículo 215** que los consejeros deben durar un año en el desempeño de sus funciones y prohíbe la reelección de estos para el periodo inmediato de los consejos técnicos consultivos escolares.

12. ¿Pueden prestar el servicio social alumnos que hayan acreditado el 70% de créditos pero adeuden materias?
Artículo 2 Reglamento servicio social para la correcta interpretación y aplicación del presente Reglamento, se entenderá por:
V.- Prestador. Al alumno regular que haya acreditado como mínimo el setenta por ciento de sus créditos académicos o el cien por ciento de los mismos, según sea el caso, y que este realizando su servicio?.
De lo anterior se concluye que la normatividad es muy clara y por lo tanto se debe aplicar con estricto apego a las disposiciones citadas.

13. ¿Pueden votar los profesores del colegio que se encuentran en -año sabático-?
El **Reglamento de las Condiciones Interiores de Trabajo del Personal Académico del IPN**, en su **artículo 80** dispone que el año sabático tenga como finalidad permitir al personal académico -dedicarse al estudio y a la realización de actividades que le permitan superarse académicamente en beneficio propio y del IPN-. Es decir, quien ejerce su derecho al período sabático se encuentra en

activo. Además, no existe disposición expresa que limite a quienes se encuentren gozando año sabático para votar en los procesos de elección de ternas para la designación de director, de ahí que pueden hacerlo.

14. ¿Qué debe entenderse por tener estudios de posgrado?
Debemos entender la palabra -tener- como poseer o haber realizado, en este caso, estudios de posgrado, en los términos previstos en el **artículo 4 del Reglamento de Estudios de Posgrado**:
"Son considerados estudios de posgrado en el Instituto aquellos que se realizan después de la licenciatura y se imparten por las unidades académicas de acuerdo con lo establecido en el presente Reglamento y en las políticas y lineamientos que para tal efecto emita la Secretaría, cuya finalidad es la de formar recursos humanos del más alto nivel científico y tecnológico en los niveles de especialidad, maestría y doctorado y podrán realizarse en las modalidades y orientaciones siguientes:
I. Especialidad y especialidad médica, que se realizarán con orientación profesional;
II. Maestría en Administración, que se realizará con orientación profesional;
III. Maestría en Ingeniería, que podrá realizarse con orientación profesional o científica;
IV. Maestría en Ciencias, que se realizará con orientación científica;
V. Doctorado en Ingeniería y Doctorado en Ciencias, que se realizarán con orientación científica, y
VI. Las demás que se establezcan con posterioridad, de conformidad con el presente Reglamento y el marco normativo aplicable."

15. ¿Respecto del perfil de Subdirector de Investigación Aplicada, procede incluir en la convocatoria para ocupar el cargo la mención ?poseer estudios de posgrado? como requisito indispensable, por analogía con las subdirecciones Académica y Científica?
El Reglamento Interno del I.P.N. establece que los subdirectores deben reunir entre otros requisitos el de poseer titulo profesional; sólo a los **subdirectores académicos de centros de investigación** les exige tener grado académico de doctor o equivalente (**artículos 174 fracción II 175 fracción I**).
La normatividad vigente del IPN no prevé requisitos particulares para ser titular de la Subdirección de Investigación Aplicada a que se alude en la consulta.

16. ¿Tienen derecho de votar o no los profesores interinos adscritos a alguna escuela?
El **artículo 208 del Reglamento Interno del IPN** establece que el personal académico de cada programa académico, o equivalente, elegirá en forma directa y por mayoría de votos a sus respectivos representantes.
Por su parte, el **artículo 6 del Reglamento de las Condiciones Interiores de Trabajo del Personal Académico en sus fracciones XV y XXII** definen a quien se debe considerar y que funciones desempeña el personal académico clasificado como interino. En tal virtud y, dada la claridad del texto de los artículos referidos, que no dejan lugar a duda, y al principio general de derecho que dice que cuando la ley no distingue, el interprete no tiene porque distinguir, es procedente que los profesores interinos voten en el proceso de elección de sus representantes del personal académico ante los Consejos Técnicos Consultivos Escolares.

17. ¿Un alumno que adeuda cinco materias puede seguir fungiendo como consejero representante de los alumnos?
Aquel alumno que adeude cinco materias se ubica en las hipótesis descritas en los **artículos 81, fracción III del Reglamento Interno** y **44 del Reglamento de Estudios Escolarizados para los Niveles Medio Superior y Superior**, de ahí que causaría baja, salvo que obtenga resolución favorable de la Comisión de Situación Escolar del Consejo Técnico Consultivo Escolar, en caso de no ocurrir lo último, estaría suspendida su calidad de alumno y, por consiguiente, de consejero.

18. ¿Un consejero que ha faltado a varias sesiones del Consejo Técnico Consultivo Escolar puede participar en el proceso de elección de terna?
El consejero que se ubique en alguno de los siguientes supuestos (**Artículo 200 del Reglamento Interno**), no podrá participar con ese carácter en el proceso de elección de terna:
I. El consejero deje de tener la categoría de personal de académico o alumno;
II. El consejero renuncie a su representación;
III. La ausencia se produzca por enfermedad que amerite incapacidad médica de más de tres meses, y
IV. El consejero falte, sin causa justificada, en el término de seis meses a tres sesiones ordinarias del Consejo o a tres sesiones de las comisiones permanentes para las que fue electo.

19. ¿Un Maestro Decano en período de prejubilación puede presidir el Consejo Técnico Consultivo Escolar o el Colegio de Profesores para dirigir el proceso de elección de ternas para la designación de director?
Los maestros decanos que estén en goce de licencia pre - pensionaria sólo deben ocuparse de los trámites de jubilación, toda vez que éste permiso es improrrogable e irrenunciable. Cuando se presente un proceso de elección de terna para director de alguna escuela, centro o unidad de enseñanza y de investigación y, ante la eventualidad de que el Maestro Decano correspondiente se encuentre con licencia pre - pensionaria, es pertinente aplicar lo preceptuado en el segundo párrafo del **artículo 15 del Reglamento del Cuerpo de Maestros Decanos**.

20. ¿Una persona puede ser director por tercera ocasión?
El **artículo 21 de la Ley Orgánica del IPN** vigente dispone que los directores de escuelas, centros y unidades de enseñanza y de investigación durarán en su cargo tres años y podrán ser designados por una sola vez, para otro periodo.

21. ¿Cuáles son los requisitos que debe cumplir un director de escuela, centro o unidad de enseñanza y de investigación, para designar a los jefes de división o de departamento a su cargo?
De conformidad con lo establecido en la fracción IV del **artículo 173 del Reglamento Interno** del Instituto Politécnico Nacional, es facultad del director de la escuela, centro o unidad de enseñanza y de investigación, designar, en los términos de las disposiciones jurídicas y administrativas aplicables, al personal adscrito a la escuela, centro o unidad a su cargo. En razón de lo anterior, para efectuar dicha designación, el director de la escuela, centro o unidad de que se trate, le corresponde cumplir con lo dispuesto en la fracción XV del artículo 173 del citado Reglamento, es decir, se deberá acordar con el Director General y con las demás autoridades centrales de esta Casa de Estudios.

C

Colección de documentos que el *SBR* emplea (se pueden descargar de la página del Abogado General del IPN, http://www.abogadogeneral.ipn.mx)

1. Ley Orgánica
2. Reglamento orgánico
3. Reglamento interno
4. Reglamento interno de la Comisión de Operación y Fomento de Actividades Académicas (COFAA)
5. Reglamento del Decanato
6. Reglamento de las condiciones generales de trabajo del personal no docente
7. Reglamento de las condiciones interiores de trabajo del personal académico
8. Reglamento del consejo general consultivo

9. Reglamento de Academias
10. Reglamento del Archivo histórico
11. Reglamento de becas del Consejo Nacional de Ciencias y Tecnología (CONACyT)
12. Reglamento de becas, estímulos y otros medios de apoyo para alumnos del IPN
13. Reglamento de becas estudio, apoyos económicos y licencias con goce de sueldo del personal académico
14. Reglamento de diplomados
15. Reglamento de distinciones al mérito politécnico
16. Reglamento de estudios escolarizados para los niveles medio y superior del IPN
17. Reglamento de evaluación del IPN
18. Reglamento del programa de estímulo al desempeño docente
19. Reglamento de planeación
20. Reglamento de prácticas y visitas escolares
21. Reglamento de estudios de posgrado
22. Reglamento para la operación, administración y uso de la red institucional de cómputo y telecomunicaciones del IPN
23. Reglamento de incorporación, reconocimiento de validez oficial, equivalencia y revalidación de estudios
24. Reglamento del sistema de becas por exclusividad
25. Reglamento de servicio social
26. Reglamento de titulación profesional

D

En este anexo se encuentran los Scripts escritos en Perl y utilizados en los experimentos.

Script que genera el grafo empleado en el *SBR*.

```
   #!/usr/bin/perl
use strict;
   my $year;
   my @months = qw(Jan Feb Mar Apr May Jun Jul Aug Sep Oct Nov Dec);
   my @weekDays = qw(Sun Mon Tue Wed Thu Fri Sat Sun);
   my ($second, $minute, $hour, $dayOfMonth, $month, $yearOffset, $dayOfWeek,
$dayOfYear, $daylightSavings) = localtime();
   $year = 1900 + $yearOffset;
   my $theTime = "Inicio de proceso: $hour:$minute:$second, $weekDays[$dayOfWeek]
$months[$month] $dayOfMonth, $year\n";
   print $theTime;
   my @ToT;

     print "Pre-procesamiento archivos reglamentos\n";
#Se transfieren los artículos de cada uno de los reglamentos a un arreglo
   my $n_a = 'c:/sp/genarch/regs/sinlemm/RPSG_IPN.txt';
   my ($nta_r1, $n_arts, @AoA) =  obt_arts ($n_a, 0, 1, @ToT);
   print "$nta_r1\n";
   $n_a = 'c:/sp/genarch/regs/sinlemm/RI_IPN.txt';
   my $nta_r2;
   ($nta_r2, $n_arts, @ToT) =  obt_arts ($n_a, $n_arts, 1, @AoA);
   print "$nta_r2\n";
   $n_a = 'c:/sp/genarch/regs/sinlemm/LO_IPN.txt';
   my $nta_r3;
   ($nta_r3, $n_arts, @AoA) =  obt_arts ($n_a, $n_arts, 1, @ToT);
```

```
print "$nta_r3\n";
$n_a = 'c:/sp/genarch/regs/sinlemm/RORGANICO.txt';
my $nta_r4;
($nta_r4, $n_arts, @ToT) =  obt_arts ($n_a, $n_arts, 1, @AoA);
print "$nta_r4\n";
$n_a = 'c:/sp/genarch/regs/sinlemm/RACADEMIAS.txt';
my $nta_r5;
($nta_r5, $n_arts, @AoA) =  obt_arts ($n_a, $n_arts, 1, @ToT);
print "$nta_r5\n";
$n_a = 'c:/sp/genarch/regs/sinlemm/RAHIPN.txt';
my $nta_r6;
($nta_r6, $n_arts, @ToT) =  obt_arts ($n_a, $n_arts, 1, @AoA);
print "$nta_r6\n";
$n_a = 'c:/sp/genarch/regs/sinlemm/RBECASCONACYT.txt';
my $nta_r7;
($nta_r7, $n_arts, @AoA) =  obt_arts ($n_a, $n_arts, 1, @ToT);
print "$nta_r7\n";
$n_a = 'c:/sp/genarch/regs/sinlemm/RBECESTYOMAPOYO.txt';
my $nta_r8;
($nta_r8, $n_arts, @ToT) =  obt_arts ($n_a, $n_arts, 1, @AoA);
print "$nta_r8\n";
$n_a = 'c:/sp/genarch/regs/sinlemm/RCGC.txt';
my $nta_r9;
($nta_r9, $n_arts, @AoA) =  obt_arts ($n_a, $n_arts, 1, @ToT);
print "$nta_r9\n";
$n_a = 'c:/sp/genarch/regs/sinlemm/RCGTPNoD.txt';
my $nta_r10;
($nta_r10, $n_arts, @ToT) =  obt_arts ($n_a, $n_arts, 1, @AoA);
print "$nta_r10\n";
$n_a = 'c:/sp/genarch/regs/sinlemm/RCITPA.txt';
my $nta_r11;
($nta_r11, $n_arts, @AoA) =  obt_arts ($n_a, $n_arts, 1, @ToT);
print "$nta_r11\n";
$n_a = 'c:/sp/genarch/regs/sinlemm/RCOTEPABE.txt';
my $nta_r12;
($nta_r12, $n_arts, @ToT) =  obt_arts ($n_a, $n_arts, 1, @AoA);
print "$nta_r12\n";
$n_a = 'c:/sp/genarch/regs/sinlemm/RDIPLOMADOS.txt';
my $nta_r6;
($nta_r13, $n_arts, @AoA) =  obt_arts ($n_a, $n_arts, 1, @ToT);
print "$nta_r13\n";
$n_a = 'c:/sp/genarch/regs/sinlemm/RDISTM.txt';
my $nta_r14;
($nta_r14, $n_arts, @ToT) =  obt_arts ($n_a, $n_arts, 1, @AoA);
print "$nta_r14\n";
$n_a = 'c:/sp/genarch/regs/sinlemm/REVAL.txt';
my $nta_r15;
($nta_r15, $n_arts, @AoA) =  obt_arts ($n_a, $n_arts, 1, @ToT);
print "$nta_r15\n";
$n_a = 'c:/sp/genarch/regs/sinlemm/RICOFFA.txt';
my $nta_r16;
($nta_r16, $n_arts, @ToT) =  obt_arts ($n_a, $n_arts, 1, @AoA);
print "$nta_r16\n";
$n_a = 'c:/sp/genarch/regs/sinlemm/RPESTIMULOSDOCENTE.txt';
my $nta_r17;
($nta_r17, $n_arts, @AoA) =  obt_arts ($n_a, $n_arts, 1, @ToT);
print "$nta_r17\n";
$n_a = 'c:/sp/genarch/regs/sinlemm/RPLAN.txt';
my $nta_r18;
($nta_r18, $n_arts, @ToT) =  obt_arts ($n_a, $n_arts, 1, @AoA);
print "$nta_r18\n";
$n_a = 'c:/sp/genarch/regs/sinlemm/RTITULACION.txt';
my $nta_r19;
```

```
   ($nta_r19, $n_arts, @AoA) =  obt_arts ($n_a, $n_arts, 1, @ToT);
   print "$nta_r19\n";
   $n_a = 'c:/sp/genarch/regs/sinlemm/RPRACYVISITAS.txt';
   my $nta_r20;
   ($nta_r20, $n_arts, @ToT) =  obt_arts ($n_a, $n_arts, 1, @AoA);
   print "$nta_r20\n";
   $n_a = 'c:/sp/genarch/regs/sinlemm/RRED.txt';
   my $nta_r21;
   ($nta_r21, $n_arts, @AoA) =  obt_arts ($n_a, $n_arts, 1, @ToT);
   print "$nta_r21\n";
   $n_a = 'c:/sp/genarch/regs/sinlemm/RRVOE.txt';
   my $nta_r22;
   ($nta_r22, $n_arts, @ToT) =  obt_arts ($n_a, $n_arts, 1, @AoA);
   print "$nta_r22\n";
   $n_a = 'c:/sp/genarch/regs/sinlemm/RSIBE.txt';
   my $nta_r23;
   ($nta_r23, $n_arts, @AoA) =  obt_arts ($n_a, $n_arts, 1, @ToT);
   print "$nta_r23\n";
   $n_a = 'c:/sp/genarch/regs/sinlemm/RSS.txt';
   my $nta_r24;
   ($nta_r24, $n_arts, @ToT) =  obt_arts ($n_a, $n_arts, 1, @AoA);
   print "$nta_r24\n";
   $n_a = 'c:/sp/genarch/regs/sinlemm/REE.txt';
   my $nta_r25;
   ($nta_r25, $n_arts, @AoA) =  obt_arts ($n_a, $n_arts, 1, @ToT);
   print "$nta_r25\n";
   $n_a = 'c:/sp/genarch/regs/sinlemm/RDEC.txt';
   my $nta_r26;
   ($nta_r26, $n_arts, @ToT) =  obt_arts ($n_a, $n_arts, 1, @AoA);
   print "$nta_r26\n\n$n_arts\n";
   @AoA = @ToT;
#######################################################################
   print "Generando archivo y base de datos tipos TyTR\n";
   dbmopen my %tyt, "TyTR" , 0666 or die "Imposible crear BD !!!";
   (%tyt) = obt_tyt  ($n_arts , @AoA);
   dbmclose %tyt;
#######################################################################
   print "Generando archivo tipos con peso por artículo TxAW.txt\n";
   my (%ppxa) = obt_tyt1  ($n_arts , @AoA);
   open(TXTO_1,"+> TxAW.txt") || die "Imposible crear TxAW.txt !!!\n";
   for my $ctr (1 .. $n_arts) {
      my @tmp = split /\#/, $ppxa{$AoA[$ctr][0]};
      print TXTO_1"$AoA[$ctr][0]\n";
      for my $x (0 .. @tmp - 1) {
         print TXTO_1"\t$tmp[$x]\n";
      }
   }
   close(TXTO_1);
#######################################################################
   print "Generando Matriz de adyacencia MS.txt\n";
   use BerkeleyDB ;
   use vars qw( %A ) ;
   my $filename = "BD_1" ;
   unlink $filename ;
   tie %A, "BerkeleyDB::Hash",
               -Filename => $filename,
               -Flags    => DB_CREATE
       or die "Cannot open file $filename: $! $BerkeleyDB::Error\n" ;
   open(TXTO_1,"+> MS.txt") || die "Imposible crear MS.txt\n";
   for my $i1 (1 .. $n_arts - 1) {
      print TXTO_1"A$i1\n";
      my $stc1 = "A".$i1;
      my %v_tc1;
```

```
      my $sc1 = 0;
      my $lva;
      my @tmp = split /\#/, $ppxa{$stc1};
      for $lva (0 .. @tmp - 1) {
         my @tmp1 = split /\|/, $tmp[$lva];
         $v_tc1{$tmp1[0]} = $tmp1[1];
         $sc1 = $sc1 + $tmp1[1]*$tmp1[1];
      }
      $A{$stc1."#".$stc1} = 1e15;
      for my $num_art2 ($i1 + 1 .. $n_arts) {
         my $stc2 = "A".$num_art2;
         my $lvb;
         my %v_tc2;
         my $sc2 = 0;
         my @tmp = split /\#/, $ppxa{$stc2};
         for $lvb (0 .. @tmp - 1) {
            my @tmp1 = split /\|/, $tmp[$lvb];
            $v_tc2{$tmp1[0]} = $tmp1[1];
            $sc2 = $sc2 + $tmp1[1]*$tmp1[1];
         }
         my $pp = 0;
         if ($lva <= $lvb) {
            foreach my $llave (keys %v_tc1) {
               $pp = $pp + ($v_tc1{$llave}*$v_tc2{$llave});
            }
         } else {
            foreach my $llave (keys %v_tc2) {
               $pp = $pp + ($v_tc1{$llave}*$v_tc2{$llave});
            }
         }
         my $sim = $pp/(sqrt($sc2)*sqrt($sc1));
         if ($sim != 0) {
            my $sim_1 = sprintf ("%.3f",1/(100*($pp/(sqrt($sc2)*sqrt($sc1)))));
            if ($sim_1 <= 1) {
               $A{$stc1."#".$stc2} = $sim_1;
               $A{$stc2."#".$stc1} = $sim_1;
               print TXTO_1"\t",$sim_1,"\n";
            } else {
               $A{$stc1."#".$stc2} = 1e15;
               $A{$stc2."#".$stc1} = 1e15;
               print TXTO_1"\t",1e15,"\n";
            }
         } else {
            $A{$stc1."#".$stc2} = 1e15;
            $A{$stc2."#".$stc1} = 1e15;
            print TXTO_1"\t",1e15,"\n";
         }
      }
   }
   $A{"A$n_arts#A$n_arts"} = 1e15;
   print TXTO_1"A$n_arts ";
   close(TXTO_1);
   untie %A ;
print "Proceso finalizado exitosamente a las: ";
($second, $minute, $hour, $dayOfMonth, $month, $yearOffset, $dayOfWeek, $dayO-
fYear, $daylightSavings) = localtime();
$year = 1900 + $yearOffset;
$theTime = "$hour:$minute:$second, $weekDays[$dayOfWeek] $months[$month] $dayOf-
Month, $year\n\n";
print $theTime;
#*****************************************************************
#SUB RUTINAS
#*****************************************************************
```

```
sub obt_arts {
   my ($n_arch, $r, $salida, @AoT) = @_;
   my $art;
   my $c;
   my %A;
   my $nar = 0;
   my @line;
   my $line;
   my $i = 0;
   open (fhtxt, $n_arch) || die "\nImposible abrir $n_arch\n";
#Cada línea del archivo se texto se transfiere a un arreglo: @line.
   while ($line = <fhtxt>) {
      chomp($line);
#Se eliminan los espacios en blanco iníciales de cada línea.
#No se almacenan líneas en blanco.
      $line =~ s/^\s*(.*)/$1\n/g;
      if ($line ne "\n"){
         $line[$i++] = $line;
      }
   }
   close (fhtxt);
   for my $q (0 .. $i - 1) {
#Se almacena en un arreglo y en un hash el texto
#de cada artículo contenido en el reglamento
#@AoA[1 .. n][0] = "A1 .. An"
#@AoA[1 .. n][1] = Contenido artículo
#%A{"A1 .. An"} = Contenido artículo
chomp ($line[$q]);
if($line[$q] =~ /[a-z]+/){
if ($line[$q] =~ m/(Artículo \d+\W*)\.(.*)/) {
$c = 0;
$r = $r + 1;
#Se obtiene el número de artículos del reglamento
            $nar = $nar + 1;
            $art = "A".$r;
            $A{$art} = $2;
            $AoT[$r][$c++] = $art;
            $AoT[$r][$c] = $2;
         }
         else {
            if ($c == 1) {
               $A{$art} = $A{$art}."\n".$line[$q];
               $AoT[$r][$c] = $A{$art};
            }
         }
      }
   }
   close (fhtxt);
   delete $A{''};
   if ($salida == 0) {
      return ($r, $nar, %A);
#El hash se retorna con el texto sin cambios
   } else {
#El texto almacenado en el arreglo se convierte
#a minúsculas y se eliminan signos de puntuación
#y números.
for my $i (1 .. $r) {
$AoT[$i][1] =~ s/\-/ /g;
$AoT[$i][1] =~ s/\-//g;
         $AoT[$i][1] =~ s/\[gtah]//g;
         $AoT[$i][1] =~ s/\•//g;
         $AoT[$i][1] =~ s/\‟//g;
         $AoT[$i][1] =~ s/\”//g;
```

```
            $AoT[$i][1] =~ s/\"//g;
            $AoT[$i][1] =~ s/\+//g;
            $AoT[$i][1] =~ s/\=//g;
            $AoT[$i][1] =~ s/\_+//g;
            $AoT[$i][1] =~ s/\://g;
            $AoT[$i][1] =~ s/\;//g;
            $AoT[$i][1] =~ s/\.//g;
            $AoT[$i][1] =~ s/\,//g;
            $AoT[$i][1] =~ s/\°//g;
            $AoT[$i][1] =~ s/\(//g;
            $AoT[$i][1] =~ s/\)//g;
            $AoT[$i][1] =~ s/\///g;
            $AoT[$i][1] =~ s/\d+//g;
            $AoT[$i][1] =~ s/\Á/á/g;
            $AoT[$i][1] =~ s/\É/é/g;
            $AoT[$i][1] =~ s/\Í/í/g;
            $AoT[$i][1] =~ s/\Ó/ó/g;
            $AoT[$i][1] =~ s/\Ú/ú/g;
            $AoT[$i][1] = lc($AoT[$i][1]);
         }
         return ($nar, $r, @AoT);
      }
   }
   sub obt_tyt {
      my ($r , @AoT) = @_;
   #Se obtiene en cuantos artículos aparece
   #cada tipo del reglamento y se eliminan los tipos
   #de la "stoplist"
      my %wrds;
      my @line;
      my $i = 0;
      my %arts_p;
      my %wf;
      my $ntr = 1;
      my @tmp = split /\s+/, $AoT[1][1];
      for my $x (0 .. @tmp - 1) {
         $wrds{$tmp[$x]} = 1;
         $wf{$tmp[$x]} = $wf{$tmp[$x]} + 1;
      }
      for my $i (2 .. $r) {
         my @tmp = split /\s+/, $AoT[$i][1];
         my %at;
         for my $x (0 .. @tmp - 1) {
            $wrds{$tmp[$x]} = $wrds{$tmp[$x]} + 1;
            $wf{$tmp[$x]} = $wf{$tmp[$x]} + 1;
            $at{$tmp[$x]} = $at{$tmp[$x]} + 1;
         }
         foreach my $llave (keys %at) {
            $wrds{$llave} = $wrds{$llave} - $at{$llave}+1;
         }
      }
      my $stopl;
      open (fhtxt1, 'stopl.txt') || die "stopl.txt no puede abrirse";
      while ($stopl = <fhtxt1>){chomp($stopl);delete $wrds{$stopl};}
      close('stopl.txt');
      delete $wrds{''};
      open(TXTO_1,"+> Tipos.txt") || die "Imposible crear Tipos.txt !!!\n";
      foreach my $llave (keys %wrds) {
         print TXTO_1"",$ntr++,"\t$llave\t$wrds{$llave}\t$wf{$llave}\n";
      }
      close(TXTO_1);
      return (%wrds);
   }
```

```
sub obt_tyt1 {
#Se obtiene la frecuencia de cada tipo x artículo
#el número de tipos del artículo
my ($r , @AoT) = @_;
my $ntr = 0;                                                                    #Se
realiza el cálculo de tfidf
   my %arts_p;
   my $bdn = "TyTR";
   dbmopen my %wrds, $bdn , 0666 or die "El archivo no se puede crear/abrir!!!";
   open(TXTO_1,"+> W.txt") || die "Imposible crear TxAW.txt !!!\n";
   for my $ctr (1 .. $r) {
      my %ppa;
      $AoT[$ctr][1] =~ s/\d+\w+//g;
      $AoT[$ctr][1] =~ s/\d+//g;
      my @tmp = split /\s+/, $AoT[$ctr][1];
      for my $x (0 .. @tmp - 1) {
         chomp($tmp[$x]);
         $ppa{$tmp[$x]} = $ppa{$tmp[$x]} + 1;
      }
      my $stopl;
      open (fhtxt1, 'stopl.txt') || die "stopl.txt no puede abrirse";
      while ($stopl = <fhtxt1>){chomp($stopl);delete $ppa{$stopl};}
      close('stopl.txt');
      delete $ppa{''};
      my $x = 0;
      foreach my $llave (keys %ppa) {
         $x++;
      }
      my $tmpl = "";
      print TXTO_1"A$ctr\t$x\t$r\n\n";
      foreach my $llave (keys %ppa) {
         my $aux = sprintf ("%.3f",
(($ppa{$llave})/$x)*((log($r/$wrds{$llave}))/log(10)));
         print TXTO_1"$llave\t$ppa{$llave}\t$wrds{$llave}\t$aux\n";
         $ppa{$llave} = $llave."|".$aux;
         $tmpl = $tmpl.$ppa{$llave}."#";
      }
      print TXTO_1"\n";
      $arts_p{$AoT[$ctr][0]} = $tmpl;
   }
   dbmclose %wrds;
   close(TXTO_1);
   my %arts_p1 = %arts_p;
   return (%arts_p1);
}
```

Script que implementa el método propuesto para la búsqueda de respuestas.

```
#!C:\perl\bin\perl.exe -w
use strict;
use BerkeleyDB ;
#La base de datos TyTR contiene un índice invertido que contiene además en cuántos artículos aparece cada tipo
   dbmopen my %tyt, "C:/Sp/GenArch/Lemm/Sesw/TyTR" , 0666 or die " Imposible abrir base de datos!!!"
   my @AR;
   my %hs;
   my $filename = "C:/Sp/GenArch/Lemm/Sesw/BD_1";
   my $p;
   my $c;
   my $nta_r1 = 107;              #RPSG
   my $nta_r2 = 295;              #RI
   my $nta_r3 = 34;               #LO
   my $nta_r4 = 84;               #RO
   my $nta_r5 = 34;               #RAcademias
   my $nta_r6 = 72;               #RAHIPN
   my $nta_r7 = 31;               #RBECASCONACYT
   my $nta_r8 = 33;               #RBECESTYOMAPOYO
   my $nta_r9 = 54;               #RCGC
   my $nta_r10 = 144;             #RCGTPNoD
   my $nta_r11 = 153;             #RCITPA
   my $nta_r12 = 39;              #RCOTEPABE
   my $nta_r13 = 42;              #RDIPLOMADOS
   my $nta_r14 = 39;              #RDISTM
   my $nta_r15 = 28;              #REVAL
   my $nta_r16 = 28;              #RICOFFA
   my $nta_r17 = 46;              #RPESTIMULOSDOCENTE
   my $nta_r18 = 23;              #RPLAN
   my $nta_r19 = 46;              #RTITULACION
   my $nta_r20 = 34;              #RPRACYVISITAS
   my $nta_r21 = 61;              #RRED
   my $nta_r22 = 85;              #RRVOE
   my $nta_r23 = 57;              #RSIBE
   my $nta_r24 = 25;              #RSS
   my $nta_r25 = 78;              #REE
   my $nta_r26 = 26;              #RDEC
   my $n_arts = 1698;             #TOTAL DE ARTICULOS
#Se crean índices para almacenamiento de query1 y query2
#como nuevos artículos y posteriormente compararlos con el resto
   my $fq = $n_arts + 1;
   my $sq = $n_arts + 2;
   my $ni = $fq;
   my $nf = $sq;
   my $nn = $sq;
   my %N_E;
   my %N_E1;
   my %N_E2;
#Se obtiene en un hash los tipos de cada artículo con peso
```

```
   my (%ppxa) = obt_catcp ('C:/Sp/GenArch/Lemm/Sesw/TxAW.txt');
   my @Preguntas;
   $Preguntas[1] = "llevar a cabo procedimiento de elección de representantes alumnos ante Consejo Técnico Consultivo Escolar";
   $Preguntas[2] = "ser consejeros deber abstener de participar en proceso de elección de terna para designación de Director";
   $Preguntas[3] = "ser procedente que alumno solicite mención honorífica a nivel licenciatura si escoger Titulación por opción de
Escolaridad";
   $Preguntas[4] = "estar facultado para actuar como asesores y/o sinodales en exámenes profesionales de licenciatura profesores de
nivel superior ser asesores de tesis de maestría y doctorado a pesar de que poseer no titulo profesional de licenciatura , pero con-
tar con documentación probatoria de cursar estudio de posgrado en instituciones públicas nacionales como CINVESTAV , o bien que estu-
diar a nivel licenciatura en extranjero y títulos estar no registrado ante Dirección General de Profesiones";
   $Preguntas[5] = "esposa de director actual poder ser  a cargo dicho";
   $Preguntas[6] = "Oficina de Abogado General o autoridad de IPN poder pronunciar sobre constitución , estructuración , organización
y funcionamiento de sociedades , asociaciones o agrupaciones de alumnos y/o egresados";
   $Preguntas[7] = "Colegios de Profesores o los  Escolares estar facultado para calificar sí aspirantes  o no requiso para ser di-
rector de escuela , centro o unidad de enseñanza y de investigación";
   $Preguntas[8] = "proceder integrar consejo técnico consultivo escolar en centro de investigación a efecto de llevar a cabo proceso
de elección de terna para designación de subdirectores a pesar de que centro ser sólo de investigación , no así de enseñanza";
   $Preguntas[9] = "proceder que aspirante proponer para integrar más de terna en proceso de elección de ternas para designación de
subdirectores técnico , administrativo académico";
   $Preguntas[10] = "proceder que persona que en ocasiones ser subdirector participe en proceso de elección de ternas para designa-
ción de subdirectores técnico , administrativo académico y en caso ejercer vez cargo";
   $Preguntas[11] = "poder individuo por año segundo consecutivo ser electo consejero representante de personal académico";
   $Preguntas[12] = "poder prestar servicio social alumnos acreditar 70% de créditos pero adeudar materias";
   $Preguntas[13] = "poder votar profesores de colegio encontrar en año sabático";
   $Preguntas[14] = "deber entender por tener estudio de posgrado";
   $Preguntas[15] = "respecto de perfil de Subdirector de Investigación Aplicada , proceder incluir en convocatoria para ocupar cargo
mención poseer estudio de posgrado como requiso indispensable , por analogía con subdirecciones Académica y Científica";
   $Preguntas[16] = "tener derecho de votar o no profesores interino adscrito a escuela";
   $Preguntas[17] = "alumno adeudar materias poder seguir fungir como representante consejero de alumnos";
   $Preguntas[18] = "consejero faltar a sesiones de Consejo Técnico Consultivo poder Escolar participar en proceso de elección de
terna";
   $Preguntas[19] = "Decano en período de prejubilación poder presidir Consejo Técnico Consultivo Escolar o Colegio de Profesores pa-
ra dirigir proceso de elección de ternas para designación de director";
   $Preguntas[20] = "persona poder ser director por ocasión tercera";
   $Preguntas[21] = "ser requiso deber  director de escuela , centro o unidad de enseñanza y de investigación , para designar a jefes
de división o de departamento a cargo";
   my $sr = "";
   for my $z (1 .. 1698) {
      $sr = $sr."A$z ";
   }
open(TXTO_1,"+> RSBdaseswL.txt") || die "Imposible crear ResultadosIRS.txt !!!\n";
for my $np (1 .. 21) {
#Inicialización Algoritmo Dijkstra
   for my $i (1 .. $nn) {
      if ($i != $ni) {
         $N_E{"A".$i} = 1e25;
         $N_E1{"A".$i} = 1e25;
         $N_E2{"A".$i} = "";
      }
      else{
```

```
            $N_E{"A".$ni} = 0;
            $N_E1{"A".$ni} = 0;
            $N_E2{"A".$ni} = "";
         }
      }
   #Se transfiere a un hash la matriz de adyacencia
      my $status;
      my $value;
      my $db = new BerkeleyDB::Unknown
               -Filename => $filename,
               -Flags => DB_RDONLY,
               or die "Cannot open file $filename: $! $BerkeleyDB::Error\n" ;
         my @w1 = split /\s+/, $sr;
         for my $i (0 .. @w1 - 1) {
            for my $j (1 .. $n_arts) {
               $w1[$i] =~ m/\w(\d+)/;
               if ($1 <= $j) {
                  $status = $db->db_get($w1[$i]."#A".$j,$value);
               } else {
                  $status = $db->db_get("A".$j."#$w1[$i]",$value);
               }
               $hs{$w1[$i]."#A".$j} = $value;
               $hs{"A".$j."#$w1[$i]"} = $value;
            }
         }
      $status = $db->db_close() ;
      $sr = "";
      for my $i1 ($fq .. $sq) {
         for my $i2 (1 .. 1700) {
            $hs{"A".$i1."#A".$i2} = 1e15;
            $hs{"A".$i2."#A".$i1} = 1e15;
         }
      }
      my $qst = $Preguntas[$np];
      print TXTO_1"P $qst\n";
      print "Pregunta $np\n\n$qst\n\n";
   #Procesamiento Pregunta -> Solicitud SR
   #Convertir a minúsculas
   #Se eliminan ¿ ? , . / %
      $qst = lc($qst);
      $qst =~ s/\¿//g;
      $qst =~ s/\?//g;
      $qst =~ s/\,//g;
      $qst =~ s/\.//g;
      $qst =~ s/\// /g;
      $qst =~ s/\%//g;
      my @tq;
      my $s = 0;
      my $qry = "";
   #Se eliminan tipos que no existen en los reglamentos
```

```
   my @w = split /\s+/, $qst;
   for my $i (0 .. @w - 1) {
      if ($tyt{$w[$i]}) {
         $tq[$s++] = $w[$i];
         $qry = $qry."$w[$i] ";
      }
   }
#Se divide en dos Solicitud SR. Si el total de tipos es impar un tipo más se coloca en el primer nodo del grafo.
#Se colocan en $p_a y $p_b que serán los nuevos nodos del grafo
   my $sizeq = @tq;
   my $df = $sizeq % 2;
   if ($df) {
      $df = $sizeq / 2;
      my @tem = split /\./, $df;
      $df = $tem[0] + 1;
   } else {
      $df = $sizeq / 2;
   }
   my $p_a = "";
   my $p_b = "";
   for my $z (0 .. $df - 1) {
      $p_a = $p_a.$tq[$z]." ";
   }
   for my $y ($df .. @tq - 1) {
      $p_b = $p_b.$tq[$y]." ";
   }
   print TXTO_1"Q $p_a\t&\t$p_b\n\n";
#Se otiene lista de tipos y frecuencia de las dos
#partes de la Solicitud SR
   my %tipq_a;
   my %tipq_b;
   my @ws_a = split /\s+/, $p_a;
   for my $i (0 .. @ws_a - 1) {
      $tipq_a{$ws_a[$i]} ++;
   }
   my @ws_b = split /\s+/, $p_b;
   for my $i (0 .. @ws_b - 1) {
      $tipq_b{$ws_b[$i]} ++;
   }
   my $p_1 = '';
   my $p_2 = '';
   foreach my $llave (keys %tipq_a) {
      $p_1 = $p_1.$llave." ";
   }
   foreach my $llave (keys %tipq_b) {
      $p_2 = $p_2.$llave." ";
   }
#Se obtiene el total de artículos en los que el aparece el término i-ésimo
#de query1 para asignarle peso a cada uno y agregarlo al hash de tipos por
#artículo $ppxa para post. compararlo con los arts. del reg.
```

```
   my $t_1 = "";
   my $t_2 = "";
   my @t = split /\s+/, $p_1;
   my @t1 = split /\s+/, $p_2;
   my $size = @t;
   my $size1 = @t1;
   for my $v (0 .. @t - 1) {
      my $a1 = ($tipq_a{$t[$v]}/($size))*(log($n_arts/$tyt{$t[$v]}))/log(10);
      $t_1 = $t_1.$t[$v]."|".$a1."#";
   }
   for my $v1 (0 .. @t1 - 1) {
      my $a2 = ($tipq_b{$t1[$v1]}/($size1))*(log($n_arts/$tyt{$t1[$v1]}))/log(10);
      $t_2 = $t_2.$t1[$v1]."|".$a2."#";
   }
#Se agregan query1 y query2 a $ppxa
   $ppxa{'A'.$fq} = $t_1;
   $ppxa{'A'.$sq} = $t_2;
#Se comparan los nuevos artículos (query1 y query2) con el resto y se agregan
#como nuevos nodos a la matriz de adyacencia
#$ppxa1{artículo j} = $tipo(1):fartj|NTTArtj|freg|farts|tf*idf# ... #$tipo(k):fartj|NTTArtj|freg|farts|tf*idf#
#Comparación de vectores por la medida del coseno
   for my $i1 ($fq .. $sq) {
      my $stc1 = "A".$i1;
      my %v_tc1;
      my $sc1 = 0;
      my @tmp = split /\#/, $ppxa{$stc1};
      foreach my $x (0 .. @tmp - 1) {
         my @tmp1 = split /\|/, $tmp[$x];
         $v_tc1{$tmp1[0]} = $tmp1[1];
         $sc1 = $sc1 + $tmp1[1] * $tmp1[1];
      }
      delete $v_tc1{''};
      for my $num_art2 (1 .. $sq) {
         my $stc2 = "A".$num_art2;
         my %v_tc2;
         my $sc2 = 0;
         @tmp = split /\#/, $ppxa{$stc2};
         foreach my $x (0 .. @tmp - 1) {
            my @tmp1 = split /\|/, $tmp[$x];
            $v_tc2{$tmp1[0]} = $tmp1[1];
            $sc2 = $sc2 + $tmp1[1] * $tmp1[1];
         }
         delete $v_tc2{''};
         my $pp = 0;
         foreach my $llave (keys %v_tc1) {
           if ($v_tc1{$llave} && $v_tc2{$llave}) {
            $pp = $pp + $v_tc1{$llave} * $v_tc2{$llave};
           }
         }
         my $sim = 0;
```

```
            $sim = $pp/(sqrt($sc2)*sqrt($sc1));
            my $expt = "A".$i1."#A".$num_art2;
            my $expt1 = "A".$num_art2."#A".$i1;
            if ($sim != 0) {
               $hs{$expt} = sprintf( "%.3f", 1/(100*$sim));
               $hs{$expt1} = $hs{$expt};
            } else {
               $hs{$expt} = 1e15;
               $hs{$expt1} = 1e15;
            }
         }
      }
      delete $hs{''};
#Obtiene lista de articulos relacionados por cada artículo
      for my $i1 (1 .. $sq) {
         my $re = "";
         for my $i2 (1 .. $sq) {
            if ($hs{'A'.$i1.'#A'.$i2} < 0.2) {
               $re = $re."A".$i2."#";
            }
         }
         $AR[$i1] = $re;
      }
#Se ejecuta algoritmo Dijkstra sobre la matriz de adyacencia
#nodo inicial: query1. nodo final: query2
      do {
         my $PN;
      A: foreach my $llave (sort {$N_E1{$b} <= $N_E1{$a}} keys %N_E1){
            $PN = $llave;
            delete $N_E1{$PN};
            last A;
         }
         $PN =~ m/A(\d+)/;
         my @tmp = split /\#/, $AR[$1];
         for my $x (0 .. @tmp - 1) {
            my $vt = $N_E{$PN} + $hs{$PN.'#'.$tmp[$x]};
            if ($vt < $N_E{$tmp[$x]}) {
               $N_E{$tmp[$x]} = $vt;
               $N_E1{$tmp[$x]} = $vt;
               $N_E2{$tmp[$x]} = $N_E2{$PN}.$PN." ".$tmp[$x]." ";
            }
         }
      } while ($N_E1{"A".$nf});
      $p = $N_E2{"A".$nf};          #Ruta de peso mínimo
      $c = $N_E{"A".$nf};           #Costo
      print "$p\t$c\n";
      my $cre = 1;
      my @n = split /\s+/, $p;
      if ($n[0] eq 'A'.$ni && $n[1] eq 'A'.$nf || $c > 1e15) {
         $hs{'A'.$ni.'#A'.$nf} = 1e15;
```

```
   } else {
      my $nr = 0;
#Retorno de resultados en la posición inicial de un arreglo. $res[0] = 'ni->n1->n2->...->nn->nf'
      my @res;
      for my $u (1 .. @n - 2) {
         if ($n[$u] ne $n[$u-1]) {
            $res[$nr++] = $n[$u];
            $n[$u] =~ m/\w(\d+)/;
            print TXTO_1 "\t",$cre++,"\t";
            $sr = $sr."$n[$u] ";
            if ($1 <= 107) {
            print TXTO_1 "A$1\tRP\n";
            } else {
################################################################################
#Se obtiene el número de artículo y documento al cual corresponde el vértice dado
            if ($1 <= 107) {
            print TXTO_1 "A$1\tRP\n";
            } else {
            if ($1 > 107 && $1 < 403) {
            my $vtem_1 = $1 - $nta_r1;
            print TXTO_1 "A$vtem_1\tR Interno\n";
            } else {
            if ($1 > 402 && $1 < 437) {
            my $vtem_1 = $1 - $nta_r1 - $nta_r2;
            print TXTO_1 "A$vtem_1\tLey Orgánica\n";
            } else {
            if ($1 > 436 && $1 < 521) {
            my $vtem_1 = $1 - $nta_r1 - $nta_r2 - $nta_r3;
            print TXTO_1 "A$vtem_1\tR Orgánico\n";
            } else {
            if ($1 > 520 && $1 < 555) {
            my $vtem_1 = $1 - $nta_r1 - $nta_r2 - $nta_r3 - $nta_r4;
            print TXTO_1 "A$vtem_1\tR Academias\n";
            } else {
            if ($1 > 554 && $1 < 627) {
            my $vtem_1 = $1 - $nta_r1 - $nta_r2 - $nta_r3 - $nta_r4 - $nta_r5;
            print TXTO_1 "A$vtem_1\tRAHIPN\n";
            } else {
            if ($1 > 626 && $1 < 658) {
            my $vtem_1 = $1 - $nta_r1 - $nta_r2 - $nta_r3 - $nta_r4 - $nta_r5 - $nta_r6;
            print TXTO_1 "A$vtem_1\tRBECASCONACYT\n";
            } else {
            if ($1 > 657 && $1 < 691) {
            my $vtem_1 = $1 - $nta_r1 - $nta_r2 - $nta_r3 - $nta_r4 - $nta_r5 - $nta_r6 - $nta_r7;
            print TXTO_1 "A$vtem_1\tRBECESTYOMAPOYO\n";
            } else {
            if ($1 > 690 && $1 < 745) {
            my $vtem_1 = $1 - $nta_r1 - $nta_r2 - $nta_r3 - $nta_r4 - $nta_r5 - $nta_r6 - $nta_r7 - $nta_r8;
            print TXTO_1 "A$vtem_1\tRCGC\n";
            } else {
```

```
            if ($1 > 744 && $1 < 889) {
            my $vtem_1 = $1 - $nta_r1 - $nta_r2 - $nta_r3 - $nta_r4 - $nta_r5 - $nta_r6 - $nta_r7 - $nta_r8 - $nta_r9;
            print TXTO_1 "A$vtem_1\tRCGTPNoD\n";
            } else {
            if ($1 > 888 && $1 < 1042) {
            my $vtem_1 = $1 - $nta_r1 - $nta_r2 - $nta_r3 - $nta_r4 - $nta_r5 - $nta_r6 - $nta_r7 - $nta_r8 - $nta_r9 - $nta_r10;
            print TXTO_1 "A$vtem_1\tRCITPA\n";
            } else {
            if ($1 > 1041 && $1 < 1081) {
            my $vtem_1 =$1-$nta_r1-$nta_r2-$nta_r3 - $nta_r4 - $nta_r5 - $nta_r6 - $nta_r7 - $nta_r8 - $nta_r9 - $nta_r10 - $nta_r11;
            print TXTO_1 "A$vtem_1\tRCOTEPABE\n";
            } else {
            if ($1 > 1080 && $1 < 1123) {
            my $vtem_1 = $1-$nta_r1-$nta_r2-$nta_r3-$nta_r4-$nta_r5-$nta_r6-$nta_r7-$nta_r8-$nta_r9 - $nta_r10 - $nta_r11 - $nta_r12;
            print TXTO_1 "A$vtem_1\tRDIPLOMADOS\n";
            } else {
            if ($1 > 1122 && $1 < 1162) {
           my $vtem_1=$1-$nta_r1-$nta_r2-$nta_r3-$nta_r4-$nta_r5-$nta_r6-$nta_r7-$nta_r8-$nta_r9-$nta_r10-$nta_r11-$nta_r12-$nta_r13;
            print TXTO_1 "A$vtem_1\tRDISTM\n";
            } else {
            if ($1 > 1161 && $1 < 1190) {
            my $vtem_1 = $1 - $nta_r1 - $nta_r2 - $nta_r3 - $nta_r4 - $nta_r5 - $nta_r6 - $nta_r7 - $nta_r8 - $nta_r9 - $nta_r10 -
$nta_r11 - $nta_r12 - $nta_r13 - $nta_r14;
            print TXTO_1 "A$vtem_1\tREVAL\n";
            } else {
            if ($1 > 1189 && $1 < 1218) {
            my $vtem_1 = $1 - $nta_r1 - $nta_r2 - $nta_r3 - $nta_r4 - $nta_r5 - $nta_r6 - $nta_r7 - $nta_r8 - $nta_r9 - $nta_r10 -
$nta_r11 - $nta_r12 - $nta_r13 - $nta_r14 - $nta_r15;
            print TXTO_1 "A$vtem_1\tRICOFFA\n";
            } else {
            if ($1 > 1217 && $1 < 1264) {
            my $vtem_1 = $1 - $nta_r1 - $nta_r2 - $nta_r3 - $nta_r4 - $nta_r5 - $nta_r6 - $nta_r7 - $nta_r8 - $nta_r9 - $nta_r10 -
$nta_r11 - $nta_r12 - $nta_r13 - $nta_r14 - $nta_r15 - $nta_r16;
            print TXTO_1 "A$vtem_1\tRPESTIMULOSDOCENTE\n";
            } else {
            if ($1 > 1263 && $1 < 1287) {
            my $vtem_1 = $1 - $nta_r1 - $nta_r2 - $nta_r3 - $nta_r4 - $nta_r5 - $nta_r6 - $nta_r7 - $nta_r8 - $nta_r9 - $nta_r10 -
$nta_r11 - $nta_r12 - $nta_r13 - $nta_r14 - $nta_r15 - $nta_r16 - $nta_r17;
            print TXTO_1 "A$vtem_1\tRPLAN\n";
            } else {
            if ($1 > 1286 && $1 < 1333) {
            my $vtem_1 = $1 - $nta_r1 - $nta_r2 - $nta_r3 - $nta_r4 - $nta_r5 - $nta_r6 - $nta_r7 - $nta_r8 - $nta_r9 - $nta_r10 -
$nta_r11 - $nta_r12 - $nta_r13 - $nta_r14 - $nta_r15 - $nta_r16 - $nta_r17 - $nta_r18;
            print TXTO_1 "A$vtem_1\tRTITULACION\n";
            } else {
            if ($1 > 1332 && $1 < 1367) {
            my $vtem_1 = $1 - $nta_r1 - $nta_r2 - $nta_r3 - $nta_r4 - $nta_r5 - $nta_r6 - $nta_r7 - $nta_r8 - $nta_r9 - $nta_r10 -
$nta_r11 - $nta_r12 - $nta_r13 - $nta_r14 - $nta_r15 - $nta_r16 - $nta_r17 - $nta_r18 - $nta_r19;
            print TXTO_1 "A$vtem_1\tRPRACYVISITAS\n";
            } else {
```

```
            if ($1 > 1366 && $1 < 1428) {
            my $vtem_1 = $1 - $nta_r1 - $nta_r2 - $nta_r3 - $nta_r4 - $nta_r5 - $nta_r6 - $nta_r7 - $nta_r8 - $nta_r9 - $nta_r10 -
$nta_r11 - $nta_r12 - $nta_r13 - $nta_r14 - $nta_r15 - $nta_r16 - $nta_r17 - $nta_r18 - $nta_r19 - $nta_r20;
            print TXTO_1 "A$vtem_1\tRED\n";
            } else {
            if ($1 > 1427 && $1 < 1513) {
            my $vtem_1 = $1 - $nta_r1 - $nta_r2 - $nta_r3 - $nta_r4 - $nta_r5 - $nta_r6 - $nta_r7 - $nta_r8 - $nta_r9 - $nta_r10 -
$nta_r11 - $nta_r12 - $nta_r13 - $nta_r14 - $nta_r15 - $nta_r16 - $nta_r17 - $nta_r18 - $nta_r19 - $nta_r20 - $nta_r21;
            print TXTO_1 "A$vtem_1\tRRVOE\n";
            } else {
            if ($1 > 1512 && $1 < 1570) {
            my $vtem_1 = $1 - $nta_r1 - $nta_r2 - $nta_r3 - $nta_r4 - $nta_r5 - $nta_r6 - $nta_r7 - $nta_r8 - $nta_r9 - $nta_r10 -
$nta_r11 - $nta_r12 - $nta_r13 - $nta_r14 - $nta_r15 - $nta_r16 - $nta_r17 - $nta_r18 - $nta_r19 - $nta_r20 - $nta_r21 - $nta_r22;
            print TXTO_1 "A$vtem_1\tRSIBE\n";
            } else {
            if ($1 > 1569 && $1 < 1595) {
            my $vtem_1 = $1 - $nta_r1 - $nta_r2 - $nta_r3 - $nta_r4 - $nta_r5 - $nta_r6 - $nta_r7 - $nta_r8 - $nta_r9 - $nta_r10 -
$nta_r11 - $nta_r12 - $nta_r13 - $nta_r14 - $nta_r15 - $nta_r16 - $nta_r17 - $nta_r18 - $nta_r19 - $nta_r20 - $nta_r21 - $nta_r22 -
$nta_r23;
            print TXTO_1 "A$vtem_1\tRSS\n";
            } else {
            if ($1 > 1594 && $1 < 1673) {
            my $vtem_1 = $1 - $nta_r1 - $nta_r2 - $nta_r3 - $nta_r4 - $nta_r5 - $nta_r6 - $nta_r7 - $nta_r8 - $nta_r9 - $nta_r10 -
$nta_r11 - $nta_r12 - $nta_r13 - $nta_r14 - $nta_r15 - $nta_r16 - $nta_r17 - $nta_r18 - $nta_r19 - $nta_r20 - $nta_r21 - $nta_r22 -
$nta_r23 - $nta_r24;
            print TXTO_1 "A$vtem_1\tREE\n";
            } else {
            my $vtem_1 = $1 - $nta_r1 - $nta_r2 - $nta_r3 - $nta_r4 - $nta_r5 - $nta_r6 - $nta_r7 - $nta_r8 - $nta_r9 - $nta_r10 -
$nta_r11 - $nta_r12 - $nta_r13 - $nta_r14 - $nta_r15 - $nta_r16 - $nta_r17 - $nta_r18 - $nta_r19 - $nta_r20 - $nta_r21 - $nta_r22 -
$nta_r23 - $nta_r24 - $nta_r25;
            print TXTO_1 "A$vtem_1\tRDEC\n";
            }}}}}}}}}}}}}}}}}}}}}}}}}
         }
#########################################################
      for my $nr (0 .. @res - 1) {
         for my $i1 (1 .. $sq) {
            my $expt = $res[$nr]."#A".$i1;
            my $expt_1 = "A".$i1."#".$res[$nr];
            $hs{$expt} = 1e15;
            $hs{$expt_1} = 1e15;
         }
      }
   }

W: while ($c < 1e15) {
      if ($cre > 101) {last W;}
      for my $i (1 .. $nn) {                          #Inicialización Algoritmo Dijkstra
         if ($i != $ni) {
            $N_E{"A".$i} = 1e25;
            $N_E1{"A".$i} = 1e25;
```

```
            $N_E2{"A".$i} = "";
         } else{
            $N_E{"A".$ni} = 0;
            $N_E1{"A".$ni} = 0;
            $N_E2{"A".$ni} = "";
         }
      }
      do {
         my $PN;
      A:  foreach my $llave (sort {$N_E1{$b} <= $N_E1{$a}} keys %N_E1){
            $PN = $llave;
            delete $N_E1{$PN};
            last A;
         }
         $PN =~ m/A(\d+)/;
         my @tmp = split /\#/, $AR[$1];
         for my $x (0 .. @tmp - 1) {
            my $vt = $N_E{$PN} + $hs{$PN.'#'.$tmp[$x]};
            if ($vt < $N_E{$tmp[$x]}) {
               $N_E{$tmp[$x]} = $vt;
               $N_E1{$tmp[$x]} = $vt;
               $N_E2{$tmp[$x]} = $N_E2{$PN}.$PN." ".$tmp[$x]." ";
            }
         }
      } while ($N_E1{"A".$nf});
      $p = $N_E2{"A".$nf};
      $c = $N_E{"A".$nf};
      print "$p\t$c\n";
      my @n = split /\s+/, $p;                                  #$res[0] = 'ni->n1->n2->...->nn->nf'
      my $nr = 0;
      my @res;                                                  #Retorno de resultados en la posición inicial de un arreglo
      if ($n[0] eq 'A'.$ni && $n[1] eq 'A'.$nf) {
         $hs{'A'.$ni.'#A'.$nf} = 1e15;
      } else {
         if ($c < 1e15) {
            my $nr = 0;
            my @res;                                            #Retorno de resultados en la posición inicial de un arreglo
            for my $u (1 .. @n - 2) {
               if ($n[$u] ne $n[$u-1]) {
                  $res[$nr++] = $n[$u];
                  $n[$u] =~ m/\w(\d+)/;
                  print TXTO_1 "\t",$cre++,"\t";
                  $sr = $sr."$n[$u] ";
##############################################################################
#Se obtiene el número de artículo y documento al cual corresponde el vértice dado
#En esta parte va el mismo código que se encuentra en la hoja anterior
##############################################################################
            for my $nr (0 .. @res - 1) {
               for my $i1 (1 .. $sq) {
                  my $expt = $res[$nr]."#A".$i1;
```

```
                    my $expt_1 = "A".$i1."#".$res[$nr];
                    $hs{$expt} = 1e15;
                    $hs{$expt_1} = 1e15;
                }
            }
        }
    }
    print TXTO_1"\n\n";
    print "\n\n";
}
close(TXTO_1);
dbmclose %tyt;
#*****************************************************************************
#SUB RUTINAS
#*****************************************************************************
sub obt_catcp {
    my $n_arch = shift;
    my $line;
    my %A;
    my $aux;
    open ('filhtxt', $n_arch) || die "\nEl archivo no puede abrirse\n\n";
    while ($line = <filhtxt>) {
        chomp ($line);
        if ($line =~ m/^(\w+)/) {
            $aux = $1;
            $A{$aux} = "";
        } else {
            $line =~ s/\s//g;
            $A{$aux} = $A{$aux}.$line."#";
        }
    }
    close ('filhtxt');
    delete $A{''};
    return (%A);
}
```