INSTITUTO POLITÉCNICO NACIONAL

CENTRO DE INVESTIGACIÓN EN COMPUTACIÓN

RECUPERACIÓN DE INFORMACIÓN
PARA RESPUESTA A PREGUNTAS EN DOCUMENTOS LEGALES

T E S I S
QUE PARA OBTENER EL GRADO DE
D O C T O R E N C I E N C I A S
P R E S E N T A:

M. EN C. ALFREDO LÓPEZ MONROY

DIRECTORES DE TESIS:

DR. ALEXANDER GELBUKH
DR. FRANCISCO HIRAM CALVO CASTRO

"LA TÉCNICA AL SERVICIO DE LA PATRIA"

México D.F., Enero de 2013

SIP-14 BIS

INSTITUTO POLITÉCNICO NACIONAL

SECRETARÍA DE INVESTIGACIÓN Y POSGRADO

ACTA DE REVISIÓN DE TESIS

En la Ciudad de México, D.F. siendo las 10:00 a.m. del día 28 del mes de Noviembre de 2012 se reunieron los miembros de la Comisión Revisora de la Tesis, designada por el Colegio de Profesores de Estudios de Posgrado e Investigación del:

Centro de Investigación en Computación

para examinar la tesis titulada:

"RECUPERACIÓN DE INFORMACIÓN PARA RESPUESTA A PREGUNTAS EN DOCUMENTOS LEGALES"

Presentada por el alumno:

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

Con registro: A 0 9 0 4 9 2

aspirante de: **DOCTORADO EN CIENCIAS DE LA COMPUTACIÓN**

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

LA COMISIÓN REVISORA

Directores de tesis

Dr. Alexander Gelbukh

Dr. Francisco Hiram Calvo Castro

Dr. Grigori Sidorov

Dr. Miguel Jesús Torres Ruiz

Dra. Sofía Natalia Galicia Haro

PRESIDENTE DEL COLEGIO DE PROFESORES

Dr. Luis Alfonso Villa Vargas

ESTADOS UNIDOS MEXICANOS
INSTITUTO POLITECNICO NACIONAL
CENTRO DE INVESTIGACION EN COMPUTACION
DIRECCION

INSTITUTO POLITÉCNICO NACIONAL

SECRETARÍA DE INVESTIGACIÓN Y POSGRADO

CARTA CESIÓN DE DERECHOS

En la Ciudad de **México** el día **12** del mes **Diciembre** del año **2012**, el (la) que suscribe **M. en C. Alfredo López Monroy** alumno (a) del Programa de **Doctorado en Ciencias de la Computación** con número de registro **A090492**, adscrito a **Centro de Investigación en Computación**, manifiesta que es autor (a) intelectual del presente trabajo de Tesis bajo la dirección de **Dr. Alexander Gelbukh y Dr. Francisco Hiram Calvo Castro** y cede los derechos del trabajo intitulado **Recuperación de información para respuesta a preguntas en 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.

M. en C. Alfredo López Monroy

Nombre y firma

A mí amada esposa: Georgina García Pacheco.

Agradecimientos

Al pueblo de México. A mis asesores y maestros por compartir sus conocimientos. A mi familia por su apoyo. A mis amigos por su compañía.

RESUMEN

Generalmente, los textos de corpus legales están fuertemente relacionados entre sí al grado de incluir referencias de uno a otro. Esto dificulta su consulta debido a que para satisfacer una necesidad de información, podría ser necesario complementar disposiciones de un texto con las de otros documentos. Por lo tanto, el objetivo de este trabajo es ayudar a localizar disposiciones legales mediante el desarrollo de un modelo de recuperación de información, el cual, a partir de una solicitud de información expresada en forma de una pregunta en lenguaje natural, proporcione un conjunto de artículos que la satisfagan. El modelo que se propone se basa en un grafo ponderado no dirigido. Se realizó una evaluación comparativa experimental de una implementación del modelo propuesto, para lo cual se utilizó un conjunto de 37,153 disposiciones y 40 preguntas con sus respectivas disposiciones-respuesta. Los resultados se compararon con los obtenidos con modelos de recuperación de información del estado del arte. La evaluación del desempeño del modelo propuesto muestra que la consideración de las referencias entre las disposiciones de textos legales puede ayudar en la mejora del actual acceso a esta clase de información.

ABSTRACT

The diversity and complexity of legal texts such as legislation, regulations, etc., make understanding and retrieval of its provisions a non-trivial task. One of the issues is the fact that the regulatory provisions tend to contain a large number of references to other provisions, which are cumbersome to follow and even more, because of these, it may often be necessary to combine several provisions from different documents to solve a legal problem. Due the development of the WWW retrieval algorithms that use the link structure of the Web in computing the importance or authority of a Web page, have been developed and used by various search engines. Since similar texts, case law documents, are also full of implicit and explicit references, some recent research has explored the use of link analysis for this kind of information. Therefore, this work describes a model developed to aid in the recovery of related provisions from different regulatory-legislative texts. Specifically, the aim is to identify the most strongly related provisions between regulations that satisfy an information need expressed as natural language question. To evaluate the performance of the proposed approach was conducted a comparative-experimental study based on an implementation of the model using techniques and methods from the areas of Natural Language Processing, Information Retrieval and Graph Theory. Performance evaluation of the proposed approach shows that consideration of the references between the provisions of legal texts can help improve the current access to this kind of information

Contenido

RESUMEN v
ABSTRACT vi
ÍNDICE DE FIGURAS x
ÍNDICE DE TABLAS xi
INTRODUCCIÓN 1
Contribuciones principales 5
Organización del trabajo 5
1. Estado del arte 6
1.1. Recuperación de información 6
1.1.1. Conceptos básicos 7
1.1.2. Modelo booleano 10
1.1.2.1. El MB y la documentación legal 12
1.1.3. Modelo del Espacio Vectorial 13
1.1.3.1. El MEV y la documentación legal: Lucene 16
1.1.4. Recuperación de pasajes 16
1.1.4.1. La RP y la documentación legal 17
1.2. Lenguajes de Marcado de Textos Electrónicos 17
1.2.1. Los lenguajes de marcado y la documentación legal. 19
1.3. Análisis de referencias 21
1.3.1. El Análisis de Referencias y la documentación legal 23
1.4. Combinación 24
2. Marco teórico 28
2.1. Teoría de grafos 28
2.2. Preprocesamiento 34
2.2.1. *Stop-words* 34
2.2.2. Lematización 35
2.2.3. *Stemming* 36
2.3. IRS 36
2.3.1. Lucene 37

2.3.2. JIRS 38
2.4. Modelo de combinación lineal 39
3. Modelo propuesto y estudio experimental 41
3.1. Descripción del modelo propuesto 41
3.1.1. Representación 42
3.1.2. Recuperación 43
3.2. Estudio experimental 43
3.2.1. Corpus 43
3.2.1.1. Documentos 44
3.2.1.2. Preguntas/Respuestas 45
3.2.2. Mecanismo de evaluación 47
3.2.3. Experimento I. 47
3.2.3.1. Representación: Grafo de artículos con referencias 47
3.2.3.2. Recuperación: Algoritmo de la ruta mínima 48
3.2.3.3. Corpus: 2,162 artículos. Preguntas/Respuestas: 21. 49
3.2.3.4. Comparación MEV, grafo sin referencias. 49
3.2.4. Experimento II 50
3.2.4.1. Representación: Grafo de artículos con referencias 50
3.2.4.2. Recuperación: Combinación algoritmos de recorrido de grafos. 50
3.2.4.3. Corpus: 1,117, 2,162 y 8,987 artículos. Preguntas/Respuestas: 40. 53
3.2.4.4. Comparación Lucene, JIRS. 53
3.2.5. Experimentos III 53
3.2.5.1. Representación: Grafo de documentos. Grafo de artículos. 54
3.2.5.2. Recuperación. Documentos: Dijkstra. Artículos, Similitud y PageRank. 54
3.2.5.3. Corpus. 37,153 artículos. 40 preguntas/respuestas 54
3.2.5.4. Comparación. Lucene, JIRS. 54
4. Resultados y análisis 55
4.1. Experimentos I 55
4.2. Experimento II 59
4.3. Experimento III 72

Conclusiones 88
Referencias.. 90

ÍNDICE DE FIGURAS

Figura 1. Los rectángulos representan los datos del IRS y los óvalos sus procesos. 8
Figura 2. Proceso de recuperación en el MB. 11
Figura 3. Sistema de consulta de información legal INFOJUS. 12
Figura 4. Resultados devueltos por el sistema de consulta legal INFOJUS. 13
Figura 5. Representación de los documentos en el MEV. 14
Figura 6. Grafo representando a) una "red social", y b) páginas Web relacionadas entre sí mediante sus *hipervínculos*. 22
Figura 7. Grafo no dirigido. 28
Figura 8. Grafo bipartito. 29
Figura 9. Grafo: a) conexo, b) inconexo. 30
Figura 10. Grafo dirigido 31
Figura 11. Los óvalos representan las etapas que la modelo específica y los cuadros las salidas de cada proceso. 41
Figura 12. Representación propuesta: los vértices pueden representar los documentos completos de toda una colección o el conjunto de sus artículos. 42
Figura 13. Proceso de representación de la solicitud y recuperación de artículos 51
Figura 14. Resultados de la implementación del modelo propuesto en su versión del experimento III para la solicitud *ciencia y tecnología*. 83
Figura 15. Primeros 10 resultados devueltos por el sistema de consulta de información jurídica INFOJUS para la solicitud *ciencia AND tecnología.* 85
Figura 16. Parte del artículo 56 de la Ley de Ciencia y Tecnología devuelto por INFOJUS para la solicitud *ciencia AND tecnologia.* 86
Figura 17. Resultado devuelto por INFOJUS con palabras resaltadas para la solicitud *ciencia AND tecnologia*. 87

ÍNDICE DE TABLAS

Tabla 1. Número de referencias entre artículos encontradas en la colección de textos. 44
Tabla 2. Descripción experimentos. 52
Tabla 3. Resultados GCR, GSR y MEV en función del número de artículos-respuesta recuperados. 55
Tabla 4. Preguntas/artículos-respuesta: Grupo I. 57
Tabla 5. Preguntas/artículos-respuesta: Grupo II. 57
Tabla 6. Preguntas/artículos-respuesta: Grupo III. 58
Tabla 7. Preguntas/artículos-respuesta: Grupo IV 59
Tabla 8. Resultados SSim, SPR, Sem. & PR, Lucene y JIRS. experimento II. Colección A. 59
Tabla 9. Número de preguntas cuyos artículos-respuesta se encontraron dentro de las primeras *n* posiciones. Experimento II. Colección A. 60
Tabla 10. Resultados Sem. & PR., Lucene y JIRS. Experimento II. Colecciones B y C. 60
Tabla 11. Evaluación de la respuesta a preguntas del tipo I. 61
Tabla 12. Evaluación de la respuesta a preguntas tipo II. 65
Tabla 13. Evaluación de la respuesta a preguntas del tipo III. 68
Tabla 14. Evaluación de la respuesta a preguntas del tipo IV. 71
Tabla 15. Resultados globales para el conjunto de 37,153 artículos. 73
Tabla 16. Resultados del experimento III. Preguntas del tipo I. 74
Tabla 17. Resultados del experimento III. Preguntas del tipo II. 76
Tabla 18. Resultados del experimento III. Preguntas del tipo III. 78
Tabla 19. Resultados del experimento III. Preguntas del tipo IV 80
Tabla 20. Colecciones de documentos: implementación del modelo propuesto y el sistema de consulta de información jurídica INFOJUS. 81

INTRODUCCIÓN

La *documentación jurídica* se conforma por toda aquella información contenida en la *legislación*, *jurisprudencia* y *teoría del derecho* o cualquier otro documento con contenido jurídico relevante (Ríos-Estavillo, 1997; Moens, 2006).

Específicamente, los corpus de documentos *legislativos-normativos* se conforman por una amplia variedad de textos; entre ellos se encuentran leyes, reglamentos, lineamientos, políticas, etc. Esta clase de documentos representa una importante fuente de información debido a que contienen normas que regulan desde la conducta de individuos en sociedad hasta la organización y funcionamiento de instituciones gubernamentales y privadas. Por esta razón, su contenido debería ser accesible para su consulta a todo público; sin embargo, la diversidad y cantidad existente de documentos, dificultan este hecho, y por lo tanto, su conocimiento. En nuestro país, para fines de 2011 existían más de 300 leyes a nivel federal; mientras que en la normatividad-administrativa se utilizan actualmente en las más de 290 instituciones de la administración pública federal mexicana, aproximadamente 43 tipos de documentos entre leyes, acuerdos, decretos, bandos, códigos, reglamentos, oficios, etc.

Otro factor importante que impacta la accesibilidad es que las normas que conforman la parte esencial de un documento legal suelen estar fuertemente relacionadas entre sí, lo que se refleja a través de las múltiples referencias textuales (explícitas e implícitas) que generalmente contienen incluso a normas de otros documentos, el ejemplo (1) ilustra lo anterior.

(1) Referencias explícitas e implícitas en un texto legal.

Ley Orgánica del Instituto Politécnico Nacional (IPN). Artículo 14.- Son facultades y obligaciones del Director General: …XV. Nombrar a los secretarios de área, previa consulta al Secretario de Educación Pública, quienes deberán reunir los requisitos señalados en el artículo 13 de esta propia Ley;… XIX. Ejercitar la representación legal del Instituto con las más amplias facultades a que se refieren los dos primeros párrafos del artículo 2554 del Código Civil para el Distrito Federal en materia común y para toda la República en materia federal, y XX. *Las demás que prevean esta Ley y otros ordenamientos aplicables.*

Se ha encontrado (Kerrigan y Law, 2003) que estas referencias intradocumento e interdocumentos afectan la legibilidad de los textos legales dificultando su consulta, esto debido a que para satisfacer una necesidad de información podría ser necesario el reunir disposiciones pertenecientes a uno o más documentos. Esto se conoce como 'problema de síntesis del conocimiento'. En el cual, el contenido de diferentes documentos podría resultar irrelevante para un usuario; no obstante, de su uso combinado podría inferir una respuesta (Moens, 2006). Los ejemplos (2) y (3) ilustran este problema a partir de solicitudes de información formuladas como una pregunta en lenguaje natural y cuyas respuestas se infieren a partir de disposiciones de dos textos diferentes. Los ejemplos pertenecen a un conjunto de preguntas formuladas por la comunidad de una institución gubernamental, y respondidas por el personal de la oficina legal, a partir de la legislación y normatividad, de la misma institución.

(2) ¿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?

Artículo 182 del Reglamento Interno del IPN. Los directores y subdirectores de escuelas, centros y unidades de enseñanza y de investigación cesarán en sus funciones… II. Al término de la duración de su encargo, conforme a lo dispuesto por el artículo 21 de la Ley Orgánica, …	Artículo 21 de la Ley Orgánica del IPN. 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 período.

(3) ¿Tienen derecho de votar o no los profesores interinos adscritos a alguna escuela?

Artículo 208 del Reglamento Interno del IPN. El personal académico de cada programa académico, o equivalente elegirá en forma directa y por mayoría de votos a sus respectivos representantes, en los términos del artículo 27 de la Ley Orgánica.	Artículo 6 del Reglamento de las Condiciones Interiores de Trabajo del Personal Académico. Para la debida interpretación y aplicación del presente Reglamento de las Condiciones Interiores de Trabajo del Personal Académico, en el curso de este instrumento, se denominará: … XV. Personal Académico: A la persona que presta sus servicios al Instituto…, desempeñando trabajos académicos en los términos del presente reglamento… XXII. Personal académico interino: Es aquel que cubre licencia temporal del personal académico de base, hasta por seis meses…

En el ejemplo (2), las disposiciones-respuesta se relacionan por medio de una referencia textual, mientras en el ejemplo (3), las disposiciones-respuesta, no se relacionan entre sí de forma directa, pero ambas satisfacen la solicitud de información planteada. En estos ejemplos se muestra el problema que implica la localización y consulta de no sólo uno sino de varios documentos legislativos y/o normativos para hallar solución a una determinada situación a partir de las disposiciones que contiene.

De acuerdo a Ríos-Estavillo (1997) facilitar el conocimiento de la información legal requiere de su adecuada estructuración, organización y sistematización. Bajo este enfoque han surgido diversas propuestas. Por ejemplo, Turtle (1995), Moens (2006) y Geist (2009), han sugerido aprovechar las referencias interdocumentos e intradocumento que caracterizan a los textos legales en métodos computacionales para su almacenamiento y consulta. Con el mismo objetivo, Moens (2001), Monroy, Calvo y Gelbukh (2008, 2009), entre otros, han estudiado el utilizar la estructura que los documentos poseen.

A pesar de la diversidad de textos legales, su contenido se organiza de una manera específica, tal como se ilustra en el ejemplo (4).

(4) Estructura típica de un documento legal.

CE[1]		**México**[2]
Título	Título	LEY ORGÁNICA DEL IPN
Preámbulo	Preámbulo	TEXTO VIGENTE. Nueva Ley publicada…
Parte dispositiva	Disposiciones generales	Artículo 1. …es la institución educativa del Estado creada para consolidar, a través de la Educación… Artículo 34. …
Fórmulas finales	Disposiciones transitorias	ARTÍCULOS TRANSITORIOS PRIMERO.- Esta Ley entrará en vigor …

La mayoría de los documentos comienza con el título (nombre de la ley, reglamento, estatuto, etc.) seguido de un breve preámbulo el cual establece: el objetivo, fundamento legal, ámbito de aplicación, etc., del documento. Posteriormente, las normas orientadas a

[1] Partes de los documentos legislativos-normativos de la Comunidad Europea. López-Arroyo Belén. http://itastformacion.tel.uva.es/ [consulta 10-15-2012]

[2] Guía para emitir documentos normativos. Secretaría de la Función Pública, (SFP). México. http://www.normateca.gob.mx/NF_Secciones_Otras.php?T=23&D=1182&ND=SFP [consulta 14-11-2012]

regular la conducta de individuos son comúnmente agrupadas en una sección denominada disposiciones generales, mientras normas dirigidas a las autoridades que las han de aplicar se establecen en la sección denominada disposiciones transitorias (Huerta-Ochoa, 2001). A su vez estas últimas, así como las disposiciones generales, se dividen en uno o más artículos. También la parte principal de los documentos, las disposiciones generales, se organiza de una forma específica. Las disposiciones generales se dividen en uno o más artículos, como se ilustra en el ejemplo (5). A su vez éstos se dividen en uno o más párrafos y/o una o más fracciones.

(5) Organización típica de las disposiciones generales de un documento legal

Legislación mexicana	**Comunidad Europea**				
	Español	**Francés**	**Inglés**	**Alemán**	**Observaciones**
Subdivisión básica (parte dispositiva/disposiciones generales)					
Artículo	Artículo	Article	Article	Artikel	Numeración continua
Organización de los artículos de las disposiciones generales					
Libro	Parte	Partie	Part	Teil	Subdivisiones con o sin título para textos largos
Título	Título	Titre	Title	Titel	
Capítulo	Capítulo	Chapitre	Chapter	Kapitel	
Sección	Sección	Section	Section	Abschnitt	
Organización del contenido de los artículos de las disposiciones generales					
Apartado	Apartado	Paragraphe	Paragraph	Absatz	Numerado
Párrafo	Párrafo	Alinéa	Paragraph	Absatz	No numerado
Fracción	Punto	-	Subparagraph	Nummer	1), 2), …
Fracción	Letra	Point	Point	Buchstabe	a), b), …
Fracción	Inciso	-	-	Ziffer	i), ii), …
-	Guión	Tiret	Indent	Gedankenstrich	-
-	Frase	Phrase	Phrase	Satz	-

De acuerdo con lo anterior, en esta tesis se propone un modelo computacional de recuperación de información legal basado en las características que los diferentes tipos de documentos legislativos y normativos, bajo la suposición que esto mejorará la recuperación de información legal. El enfoque presentado en este trabajo consiste en un modelo basado en un grafo ponderado no dirigido, el cual se eligió debido a que refleja la estructura de los documentos y facilita el uso de las relaciones entre disposiciones. Para investigar el desempeño del modelo propuesto se llevó a cabo un estudio experimental de diferentes implementaciones del modelo utilizando técnicas del área de Procesamiento de Lenguaje Natural (PLN) y Recuperación de Información, entre otras. Adicionalmente, los resultados fueron comparados con los obtenidos con los principales enfoques del estado del arte, el Modelo del Espacio Vectorial, y los sistemas de recuperación de información Lucene y JIRS (*Java Information Retrieval System*).

Contribuciones principales

- La propuesta de un modelo para la recuperación de información legal basado en rasgos característicos de documentos legislativos-normativos como son su estructura y las referencias que suelen contener.
- Métodos para la implementación del modelo que aprovechan las características de los textos.

Organización del trabajo

El trabajo se organiza como sigue: en el Capítulo I, se realizó una reseña de literatura sobre métodos aplicados recientemente a documentación legal. En el Capítulo II, se proporcionaron conceptos básicos que faciliten la comprensión del modelo propuesto, el cual se describe en el Capítulo III, además de los aspectos principales de su implementación, los experimentos realizados y la forma de evaluación. En el cuarto Capítulo se presentan y analizan los resultados obtenidos y finalmente se proporcionan las principales conclusiones derivadas del estudio.

1. Estado del arte

El presente trabajo se enfoca básicamente en la tarea de recuperación de información (legal) textual a partir de una solicitud de información dada por un usuario. Sobre dicha tarea ha habido desde hace tiempo un amplio interés, por lo que la revisión de la literatura se divide en tres partes. En la sección 1.1, se examinan los principales enfoques del área de recuperación de información aplicados al área legal. Para ello se describen brevemente los modelos más representativos del área de Recuperación de Información con base en los procesos que generalmente especifican: el de representación del contenido de los documentos y el mecanismo de recuperación. En la sección 1.2, se da una breve descripción de investigación realizada sobre el empleo de técnicas basadas en los Lenguajes de Marcación de Textos Electrónicos, recientemente utilizados para el manejo de información en la WWW (*World Wide Web*). Una de las características principales de los textos legislativos y normativos es la organización jerárquica de su contenido y las referencias que poseen, por lo cual, en la sección 1.3, se revisan brevemente técnicas utilizadas en el denominado análisis de referencias y diferentes trabajos basados en la estructura de la Web. Finalmente, en la sección 1.4 se describen los trabajos realizados sobre la técnica de combinación o fusión en el área de Recuperación de Información y en el dominio legal.

1.1. Recuperación de información

Muchas de las aplicaciones orientadas al manejo de texto que actualmente se utilizan, tales como los modernos motores de búsqueda, no hubieran sido posibles sin los estudios realizados en el área de la Recuperación de Información, IR (por sus siglas en inglés). Una parte considerable de la tecnología desarrollada para el manejo de información textual ha requerido de no solo una exhaustiva experimentación sino también de un amplio trabajo teórico. La verificación empírica rigurosa es necesaria para el desarrollo de nuevas herramientas; sin embargo, si la experimentación no es guiada por la teoría se puede caer en el enfoque de prueba y error el cual hoy en día resulta insuficiente debido al constante surgimiento de nuevos retos y problemas (Hiemstra, 2009). Pero, ¿cuál es la teoría tras la cual subyacen las actuales herramientas de recuperación de información?,

desafortunadamente, por el momento, no existe una única respuesta a esta cuestión. Hasta ahora, se han desarrollado muchas teorías, o más bien lo que se ha denominado modelos formales. Cada modelo puede resultar de utilidad para el desarrollo de algunas herramientas, aunque no para todas. Sin embargo, en conjunto los diversos modelos permiten adquirir un mejor entendimiento sobre la recuperación de información, lo que repercute a su vez en el desarrollo de nuevas herramientas o en la mejora de las existentes. A fin de comprender la IR, es esencial un conocimiento adecuado de los diferentes modelos de recuperación hasta ahora desarrollados o al menos de los más importantes. En este capítulo se describieron algunos de los modelos más representativos de la IR relacionados con el presente trabajo.

1.1.1. Conceptos básicos

En su forma más básica un *Sistema de Recuperación de Información* o IRS (por sus siglas en inglés) es un programa computacional para la consulta y usualmente el almacenamiento de documentos. Generalmente, tales sistemas sólo ayudan a sus usuarios a localizar y recuperar la información que necesitan. Esto, en el sentido de que no devuelven en sí información o responden preguntas, sino que sólo informan sobre la existencia y ubicación de documentos que pudieran contener la información requerida. Algunos sistemas suelen adjuntar información extraída directamente de los textos. Sin embargo, la información proporcionada por tales sistemas, como por ejemplo Google, sólo consiste en un fragmento de texto en donde aparecen los términos de la solicitud. Por lo tanto, un usuario sólo tiene una mínima parte del contenido de los documentos como ayuda para decidir si le son interesantes. Aún más, el texto proporcionado no contiene información sobre si el documento es útil para la necesidad de información del usuario. Esto puede obligar al usuario a prácticamente leer cada documento recuperado hasta hallar información que satisfaga su necesidad de información, lo que puede convertirse en una rutina tediosa y aún más importante, se ha encontrado (Ferreira y Atkinson, 2009) que podría demandar la inversión de tiempo valioso para realizarse.

Los documentos sugeridos que eventualmente logran satisfacer la necesidad de información del usuario se denominan comúnmente como *documentos relevantes*. Un Sistema de

Recuperación de Información perfecto recuperaría sólo documentos relevantes y no los irrelevantes. Sin embargo, tal sistema no es posible debido a que principalmente la relevancia depende de la opinión subjetiva de los usuarios. En la práctica, dos usuarios podrían utilizar la misma solicitud de información en un sistema de recuperación de información y juzgar diferente la relevancia de los documentos devueltos por el sistema. Para algunos usuarios podrían ser de utilidad mientras que para otros no.

Con respecto al funcionamiento de un IRS, hay tres procesos básicos que el sistema debe realizar: el de la *representación del contenido de los documentos*, el de la *representación de la necesidad de información de los usuarios*, y el de la *recuperación* en sí misma, en la **Figura 1** se ilustran estos procesos y su interrelación.

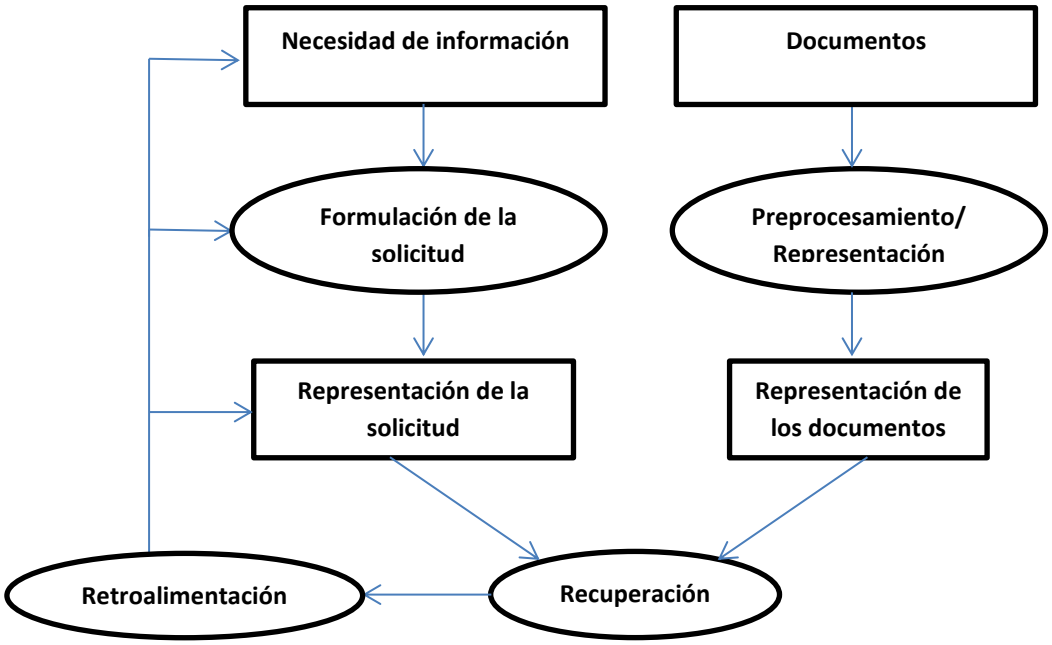


Figura 1. Los rectángulos representan los datos del IRS y los óvalos sus procesos.

Con respecto a los documentos a utilizar, éstos podrían encontrarse en diferentes formatos de almacenamiento como *Portable Document Format*, PDF, *HiperText Mark-up Language*, HTML, *eXtensible Mark-up Language*, XML debido a lo cual un IRS primero los convierte a un formato común, usualmente texto plano en alguna codificación estándar. Adicionalmente, un IRS realiza un procesamiento del contenido de los textos antes del proceso de representación.

A pesar de que la mayoría de sistemas de recuperación emplean las palabras como unidades lingüísticas mínimas, también se ha estudiado el uso de secuencias de palabras, entre otras opciones. Dependiendo de la unidad lingüística mínima elegida, es común que se extraigan y almacenen de una forma adecuada para su posterior uso, dichas unidades por documento o para toda la colección. Generalmente, se emplean algoritmos relativamente simples para esta tarea, como podría ser uno que sólo distinguiera entre palabras y signos de ortografía, eliminando éstos y convirtiendo a minúsculas las palabras.

Una vez que se tiene el contenido de los textos en la forma requerida un IRS lleva todo el contenido del documento (o sólo ciertos elementos del mismo) a una *representación* sobre la cual realizará el proceso de *recuperación*. Una característica del proceso de representación de los documentos es que generalmente este se realiza "fuera de línea", esto es, un proceso en el cual el usuario final del sistema de recuperación de información no está directamente involucrado, al igual que en el proceso de preprocesamiento de los documentos. En resumen, de estos procesos (preprocesamiento y representación) resulta una representación de los documentos que el IRS utilizará como base para el proceso de recuperación. Debido a que en el manejo tradicional de libros escritos una etapa consiste en crear un índice de las obras que facilite su almacenamiento y consulta, el preprocesamiento y representación de los documentos se denomina en conjunto como etapa de *indexado*.

El objetivo de los usuarios de un IRS es satisfacer una *necesidad de información*. El proceso de representar la necesidad de información de los usuarios se conoce también como etapa de *formulación de la solicitud de información*. La representación resultante se denomina simplemente *solicitud*. En un sentido amplio, la formulación de la solicitud podría denotar el diálogo completo de interacción entre el sistema y el usuario, algo que se ha sido estudiado con la finalidad de conducir a los usuarios a no sólo formular una mejor solicitud sino también, posiblemente, a comprender mejor su necesidad de información (Ferreira y Atkinson, 2009). Esto, en la Figura 1, está ilustrado como el proceso de *retroalimentación*.

El proceso de *recuperación* usualmente se basa en la comparación entre la solicitud y los documentos de la colección, más específicamente, entre sus correspondientes representaciones, resultando en una lista ordenada de documentos la cual los usuarios

podrán recorrer en forma descendente en busca de la información que necesiten. Los IRS que proporcionan como salida una lista ordenada de documentos, devolverán los documentos que considere relevantes dentro de las primeras posiciones; esto con el objetivo de disminuir el tiempo que un usuario podría invertir en la lectura de los documentos devueltos por el sistema. Actualmente existen una amplia variedad de algoritmos, algunos simples pero efectivos, que se basan desde la distribución de términos en cada documento de una colección, hasta estadísticas sobre el número de conexiones que tiene un documento con otros de la colección. El funcionamiento de muchos de los sistemas de recuperación de información se basa directa o indirectamente en modelos de recuperación de información. Estos modelos de recuperación generalmente especifican la representación de la solicitud de los usuarios como la de los documentos, así como el proceso de recuperación. Precisamente es que, con base en estas tres etapas, se describen los modelos relacionados con el presente trabajo. Primero se describe cada modelo y posteriormente se proporciona una breve revisión de trabajos en los cuales han sido utilizados en el dominio jurídico.

1.1.2. Modelo booleano

El Modelo Booleano, MB, a pesar de ser uno de los primeros modelos utilizados para la recuperación de información, debido a su simplicidad y resultados aceptables para muchas tareas, es aún hoy la base de funcionamiento de importantes IRS.

Representación

El modelo booleano se basa en la teoría clásica de conjuntos, pues los documentos son concebidos precisamente como conjuntos de términos, es decir: $D = \{D_1, D_2, \dots, D_n\}$ y $D_i = \{t_1, t_2, \dots, t_k\}$. Donde cada D_i es un documento de la colección D y cada t_j los términos del documento D_i.

Recuperación

Para la formulación de la solicitud se emplea la lógica de Boole. Básicamente, la solicitud de información se formula como una combinación de palabras y operadores de la lógica booleana (ejemplo 6).

(6) 3 diferentes ejemplos de formulaciones de solicitudes en el MB

social Y economía	
social O política	(social O política) Y (NO(social Y economía))

A partir de las indicaciones de la solicitud de información se forma un conjunto con los términos que contiene y con base en operaciones de conjuntos se recupera un conjunto de documentos como resultado final del proceso. En la **Figura 2,** se ilustra el resultado del proceso de recuperación para las solicitudes del ejemplo (6). Suponiendo 7 documentos (D_1, D_2,…, D_7) con los términos que se ilustran, cada conjunto se conforma de los documentos que contienen los términos de búsqueda: social, económica, política. Las áreas sombreadas representan los documentos que se obtienen con cada solicitud del ejemplo (6).

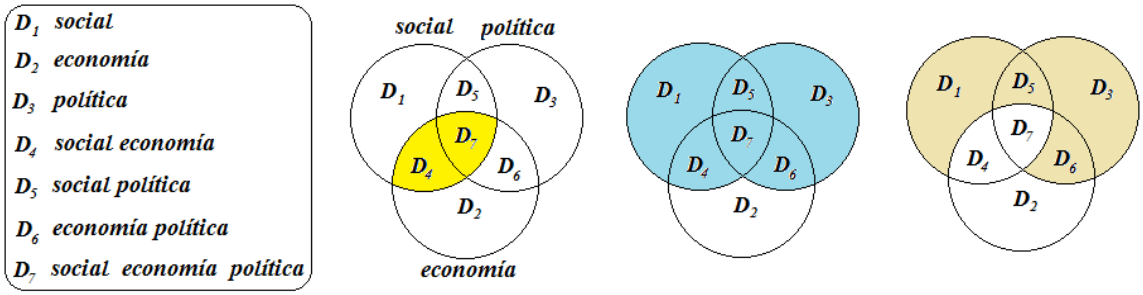


Figura 2. Proceso de recuperación en el MB.

Debido a su simplicidad, el modelo booleano presente diversos inconvenientes:

- Las solicitudes resultan ya sea en muy pocos (0) o una cantidad considerable de documentos (1000), a esto último se le ha denominado sobrecarga de información.
- Toma tiempo de aprendizaje el formular solicitudes que produzcan una cantidad manejable de textos, y aún, en tal caso, es común la situación anterior.
- No hay forma de ordenar los resultados.

Los sistemas actuales de recuperación basados en el MB no sólo emplean las operaciones lógicas (*AND*, *OR*, *NOT*). La recuperación basada sólo en una formulación de solicitudes con operaciones lógicas básicas y una salida de documentos no ordenada resulta insuficiente para muchas de las necesidades de información que los usuarios tienen; debido a esto, los sistemas implementan el denominado Modelo Booleano Extendido el cual incorpora operadores adicionales, por ejemplo el de proximidad. Un operador de proximidad (NEAR) es una forma de especificar que dos términos de una solicitud deben aparecer "cerca" uno del otro, donde la "cercanía" puede definirse, por ejemplo, a partir de restringir el número de palabras intermedias (Manning, Raghavan y Schtze, 2008).

1.1.2.1. El MB y la documentación legal

Hoy en día existen portales Web que proporcionan información legal a través de mecanismos todavía basados en el MB. Por ejemplo, Westlaw, uno de los principales proveedores de información legal en los EEUU, comenzó a dar servicio en 1975. En 2005, su búsqueda booleana ("términos" y "conectores") era aún la predeterminada y la utilizada por un alto número de sus usuarios, a pesar de que la búsqueda basada en la ordenación de documentos (denominada "lenguaje natural" por Westlaw) fue agregada en 1992 (Manning, *et al*. 2008). En nuestro país, el sistema de recuperación de información legal INFOJUS del Instituto de Investigaciones Jurídicas de la Universidad Nacional Autónoma de México ofrece la consulta de 279[3] leyes federales (**Figura 3**) desde simplemente seleccionar la visualización de un documento en particular a partir de su título, hasta una búsqueda basada en el MB (con los operadores básicos AND, OR y NOT y el operador extendido NEAR) la cual devuelve como resultados artículos o documentos completos (**Figura 4**).

Figura 3. Sistema de consulta de información legal INFOJUS.

[3] http://info4.juridicas.unam.mx/ijure/fed/ consulta [21-11-2012]

Figura 4. Resultados devueltos por el sistema de consulta legal INFOJUS.

1.1.3. Modelo del Espacio Vectorial

El Modelo Vectorial o comúnmente llamado del Espacio Vectorial (MEV) surgió con el objetivo de disminuir algunos de los inconvenientes del MB. Específicamente flexibiliza la formulación de la solicitud de información al no precisar de operadores de búsqueda y su mecanismo de recuperación se encarga de ordenar los textos de acuerdo a su relevancia con respecto a la solicitud de información.

Representación

En el MEV (Salton *et al*., 1975) consiste en formar un espacio vectorial de dimensión igual al número de términos (diferentes) en la colección. En este espacio cada documento se representa como un vector, el cual se forma a partir de valores asociados a los términos que contiene, **Figura 5**. De esta forma cada documento queda definido como: $d_M = (w_{1,M}, \dots, w_{n,M})$. Donde *n* es el número de términos en la colección de textos, y $w_{i,M}$ es un valor "de peso" o "ponderación" asociado del *i-ésimo* término en el documento *M*.

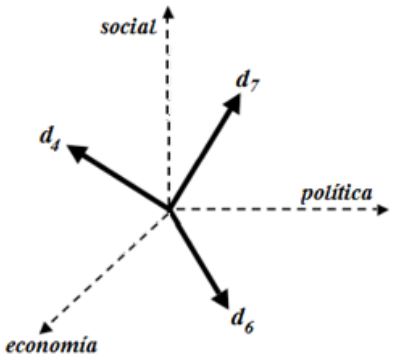


Figura 5. Representación de los documentos en el MEV.

Ponderación

Para obtener los valores asociados a los términos a partir de los cuales se forman los vectores que representaran a los documentos se han propuesto y estudiado diferentes métodos. El más simple, denominado, método binario consiste en determinar simplemente si un término i aparece o no en un documento M.

$$w(i,M) = \begin{cases} 1 \; si \; el \; término \; i \; aparece \; en \; el \; documento \; M \\ 0 \; en \; caso \; contrario \end{cases} \quad (1)$$

Se ha sugerido que no todos los términos que aparecen en un texto son igualmente importantes. Para considerar esta suposición, se propuso utilizar, en vez del método binario, la frecuencia de los términos en los documentos, es decir, el número de veces que un término aparece en un documento (Lunh, 1957). De esta forma, entre más frecuente es un término en un documento, mayor es su valor asociado.

$$w(i,M) = f_{i,M} \quad (2)$$

donde $f_{i,M}$ es la frecuencia del término i en el documento M.

A pesar de basarse en una suposición razonable, el simple conteo de términos presenta algunos problemas. El principal inconveniente que se ha documentado es que si bien un término frecuente en un texto puede ser importante en ese mismo texto, en el caso de ser igual de frecuente en el resto de documentos, dicho término podría no ser de utilidad para discriminar entre textos relevantes e irrelevantes. Orientada a tal inconveniente surgió la denominada frecuencia inversa del documento (Salton, 1988), definida como:

$$idf_{i,M}(i,N) = \log([N]/m\,(i,N)) \quad (3)$$

donde, $idf_{i,M}$ es la frecuencia inversa del documento, *idf* (por sus siglas en inglés) para el término *i* en el documento *M*, $[N]$ es el número de documentos en la colección N, y:

$$m(i,N) = f_{i,N} \tag{4}$$

donde $f_{i,N}$ es el número de documentos en la colección de documentos N que contienen al término i. Así el valor asociado al término i del documento *M* considerando su frecuencia en la colección de documentos N, está dado por:

$$w(i,M,N) = idf_{i,M}(i,N) \tag{5}$$

Mediante la *idf* si un término aparece en una cantidad menor de textos recibe un mayor peso en comparación de un término que aparezca en un mayor número de documentos. De esta forma la *idf* implementa la idea que si término resulta relativamente común en una colección de documentos, éste podría no ser de utilidad para distinguir entre documentos relevantes y no-relevantes.

Hoy en día, una de las técnicas más populares de ponderación es la que combina la frecuencia de un término en un documento, tf (por sus siglas en inglés), con su frecuencia inversa del documento (idf, inverse document frequency). Mientras tf mide la densidad de un término dentro de un documento, idf mide su frecuencia en todo el corpus. De esta forma $tfidf$, como se denomina a dicha combinación, concede un mayor valor a los términos menos frecuentes en la colección y a su vez más frecuentes en un documento (Brunzel, 2007). La ecuación que define esta medida está dada por:

$$w(i,M,N) = tfidf_{i,M}(i,M,N) = f_{i,M} * \log\left(\frac{N}{m(i,N)}\right) \tag{6}$$

Recuperación

En el MEV, una solicitud de información es tratada como un nuevo documento, es decir, de forma similar a éstos también se construye un vector a partir de los términos que la componen el que después se incorpora al mismo espacio vectorial. Posteriormente, la recuperación se lleva a cabo a través de evaluar el grado de similitud entre los vectores correspondientes a los documentos, y el vector correspondiente a la solicitud de información. Para medir el grado de similitud se utiliza comúnmente la correlación entre vectores, dada por el coseno del ángulo entre vectores (Baeza-Yates y Ribeiro-Neto, 1999).

$$sim_{U,V} = cos(d_U, d_V) = \frac{d_U \cdot d_V}{\|d_U\| * \|d_V\|} \quad (7)$$

donde $sim_{U,V}$ es la similitud entre los documentos *U* y *V* en el MEV, $d_U \cdot d_V$ corresponde al producto punto entre los vectores, generalmente normalizados, *U* y *V*, y $\|a\|$ denota la norma de los vectores.

1.1.3.1. El MEV y la documentación legal: Lucene.

Los métodos más conocidos en el área de Recuperación de Información: el llamado Modelo Booleano y el Modelo del Espacio Vectorial (MEV) han sido aplicados a una amplia variedad de documentos, entre ellos los legales (Moens, Uyttendaele y Dumortier, 1997; Osborn y Sterling, 1999; Schweighofer, Rauber & Dittenbach, 2001; van Engers, van Gog y Jacobs, 2005). Por ejemplo, Lucene (popular software de IR) basado en los modelos anteriores, se ha utilizado recientemente en el dominio legal (Peñas, Forner, Sutcliffe, Rodrigo, Forăscu, Alegria, Giampiccolo, Moreau y Osenova, 2009).

Básicamente, Lucene, emplea el MB y el MEV para recuperar un conjunto de documentos relevantes para una solicitud de información expresada a partir de un conjunto de términos clave. Debido a su buen desempeño, éste es actualmente utilizado para proporcionar acceso a diferentes tipos de documentos legales pertenecientes a la Organización de la Naciones Unidas para la alimentación y la Agricultura[4]).

Uno de los inconvenientes del MB no resuelto totalmente por el MEV es el de la sobrecarga de información. Aún los IRS basados en estos modelos suelen devolver o muy pocos o una cantidad tal de documentos que puede abrumar al usuario. Problema que también se ha observado en el área legal, (Daniels & Rissland, 1997; Moens, 2006; Geist, 2009).

1.1.4. Recuperación de pasajes

La Recuperación de Información se ha enfocado principalmente en la recuperación de documentos completos, aunque también se ha enfocado en otros estudios. Por ejemplo, con la finalidad de reducir la cantidad de tiempo que un usuario debe invertir para satisfacer una necesidad de información, se han hecho trabajos en la denominada Recuperación de Pasajes

[4] Página electrónica de las FAO: http://webguide.fao.org/web_publishing/static/cms0/en/ [Consulta 10/10/2012].

(RP). La premisa básica en que se basa la RP es que algunas partes (o pasajes) de un documento podrían ser más relevantes con respecto a una solicitud que otras. Se ha encontrado que para algunas tareas, el emplear en la representación de los documentos los pasajes que los conforman, en vez de los documentos completos, proporciona mejores resultados (O'Connor, 1975; O'Connor, 1980). Como definición de pasaje existen diversas opciones (Callan, 1994). Los pasajes a nivel del discurso pueden corresponder a unidades textuales tales como oraciones, párrafos o secciones (Salton et al., 1993; Wilkinson, 1994). Los pasajes semánticos se basan en la similitud del contenido del documento. Los pasajes de ventana se basan en un número fijo de palabras. Para la RP se han utilizado tanto los modelos típicos desarrollados en la recuperación de documentos (p. ej., el MEV) como nuevos enfoques basados en otras áreas como la del Procesamiento de Lenguaje Natural (Ledeneva, 2009, Erkan, G. & Radev, D., 2004; Mihalcea 2004).

1.1.4.1. La RP y la documentación legal

El Sistema de Recuperación de Información JIRS (por sus siglas en inglés correspondientes a *Java Information Retrieval System*), es una implementación de un modelo de recuperación de información basado en *n-gramas* enfocado a la Recuperación de Pasajes. Este IRS ha sido utilizado con resultados prometedores para la recuperación de diferentes tipos de documentación legal (Rosso, Correa & Buscaldi, 2011). Si bien la recuperación de documentos y pasajes ha ayudado a facilitar la consulta de información, tales enfoques aún son insuficientes (Daniels & Rissland, 1997) y por ello son aún son temas de amplia investigación. Con el objetivo de mejorar el acceso a la información, específicamente legal, se ha sugerido el considerar la estructura de los textos y las referencias que contienen.

1.2. Lenguajes de Marcado de Textos Electrónicos

Precisamente, tratando de aprovechar la estructura de los documentos legales y las referencias que comúnmente contienen, se ha investigado en los últimos años sobre el uso de los Lenguajes de Marcación de Textos Electrónicos (LMTE). Una idea que ha resultado muy útil para la representación de textos en forma electrónica ha sido la del uso de metadatos llamados más comúnmente etiquetas o marcas. Los metadatos, son información textual que, como su nombre lo indica, se adjunta al contenido de los documentos. Los

metadatos pueden corresponder directamente al contenido de un documento. Por ejemplo, para el caso de un libro, los términos: prólogo, resumen, capítulo I, capítulo II, etc., podrían ser utilizados como metadatos. Incluso, se pueden agregar metadatos a partes que comúnmente no incluyen un nombre u identificador, como el título del libro, los párrafos, etc., y aún más, se pueden incluir datos sobre el documento, que no aparecen en el contenido mismo, como por ejemplo, el área de conocimiento al que pertenece. La finalidad de los metadatos es tanto organizar el contenido de los documentos de una forma clara con base en la propia estructura natural de los documentos, como también, posiblemente, describir su contenido.

A pesar de que los lenguajes de marcación de textos no son propiamente modelos de recuperación de información como los mostrados en la sección precedente, por ser una herramienta desarrollada para la representación y acceso a información textual, se describen de una forma similar a un modelo de recuperación de información, es decir, con base en las etapas de representación y recuperación.

Representación

Los metadatos generalmente se codifican usando un lenguaje estandarizado denominado de marcado o de etiquetado. Entre los más utilizados se encuentra el HTML (*HyperText Mark-up Language*) y el XML (*eXtensible Mark-up Language*) ambos derivados del SGML (*Standard Generalized Mark-up Language*).

(7) Ejemplo de un texto y su representación XML.

Recordatorio. **Alfredo:** **No olvides pasar por mí.** **Gina.**	<?xml version="1.0"?> <note> <to>**Alfredo:**</to> <from>**Gina.**</from> <heading> **Recordatorio** </heading> <body>**No olvides pasar por mí** </body> </note>

Mientras que HTML es un lenguaje que indica la forma en que debe lucir un documento, XML se utiliza más bien para describir su contenido. A pesar de las limitaciones de HTML, su simplicidad, y ciertas características de éste, lo han hecho la herramienta de mayor uso en la WWW, características tales como su capacidad para aplicarse a documentos de cualquier tamaño e integrarse a formatos diversos, la idoneidad para gestionar referencias

internas y externas al propio documento, así como la posibilidad de incorporar modelos de búsqueda en base en las etiquetas del lenguaje de marcas.

Por otra parte, XML ofrece además de las ventajas de HTML la posibilidad de estandarizar la estructura de un tipo de documento en particular y definir para él una gramática propia. Debido a esto, recientemente se está comenzando a usar como complemento precisamente del HTML en la Web. Mediante la gramática se posibilita la descripción precisa de la estructura de los documentos y la forma en que pueden generarse en términos de las posibles configuraciones de los atributos de sus metadatos así como de los posibles valores de éstos últimos. En XML la gramática se denomina DTD (*Document Type Definition*) o diagrama XML, lo que depende de la sintaxis utilizada para la descripción de la propia gramática. Con base en el DTD o el diagrama XML, es posible verificar de forma automática si un documento en particular cumple con la gramática definida. XML también permite la personalización del conjunto de etiquetas para aplicaciones más específicas (el conjunto de etiquetas es extensible).

Recuperación

Generalmente, con la mayoría de los lenguajes de marcado los metadatos se encuentran presentes en los documentos en forma de etiquetas invisibles a los usuarios humanos, pero los programas pueden utilizarlas para además de recuperar, visualizar e incluso clasificar documentos o partes de los mismos. Actualmente, para la recuperación de textos generados en los lenguajes de marcado, además de los motores de búsqueda, existen portales que ofrecen documentos de dominios específicos a través de modelos de recuperación de información tradicionales como el MB o el MEV con la particularidad de que permiten realizar búsquedas en el contenido no sólo del documento completo sino en el de los metadatos.

1.2.1. Los lenguajes de marcado y la documentación legal.

Precisamente los LMTE surgieron con una investigación enfocada a la documentación legal. A finales de 1970, tres investigadores: Charles Goldfarb, Ed Mosher y Ray Lorie, recibieron el encargo por parte de IBM, de diseñar un sistema de edición, almacenamiento, búsqueda y gestión de documentos legales al que respondieron con un sistema de formateo

estructural el que, en un principio, denominaron GML. Debido a su utilidad, para 1986 se convirtió en un estándar, el SGML (*Standard Generalized Markup Language*). A pesar de la enorme potencialidad que SGML ofrece, éste se relegó a la publicación, gestión e intercambio de documentos electrónicos en grandes instituciones. No obstante, HTML, una aplicación del lenguaje SGML, que indica cómo se deben codificar los documentos para su distribución en la Web, se convirtió en la tecnología con mayor presencia en Internet (Alvite-Díez, 2003).

Teniendo en cuenta las peculiaridades que caracterizan a los documentos jurídicos, como su estructura y referencias que contienen, son claras las ventajas que aporta la aplicación de la tecnología Web a la documentación jurídica. Específicamente, la utilización de los LMTE como medio de difusión de este tipo de información en lugar de las tradicionales bases de datos (Francesconi, 2006). Entre algunos de los argumentos para su aplicación se encuentran: su amplio uso en la tecnología Web, la posibilidad de aplicación a documentos de cualquier tamaño y formatos diversos, la idoneidad para gestionar las referencias internas y externas a los propio textos legales y jurisprudenciales, y la posibilidad de incorporar motores de búsqueda que puedan proceder a la indización de los documentos a partir de las etiquetas del correspondiente lenguaje de marcas y su posterior recuperación. Debido a esto, en la última década ha resurgido el interés sobre los LMTE en el dominio legal, como lo muestran diferentes estudios. Algunos de los más relevantes se describen brevemente a continuación.

Entre los trabajos más recientes, se encuentran los realizados por el CETL (*Center for Electronic Text in the Law*) de la Escuela de Derecho de la Universidad de Cincinnati, centro creado con la intención de trabajar con recursos digitales jurídicos, investigar las mejores posibilidades para la representación digital de los textos legales y, por último, publicar en Internet materiales seleccionados relacionados con el Derecho (Fitchett, 1997). En una línea de trabajo similar, comenzó el *Corpus Legis Project*. Proyecto desarrollado por el *Law and Informatics Research Institute*, de la Facultad de Derecho y el Departamento de Lingüística Computacional de la Universidad de Estocolmo, con el fin de elaborar recursos de textos legales electrónicos para la realización de estudios jurídico-lingüísticos (Sjöberg, 1998). El *Corpus Legis Project* ha generado el *Corpus Legis System*,

que comprende, además del corpus textual legal en formato SGML y en otros formatos, otros ficheros asociados (declaraciones SGML, DTD, etc.). El sistema se compone de tres aplicaciones: Panorama (navegador), PRISE (aplicación de RI) y un sistema de gestión y publicación electrónica.

Entre otros trabajos también se encuentran los de Kerrigan y Law (2003) quienes describen una representación de textos normativos y legales basada en el lenguaje XML (*eXtended Markup Language*) a la cual incorporan marcas basadas en lógica de primer orden, y emplean como base de un sistema de verificación de cumplimiento de disposiciones. Mientras que en el trabajo de Mercatali, Romano, Boschi, y Spinicci (2005) se describe lo que los autores consideraron los primeros pasos para la transformación automática de información textual legal en modelos formales. Para ilustrar su propuesta, Mercatali *et al*., utilizan el lenguaje de marcación de textos XML y el lenguaje unificado de modelado UML (*Unified Modelling Language*). Y, finalmente, el de Francesconi (2006) cuyo trabajo describe los avances del proyecto italiano *Norme in Rete* enfocado al desarrollo de herramientas orientadas a la creación y manejo de documentos legales para las cuales emplean principalmente el Lenguaje de Marcación de Textos Electrónicos.

1.3. Análisis de referencias

Un rasgo característico de diversos tipos de documentos (p. ej. Artículos científicos y técnicos, textos legales, jurídicos) es un fuerte uso de referencias internas y externas al propio documento). El aprovechar tales referencias, específicamente en artículos científicos, ha sido estudiado ya desde hace algún tiempo (Salton, 1968; Small, 1973) y más recientemente en el área legal (Turtle, 1995; Moens, 2006; Geist, 2009). Debido a que el enfoque de referencias se adoptó para la publicación de información en Internet, también ya desde hace algunos años hay un amplio interés en el desarrollo de algoritmos que aprovechen tales referencias (Frisse y Cousins, 1989; Croft y Turtle, 1989) lo que ha llevado al actualmente denominado análisis de referencias o citas. Debido a que éste ha sido desarrollado para el acceso principalmente al contenido de documentos, su descripción se realiza con base en los procesos de representación y recuperación de un modelo de recuperación de información.

Representación

Una de las representaciones que ha resultado de mucha utilidad en el análisis de citas han sido los grafos, debido a su uso común como un medio de representación de objetos conectados o relacionados entre sí. En breve, un grafo G consiste en un conjunto de nodos o vértices y un conjunto de arcos o aristas en el que cada elemento de E se asocia a un par vértices. Básicamente, en el análisis de citas en la WWW los nodos de un grafo representan las páginas electrónicas, mientras sus aristas los *hipervínculos* que contienen. De forma similar se puede representar la relación entre un conjunto de personas: los nodos representarían a las personas, mientras que las aristas representarían si una persona conoce a otra (Figura 6).

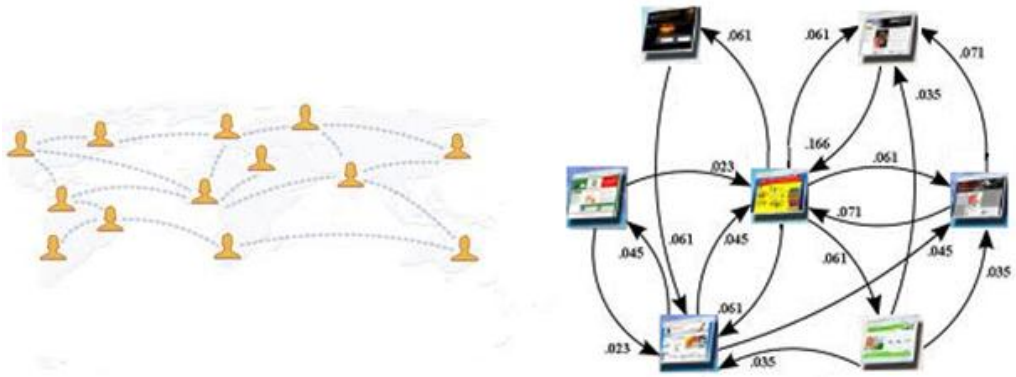


Figura 6. Grafo representando a) una "red social", y b) páginas Web relacionadas entre sí mediante sus *hipervínculos*.

Recuperación

Debido a la evolución de la WWW, ha habido una gran cantidad de trabajos de estudios relacionados con el análisis de referencias. Actualmente, se investiga sobre diferentes técnicas y modelos orientados al análisis de referencias para su aplicación en diversas áreas. Por ejemplo, los algoritmos iterativos de ordenación sobre grafos tales como PageRank, (Brin & Page 1998) o HITS (Kleinberg, 1999) desarrollados principalmente para el análisis de referencias de páginas Web, han sido utilizados también para el análisis de redes sociales y más recientemente han sido aplicados con éxito en el área de procesamiento automático de textos (Mihalcea, Tarau & Figa, 2004; Mihalcea, 2004).

1.3.1. El Análisis de Referencias y la documentación legal.

En la presente investigación, las numerosas referencias que los documentos legales contienen representan información importante, de manera similar a las páginas electrónicas en la Web. No obstante, en su trabajo Lau, Law y Wiederhold (2006), consideran que el dominio jurídico es ligeramente diferente al de la Web, debido a que mientras en el análisis de citas se asume que existe una colección de documentos con referencias de uno a otro, las disposiciones de los textos normativos forman más bien islas de información. Dentro de una isla las disposiciones se encuentran fuertemente relacionadas, mientras las referencias entre islas son raras. Bajo otro enfoque, Geist (2009) llevó a cabo una investigación sobre el uso de técnicas del área de análisis de referencias en sistemas de búsqueda de casos legales, a partir de la cual concluyó que, a pesar de las dificultades que podrían encontrarse al aplicar los algoritmos de análisis de referencias en el área jurídica, éstos podrían ser útiles para mejorar el desempeño de los actuales sistemas de búsqueda de información. Esto debido a la similitud que demuestra existe entre la estructura de las colecciones de casos legales y la de la WWW. No obstante lo anterior, hasta nuestro conocimiento, el análisis de referencias aún no ha sido explorado en el desarrollo de algoritmos para la recuperación de información jurídica[5]. Apartados del marco del análisis de referencias, Lau *et al*. (2006) describen un esquema de análisis comparativo para la recuperación de disposiciones relacionadas pertenecientes a diferentes textos normativos basado en parte en la estructura de los textos legales y las referencias entre sus disposiciones. Por otra parte, Kerrigan *et al*. (2003) consideran las referencias explícitas de los textos normativos para facilitar su consulta en su sistema de ayuda para el cumplimiento de normas legales.

En resumen, el trabajo en la tarea de recuperación de información legal se ha enfocado en el uso de estándares existentes y técnicas de diferentes áreas; sin embargo, hay un amplio terreno poco o aún no explorado en lo que respecta al uso de la estructura de los textos y las referencias que contienen. Debido a esto, en el presente trabajo se investiga un modelo para la recuperación de disposiciones legales basado precisamente en tales características.

[5] Post y Eisen (2000), examinaron las referencias entre documentos de una colección de casos legales con el objetivo de probar la hipótesis de que los argumentos legales y la doctrina legal poseen una clase de estructura fractal; lo cual, de ser así, plantearon, podría ser de utilidad para un mejor entendimiento de la naturaleza y estructura de los sistemas jurídicos.

1.4. Combinación

Como lo muestran las secciones anteriores existen diversos modelos de IR. Muchos otros han sido propuestos y aún siguen siendo tema de investigación (van Rijsbergen, 1986; Deerwester *et al*., 1990; Fuhr, 1992; Turtle y Croft, 1992). Conforme tales modelos han ido desarrollándose, éstos han sido también evaluados intensamente. Desde los primeros experimentos, se observó que diferentes modelos, o alternativamente algoritmos de recuperación, devolvían relativamente pocos documentos en común, aun cuando la efectividad de recuperación de los algoritmos era semejante (McGill *et al*., 1979; Croft y Harper, 1979). Estudios similares mostraron que la práctica de representar los documentos con múltiples representaciones basadas en sus diferentes elementos como sólo el título o el resumen, etc., ofrecía mejores resultados que únicamente todo el contenido del documento (Fisher y Elchesen, 1972; McGill *et al*., 1979; Katzer *et al*., 1982). Éstas, y otras investigaciones, sugirieron que la localización de documentos relevantes para una solicitud de información podría estar más allá de las capacidades de un único modelo de recuperación o una sola forma de representación. La ausencia de coincidencia observada entre los conjuntos de documentos relevantes con el uso de diferentes algoritmos de recuperación (o representaciones) llevó a dos distintos enfoques para el desarrollo de nuevos sistemas y modelos de recuperación. Un acercamiento ha sido a través de la creación de modelos que puedan describir explícitamente y combinar múltiples fuentes de evidencia acerca de la relevancia de los documentos. Estos modelos han sido principalmente probabilísticos, motivados por el principio de ordenación probabilístico (Robertson, 1977), el cual establece que una forma de alcanzar la efectividad de recuperación óptima es mediante la ordenación de los textos en forma descendente con respecto a su probabilidad de relevancia. El otro enfoque ha sido el diseñar sistemas que puedan combinar de forma efectiva los resultados de múltiples búsquedas, basadas en diferentes modelos de recuperación. Esta combinación puede realizarse en una única arquitectura (Croft y Thompson, 1987; Fox y France, 1987) o en un medio ambiente heterogéneo y distribuido (Lee, 1995, 1997; Voorhees, *et al*., 1995; Callan *et al*., 1995). La combinación de múltiples resultados de búsqueda se ha convertido en una técnica importante en las bases de datos multimedia (Fagin, 1996, 1998) y es actualmente la base

de los denominados "metabuscadores", sistemas de recuperación Web (p. ej. MetaCrawler[6]) que combinan los resultados de diferentes motores de búsqueda (Dwork, Kumar, Naor y Sivakumar, 2001; Aslam y Montague, 2001; Gargano y Prasad, 2006; Caputo, Basile y Semeraro, 2009). La motivación tras estos enfoques ha sido el mejorar la efectividad de la recuperación mediante la combinación de múltiples fuentes de evidencia sobre la relevancia de los textos.

Además de los resultados empíricos que muestran la viabilidad de la combinación como un enfoque prometedor para mejorar la efectividad de la recuperación de textos, hay también una fuerte investigación para encontrar su justificación teórica. Una de ellas se ha encontrado en el marco de la probabilidad Bayesiana (Pearl, 1988). En éste, es posible describir la forma en que es afectada una hipótesis H al agregarle una nueva pieza de evidencia e. Específicamente:

$$\log O(H|E,e) = logO(H|E) + logL(e|H) \qquad (8)$$

Donde E es toda la evidencia previa a e.

$O(H|E) = \frac{P(H|E)}{P(\neg H|E)}$, es la probabilidad a posteriori de H dada la evidencia E,

$O(H|E,e)$, es la probabilidad de H dada la nueva evidencia e, y

$L(e|H) = \frac{P(e|H)}{P(e|\neg H)}$, es la razón de probabilidad de la evidencia e.

Esta formulación esclarece el que cada pieza adicional de evidencia positiva incremente la validez de la hipótesis. Una pieza de evidencia con una alta probabilidad de relevancia puede tener un impacto substancial sobre las razones de probabilidad. Adicionalmente, el efecto de un error significativo de la probabilidad de una pieza de evidencia puede reducirse mediante evidencia positiva adicional. En otras palabras, mediante la adición de nueva evidencia es posible lograr una reducción del error promedio. El análisis asume que la evidencia es condicionalmente independiente, sin embargo, si la nueva evidencia puede inferirse directamente de la evidencia previa, el impacto de la nueva evidencia será mucho menor.

[6] www.metacrawler.com

En los modelos de recuperación, la hipótesis de relevancia (R) se basa en la observación (o evidencia acerca) del contenido de un documento (D) y una solicitud específica (Q). La estimación de $P(R|D,Q)$ se puede considerar como la acumulación de piezas de evidencia proporcionadas por diferentes representaciones de la solicitud y/o los documentos. La acumulación de más piezas de evidencia podría resultar en estimados de probabilidad de relevancia más precisos si la evidencia no se correlaciona. A menudo los modelos de recuperación introducen conceptos intermedios que ocultan la relación entre las observaciones y la hipótesis, no obstante, aun en tales casos este modelo simple justifica en parte el uso de la combinación de evidencia.

En lo que respecta a la combinación de los resultados proporcionados por diferentes algoritmos de recuperación sobre una misma representación o la combinación de la salida de diferentes sistemas de recuperación, ambos pueden ser modelados como una combinación de clasificadores; esto ha mostrado que reduce también los errores de clasificación (Tumer y Ghosh, 1999). Un sistema de recuperación puede ser considerado como un clasificador binario (con la clase relevante y no-relevante). Para un documento dado, la salida del sistema corresponde a la probabilidad que dicho documento pertenezca a la clase relevante. En este enfoque, los errores de clasificación reducen la efectividad de la recuperación. La cantidad de reducción del error con la combinación depende de la correlación de las salidas de los clasificadores; entre menor correlación mejores resultados. Se ha mostrado que este modelo proporciona una explicación de muchos de los fenómenos observados en experimentos de combinación (Vogt y Cottrell, 1998), como el del incremento de la probabilidad de relevancia de un documento en el caso de ser evaluado altamente relevante por diferentes sistemas. También proporciona condiciones básicas para lograr una combinación óptima.

Debido a los logros alcanzados con el enfoque de combinación (conocido también como fusión) en la tarea de recuperación de información, la combinación ha sido aplicada a tareas relacionadas como filtrado (Hull *et al*., 1996), categorización (Lewis y Hayes, 1994; Larkey & Croft, 1996), Búsqueda de Respuestas, (Aceves-Pérez, 2008), evaluación de resúmenes automáticos (Lapata, 2006) y ha sido estudiada también en otros campos como el aprendizaje automático (Mitchell, 1997; Fürnkranz y Hüllermeier, 2011), o Biología

(Chuang, Chen, Kao y Hsu, 2004; Yang, Chang, Shen, Kristal y Hsu, 2005; Lin, 2010), Deportes (Truchon y Gordon 2009). Especiales son los trabajos realizados en el área de la Teoría Económica, específicamente la Teoría de Decisión y la Teoría de Elección Social cuyos métodos han sido estudiados y aplicados a tareas de la Ciencia de la Computación (Davenport y Kalagnanam, 2004; Ukkonen, 2004; Conitzer, 2006; Roberts, 2008).

El modelo propuesto permite emplear diferentes representaciones y mecanismos de recuperación, por ello exploramos en su implementación una técnica de combinación para mejorar su efectividad.

En nuestro país el actual mecanismo de acceso a información legal es a través de portales Web gubernamentales en los que la representación y recuperación de la información se basa en el enfoque de bases de datos. Estas se utilizan para almacenar los documentos en formato PDF junto con información sobre estos como su título, tipo (ley, código, etc.), fecha de publicación, institución emisora, etc.

Hasta nuestro conocimiento no hay en nuestro país hasta el momento investigación sobre mecanismos alternos de acceso a la información legal y/o jurídica, por ello parte la motivación del presente trabajo. Adicionalmente, si bien los LMTE surgieron para el manejo de los textos legales y su uso en estos parece viable, su aplicación requiere aún de mecanismos de recuperación que hagan un uso adecuado de las marcas de los textos. En la WWW, los algoritmos de análisis de citas han sido aplicados con éxito; sin embargo, los motores de búsqueda genéricos han resultado insuficientes para la recuperación de información legal (Benamarkian, 2000). Es por ello que recientemente se ha investigado si efectivamente algoritmos desarrollados para el análisis de referencias son útiles también en la recuperación de información legal; sin embargo, hasta el momento no han sido desarrollados modelos de recuperación basados en el enfoque del análisis de citas en el área legal.

2. Marco teórico

El presente trabajo abarca conceptos, métodos, etc. de diferentes áreas, a saber: Procesamiento de Lenguaje Natural (PLN), Teoría de Grafos, Recuperación de Información, Análisis de Referencias, y la denominada Combinación o Fusión de Evidencia. En esta sección se presentan tales conceptos y métodos con el objetivo de facilitar la comprensión del trabajo realizado en la presente tesis.

2.1. Teoría de 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.

Conceptos básicos (tomados de Johnsonbaugh, 2005).

En su forma más simple un grafo G consiste en un conjunto de nodos (o vértices) $V = \{v_1, v_2, \dots, v_n\}$ y un conjunto de aristas (o arcos) $E = \{e_1, e_2, \dots, e_m\}$. Cada arista e_k se asocia a una pareja no-ordenada de nodos $e_k = \{v_i, v_j\}$. Generalmente al grafo definido de esta forma se denomina grafo no dirigido. Se dice que una arista *e* en un grafo que se asocia con el par de vértices *v* y *w* es incidente sobre *v* y *w*, o alternativamente que los vértices *v* y *w* son incidentes sobre *e*. También se suele decir que v y w son vértices adyacentes.

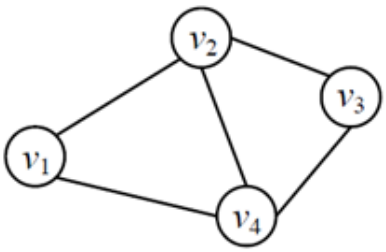


Figura 7. Grafo no dirigido.

Si *G* es un grafo 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. Un grafo es conexo en el caso de que no sea posible dividir el conjunto de nodos en componentes tales que no

existan aristas cuyos nodos incidentes ocurran en componentes diferentes; en caso contrario se denomina grafo inconexo.

Un tipo de grafo muy útil es el grafo bipartito $G = (V_1, V_2, E)$, éste consiste en dos conjuntos disjuntos de nodos V_1, V_2 tal que cada arista $e \in E$ tiene un nodo que pertenece a V_1 y el otro a V_2. Un grafo bipartito es completo si cada nodo en V_1 se conecta con cada nodo de V_2 (**Figura 8**).

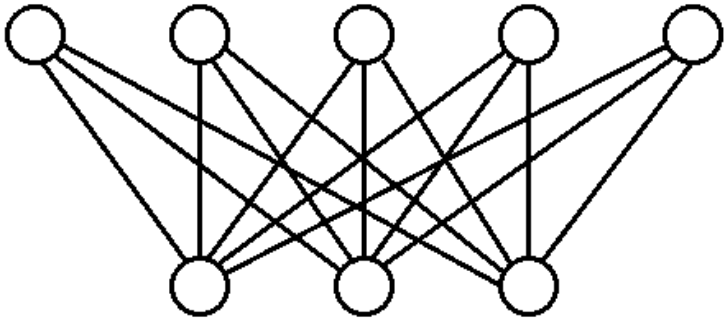


Figura 8. Grafo bipartito.

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

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*.

Un concepto muy importante para el presente trabajo es el de ruta o trayectoria sobre un grafo, la cual se define de manera formal como sigue. 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 un grafo ponderado, la longitud de una ruta es la suma de los pesos de las aristas en la ruta. 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 9 se muestran ejemplos de ambos tipos de grafos.

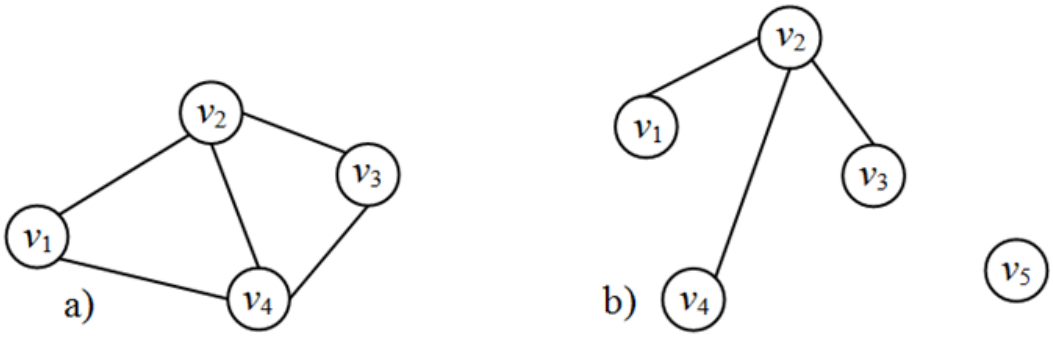


Figura 9. Grafo: a) conexo, b) inconexo.

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. Por ejemplo, sean *v* y *w* vértices en un grafo *G*, entonces:

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

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 Dijkstra

El algoritmo de Dijkstra consiste básicamente 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 son 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 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*.

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.     }
```

Un grafo dirigido consiste en un conjunto de nodos y aristas, pero esta vez una arista es una pareja ordenada de nodos (*u*,*v*), representando una conexión de *u* a *v*.

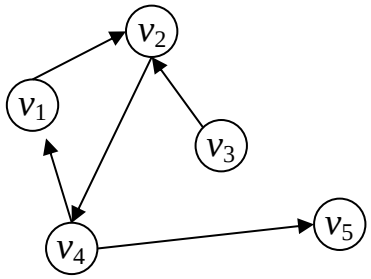


Figura 10. Grafo dirigido

Se dice que existe una ruta dirigida de u a v si existe una secuencia de nodos $u = w_0, \dots, w_k = v$ tal que (w_i; w_{i+1}) es una arista, para toda $i = 0, \dots, k-1$.

Un ciclo o lazo dirigido es una ruta dirigida no trivial de un nodo a sí mismo. Una componente fuertemente conectada de un grafo es un conjunto de nodos tales que para cada par de nodos en la componente, hay una ruta dirigida de uno a otro.

Un *grafo acíclico dirigido*, DAG (por sus siglas en inglés) es un grafo dirigido sin ciclos dirigidos. En un DAG, un *nodo sumidero* es un nodo sin una ruta dirigida a ningún otro nodo. Un uso importante de los DAG es en las denominadas cadenas de Markov, fundamento de los algoritmos de *citas*, uno de los cuales es utilizado en el presente trabajo.

Cadenas de Markov (Golubitsky & Dellnitz, 1999).

Una cadena (homogénea) de Markov para un sistema con un número finito de estados marcados de 1 a n junto con probabilidades p_{ij} de moverse del estado i al j en un paso, se define precisamente por el conjunto de estados $S = \{1,2, \dots, n\}$ y una matriz M de $n \times n$ representando las probabilidades de los movimientos. El sistema comienza en algún estado inicial en S y a cada paso se mueve de un estado a otro. Esta transición está guiada por M: a cada paso, si el sistema se encuentra en un estado i, se mueve a un estado j con una probabilidad M_{ij}. El movimiento de un estado a otro sólo depende del estado en que se encuentre el sistema y no de cómo llego ahí. Si el estado actual del sistema está dado como una distribución de probabilidad, la distribución de probabilidad del siguiente estado está dada por el producto del vector que representa la distribución del estado actual y la matriz M. En general, el estado inicial del sistema se escoge de acuerdo a cierta distribución de probabilidad x (usualmente una distribución uniforme) en S. Después de k pasos, el estado del sistema se distribuye de acuerdo a xM^k. Bajo ciertas condiciones, independientemente de la distribución inicial x, el sistema eventualmente alcanza un punto fijo donde la distribución del estado no cambia más. Esta distribución se denomina distribución estacionaria. Es posible mostrar que la distribución estacionaria del sistema está dada por el *eigenvector* principal y de M, es decir, $yM = \lambda y$. En la práctica, un algoritmo de iteración puede obtener rápidamente una aproximación razonable a y. Una observación importante es que las entradas en y definen un orden natural del conjunto de estados S del sistema. Un aspecto relevante que surge al emplear las cadenas de Markov para ordenar los elementos de S es el siguiente:

Una cadena de Markov define un grafo ponderado dirigido con n nodos tales que el peso de una arista (u, v) está dada por M_{uv}. Las componentes fuertemente conectadas de este grafo definen un DAG. Si este DAG tiene un nodo sumidero, entonces la distribución estacionaria de la cadena estará enteramente concentrada en la componente fuertemente conectada correspondiente al nodo sumidero. En este caso, solo se obtiene una ordenación de las alternativas presentes en esta componente. Si esto sucede, el proceder usual es eliminar estos estados de la cadena y repetir el proceso para ordenar los nodos restantes. Por supuesto, si esta componente tiene suficientes alternativas podría ser posible mejor

detener el proceso y conformarse con una lista parcial con las mejores alternativas. Si el DAG de componentes conectadas esta débilmente conectado y tienen más de un nodo sumidero, entonces se obtendrán dos o más agrupaciones de alternativas las cuales se podrían ordenar de acuerdo a las probabilidades de cada componente. Si el DAG tuviese varias componentes débilmente conectadas podría ser que se obtuvieran agrupaciones de alternativas incomparables.

Uno de los algoritmos más populares basados en cadenas de Markov para el análisis de citas en páginas electrónica, PageRank, resuelve el inconveniente de los nodos sumidero al incluir un mecanismo que le permite salir de un estado, en caso en que el DAG contenga nodos sumideros, permitiendo al algoritmo "saltar" de forma aleatoria de tales nodos. Un aspecto importante del algoritmo PageRank es el hecho de que conserva las propiedades de una cadena de Markov. Específicamente para dicho algoritmo, comenzando con valores arbitrarios asignados a cada vértice en el grafo, los cálculos iteran hasta converger a un umbral predeterminado. Después de ejecutar el algoritmo, cada vértice tendrá asociado un valor el cual representará la "importancia" de cada vértice en el grafo. Se destaca que los valores finales no son afectados por la distribución de probabilidad inicial, sino sólo por el número de iteraciones requerido para alcanzar la convergencia hasta cierto umbral predefinido. Se describe de forma general el algoritmo PageRank y el algoritmo HITS (concebido con un propósito similar a PageRank).

Sea $G = (V, E)$ un grafo dirigido con el conjunto de vértices V y el conjunto de aristas E, donde E es un subconjunto de $V \times V$. Para un vértice dado V_i, sea $In(V_i)$ el conjunto de vértices que llevan al nodo V_i a través de una arista y sea $Out(V_i)$ el conjunto de vértices que puede alcanzarse desde el nodo V_i a través de una arista.

PageRank (PR):

Es probablemente el algoritmo de ordenación más popular originalmente diseñado como un método para el análisis de referencias de páginas electrónicas.

$$PR(V_i) = (1 - d) + d * \sum_{V_j \in In(V_i)} \frac{PR(V_j)}{|Out(V_j)|} \qquad (9)$$

Donde d es un parámetro entre 0 y 1.

HITS (*Hiperlinked Induced Topic Search*):

La característica de este algoritmo es que por cada vértice, HITS produce dos valores los cuales lo distinguen como *authorities* (páginas fuertemente referenciadas por otras páginas), y *hubs* (páginas con numerosas referencias a otras páginas).

$$HITS_A(V_i) = \sum_{V_j \in In(V_i)} HITS_H(V_j)\,, \qquad HITS_H(V_i) = \sum_{V_j \in Out(V_i)} HITS_A(V_j) \qquad (10)$$

2.2. Preprocesamiento

Es común que en la preparación de las colecciones de documentos se eliminen de los textos ciertos símbolos que son de poca utilidad, así como caracteres especiales. Esta remoción dependerá del formato en que se encuentre la colección y de la aplicación que se le dará a ésta. Por ejemplo, si se tratara de páginas HTML, será necesario 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.

Si bien para muchas tareas de IR y PLN se emplean el conjunto de términos tal como se extrajeron de los documentos, es común que algunos además reduzcan dicho conjunto de términos; esto se hace con diferentes propósitos y consideraciones. La reducción puede consistir desde sólo eliminar términos que aparezcan en todos los documentos, o las denominadas *stop-words*, hasta el uso de técnicas como la lematización o el denominado *stemming*.

2.2.1. Stop-words

En cada lenguaje natural existe un conjunto de palabras denominadas "vacías" las cuales se ha encontrado resultan de poca utilidad para diversas tareas tanto de Procesamiento de Lenguaje Natural como de IR. Estas palabras comúnmente corresponden a conjunciones, preposiciones, pronombres, artículos, etc. Aunque también pueden ser verbos, adjetivos y adverbios (Ledeneva, 2009). También se considera que palabras que aparezcan en la mayoría de textos de un corpus definido no contribuyen significativamente en discriminar información en tareas como Recuperación de Información, Búsqueda de Respuestas, Resúmenes Automáticos, etc. De hecho, es ampliamente aceptado que una palabra que aparezca al menos en el 80% de textos en un corpus en particular no es de utilidad para las

tareas antes mencionadas, entre otras. Estas palabras se consideran también como palabras vacías. Una característica de las *stop-words* es que suelen ser en muchos casos palabras muy utilizadas en la redacción de textos. El ejemplo (8) ilustra lo anterior. Presenta una parte del conjunto de palabras de un artículo perteneciente a la Ley Orgánica del IPN. En la columna tf_i se incluye la frecuencia del cada palabra en el artículo y en la columna df_i se proporciona el número de artículos en los que aparece dentro de una colección de 1,632 artículos en total. Como se observa las preposiciones *con* y *por* aparecen en una mayor cantidad de documentos, en comparación de las palabras *educativas* o *impartan*.

(8) Frecuencia de las palabras del artículo 1 de la Ley Orgánica del IPN.

Términos	tf_i	df_i
impartan	**2**	**16**
conjunta	1	7
extranjero	1	17
país	1	31
aquellos	1	39
educativas	**1**	**42**
…		
con	**1**	**726**
por	**1**	**815**
del	2	1214
las	1	1043
el	2	1448
en	1	1259
los	1	1292
de	4	1624
la	1	1367

2.2.2. Lematización

En el área de Procesamiento de Lenguaje Natural, *lematización* se refiere al proceso de encontrar los principales fenómenos morfológicos en una palabra dada, 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 aquella proporcionada al lematizador y la de la derecha es la palabra que proporciona este mismo.

Estudiantes → estudiante, Trabaje → trabajar

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

En el presente trabajo se consideró que lematizar podría ser de mucha utilidad debido a que comúnmente los usuarios de información legal plantean su situación en presente, por ejemplo, ¿Cuáles son los requisitos que deben cumplir…?. Por otra parte, en los textos legales, el tiempo verbal más usado en español es el futuro de mandato o legislativo para expresar el carácter preceptivo: "los requisitos que deberá cumplir el personal…". Al lematizar posiblemente los términos de búsqueda concuerden con los de un documento y con ello se ayuda a su recuperación, siempre y cuando el mecanismo de recuperación considere o esté basado precisamente en tal situación.

2.2.3. Stemming

En el área de recuperación de información, a *lematizar* se le conoce como el proceso de encontrar la "raíz" de una palabra dada; esta raíz es 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 (Porter, 1997).

2.3. IRS

Para investigar la capacidad del modelo y métodos de implementación del mismo se compararon sus resultados con los de enfoques ya utilizados en el dominio legal, los cuales se eligieron también tomando en cuenta que como característica principal no consideraran las referencias o estructura de los textos legislativos. Esto se hizo con la finalidad de determinar si el desempeño del modelo propuesto aprovecha o no la inclusión de tales características de los documentos. Se eligió el modelo del espacio vectorial, el software libre de recuperación de información Lucene, y el sistema de recuperación de pasajes, también libre, JIRS. A continuación se describen brevemente tanto Lucene como JIRS debido a que posteriormente el análisis de resultados se realiza precisamente con base en su mecanismo de recuperación.

2.3.1. Lucene

El software libre Lucene combina el Modelo Booleano (MB) y el Modelo del Espacio Vectorial (MEV). Los documentos que son "aprobados" por el MB son después calificados y ordenados mediante el MEV.

Representación.

En Lucene, que usa el MEV, los documentos y las solicitudes se representan como vectores a partir de los valores asociados a sus términos.

En el MEV los valores asociados a los términos de la solicitud y los documentos se obtienen a partir de una versión modificada de la típica *tfidf*. En Lucene *tfidf* se define por:

$$tfidf_{i,M}(t,d,N) = tf(t,d) * idf(t,N) \quad (11)$$

$$tf(t,d) = \sqrt{f_{t,d}} \quad (12)$$

$$idf(t,N) = 1 + \log\left(\frac{N}{m(i,N)+1}\right) \quad (13)$$

Donde $m(i,N)$ es el número de documentos en la colección de documentos N que contienen al término i

Recuperación.

El mecanismo de recuperación se basa en la medida de similitud coseno con modificaciones realizadas con el objetivo de tanto mejorar la calidad de sus resultados como facilitar su implementación.

- Se encontró que la normalización de los vectores-documento presenta diversos inconvenientes, el principal es que elimina información relativa a la longitud del documento, en función del número de términos que contiene. Para un documento con un párrafo que se repita, por ejemplo, 10 veces con términos diferentes, esto podría ser adecuado, pero para documentos que no contengan párrafos duplicados esto podría no ser correcto. Para evitar este problema, Lucene emplea un factor de normalización que produce un vector unitario o de mayor longitud (dependiendo de ciertas características de cada documento): *doc-len-norm*(*d*).

- En el proceso de representación, los usuarios pueden especificar que ciertos documentos son más importantes que otros. Por esto, el valor de similitud se multiplica por el valor de "importancia" de cada documento *doc-boost*(*d*).
- Durante la recuperación los usuarios pueden especificar una mayor importancia a ciertos términos de la solicitud lo que se toma en cuenta mediante el factor: *q_boost*(*q*).
- Con el MEV es posible recuperar documentos que no necesariamente contengan todos los términos de la solicitud. En Lucene los usuarios pueden además recompensar documentos que contengan un mayor número de términos de la solicitud mediante un *factor de coordinación*: el que es mayor mientras más términos compartan en común documento y solicitud: *coord-factor*(*q*, *d*).

$$score(q,d) = coord(q,d) * q\_boost(q) \frac{V(q) \cdot V(d)}{\|V(q)\|} * doc\_len\_norm(d) * d\_boost(d)$$

La ecuación anterior es sólo conceptual, en el sentido que es únicamente una simplificación de la fórmula utilizada realmente en el proceso de recuperación. En la documentación de Lucene se describe su implementación[7].

2.3.2. JIRS

El Sistema de recuperación de información JIRS[8] (por sus siglas en inglés correspondientes a *Java Information Retrieval System*), es una implementación de un modelo de recuperación de información basado en *n-gramas*. Debido a la complejidad computacional que implica su mecanismo de recuperación, a diferencia de un sistema de información tradicional (el cual devuelve un conjunto de documentos) JIRS se diseñó para la recuperación de una menor cantidad de texto como párrafos (o pasajes) de un corpus relativamente pequeño. A continuación se describe de forma general su funcionamiento.

Representación

Al contrario del modelo booleano y el MEV que utilizan las palabras de los textos como unidades lingüísticas básicas, JIRS utiliza n-gramas en su modelo principal de

[7] http://lucene.apache.org/core/3_6_0/scoring.html
[8] http://sourceforge.net/projects/jirs/

recuperación. Un *n-grama* se define como una secuencia de n palabras consecutivas. Decimos que un n-grama ocurre en un texto si esas palabras aparecen en el mismo orden, inmediatamente una después de la otra (Schneider, 2002).

Recuperación

Mediante el empleo de n-gramas el mecanismo de recuperación de JIRS tiene la capacidad de localizar estructuras formadas por términos de la solicitud de información dentro de una colección de documentos (Buscaldi, Rosso, Gómez-Soriano, & Sanchis, 2010).

En JIRS, la recuperación de pasajes se realiza en dos etapas: en la primera, efectúa una reducción de la colección de pasajes mediante un método tradicional de recuperación de información. Posteriormente asigna a cada pasaje del subconjunto resultante un valor de peso en función de los n-gramas que comparte con la solicitud. JIRS básicamente considera más relevantes aquellos pasajes con un mayor peso. Estos pasajes son devueltos en forma de lista.

2.4. Modelo de combinación lineal

El modelo de combinación lineal ha sido utilizado para la recuperación de información por muchos investigadores con diferente grado de éxito (Bartell, Cottrell y Belew, 1994; Fuller, Kaszkiel, Ng, Vines, Wilkinson y Zobel, 1998; Kantor, 1995; Knaus, Mittendorf,. y Schauble, 1995; Selberg, y Etzioni, 1996; Shaw y Fox, 1995; Strzalkowski, Lin y Pérez-Carballo, 1998; Vogt, Cottrell, Belew y Bartell, 1997).

El modelo de combinación lineal calcula el valor de relevancia *p* de un documento *x* a una solicitud *q* con base en los pesos $\vec{w} = (w_1, w_2, \dots, w_s)$ dados a cada uno de los sistemas de recuperación de información *s*, y sus estimados de relevancia ρ_i.

$$\rho_{\vec{w}}(x, q) = \sum_{i=1}^{s} w_i \rho_i(x, q)$$

Este valor es posteriormente empleado para ordenar los documentos. Para sólo dos sistemas de recuperación de información, la ecuación anterior se simplifica a:

$$\rho_{w_1,w_2}(x, q) = w_1 \rho_1(x, q) + w_2 \rho_2(x, q)$$

Un aspecto importante del modelo de combinación lineal es que es equivalente al tipo de red neuronal más simple, la adición de la función de activación en la salida no agrega poder computacional, y como tal no cambia el orden inducido por la combinación de los sistemas

de recuperación. En el presente trabajo se utiliza la combinación con valores de ponderación prestablecidos, sin embargo podría ser posible utilizar técnicas de aprendizaje automático o métodos adaptativos para el cálculo de tales valores.

3. Modelo propuesto y estudio experimental

En este capítulo se presenta el modelo propuesto, su implementación y la descripción de los experimentos llevados a cabo.

3.1. Descripción del modelo propuesto

El modelo propuesto consta de dos etapas: una de representación de los textos legales y otra de recuperación de sus disposiciones a partir de una solicitud de información formulada como una pregunta en lenguaje natural. La solicitud de información plantea una situación que es posible resolver mediante una o más disposiciones normativas, las cuales en ocasiones pueden estar acompañadas por su fundamento legal, es decir, una más disposiciones pertenecientes primordialmente a leyes o a la Constitución. Debido a la naturaleza de las normas puede ser posible que las disposiciones respuesta se encuentren dispersas en uno o más textos, relacionadas de manera explícita, como se ilustra en la **Figura 11**.

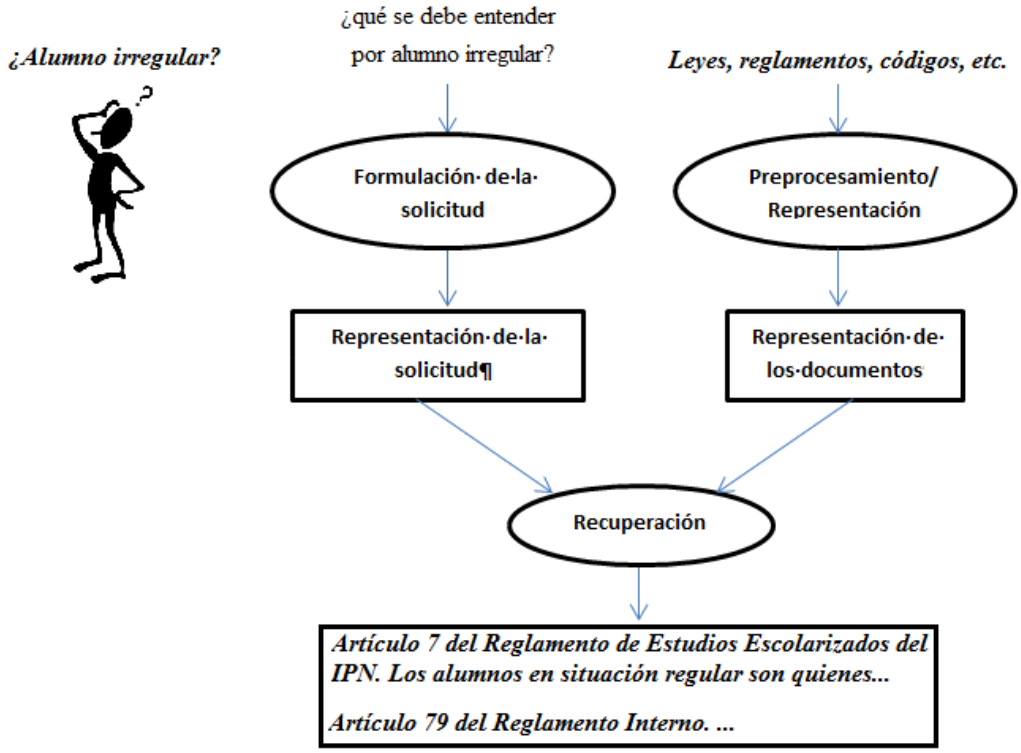


Figura 11. Los óvalos representan las etapas que la modelo específica y los cuadros las salidas de cada proceso.

El modelo propuesto se basa en un grafo ponderado no dirigido $G = (A, S)$, como se describe a continuación.

3.1.1. Representación

Los documentos y/o artículos de las disposiciones generales que contienen se representan mediante el conjunto de vértices del grafo ponderado no dirigido y a través del conjunto de aristas S, su similitud y/o las referencias entre artículos. En la **Figura 12** se ilustra la representación propuesta.

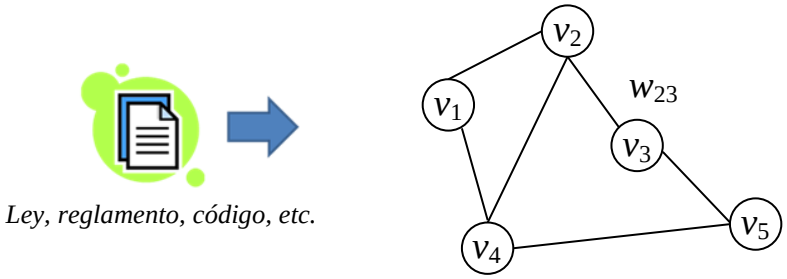


Figura 12. Representación propuesta: los vértices pueden representar los documentos completos de toda una colección o el conjunto de sus artículos.

El grafo se eligió debido a que refleja la estructura de los textos en lo que respecta a la división de su contenido principal: las disposiciones generales y las disposiciones transitorias. Si bien el preámbulo contiene información relevante para el entendimiento de la razón, objetivo, etc. del documento mismo, en el presente trabajo nos interesa y nos enfocamos a las normas orientadas a regular la conducta de los individuos. Por esta misma razón también consideramos únicamente los artículos de las disposiciones generales, ya que de acuerdo a Huerta-Ochoa (2009) el contenido de los artículos transitorios se refiere más bien a la forma de aplicación de las normas de las disposiciones generales. No obstante lo anterior, en la implementación del modelo se podría incluir una forma de permitirle al usuario la consulta de los artículos transitorios relevantes y el preámbulo de los artículos devueltos, o bien, pueden ser incluidos en el modelo. En el presente trabajo se eligió la primera opción.

3.1.2. Recuperación

Dada una solicitud de información, ésta básicamente se agrega a los grafos de artículos o documentos como si se tratara de nuevos nodos y con base en algoritmos orientados a grafos, se recuperaran ya sea documentos o artículos. En el presente trabajo se exploraron y propusieron diferentes algoritmos e incluso su combinación, como más adelante se muestra.

3.2. Estudio experimental

Con respecto a la comparación entre documentos y artículos se utilizó la función de similitud coseno. Para el uso de la similitud coseno, primero se preprocesaron los documentos de la manera siguiente: se lematizó el contenido de los textos (verbos, adjetivos, sustantivos, etc.) y luego se convirtió a minúsculas, posteriormente se eliminaron todos los signos de ortografía, números y una lista predefinida de palabras (denominadas en inglés *stop-words*). Posteriormente se extrajeron los artículos de las disposiciones principales de cada documento. Tanto los términos de cada documento como de cada artículo se ponderaron partir de la medida *tfidf* (*term frequency inverse document frequency*). De esta manera, el valor de un término x en el k-ésimo documento (o artículo), de la colección A, quedó definido por:

$$\mathrm{tfidf}(x, a_k, A) = \mathrm{tf}(x, a_k) * \mathrm{idf}(x, A) \tag{14}$$

$$\mathrm{tf}(x, a_k) = \text{Frecuencia de } x \text{ en } a_k \tag{15}$$

$$\mathrm{idf}(x, A) = \log\left(1 + \frac{|A|}{\text{Artículos (o documentos) que contienen x}}\right) \tag{16}$$

3.2.1. Corpus

Un inconveniente que se encontró en el desarrollo del presente trabajo fue que en nuestro país no hay disponibles colecciones de documentos legislativos para su uso con métodos automáticos. Por ello fue preciso primeramente conformar un conjunto sobre el cual se pudiera investigar y evaluar el desempeño del modelo propuesto y los enfoques de referencia. A continuación se describe el corpus que se integró así como las solicitudes de información que se emplearon en el trabajo.

3.2.1.1. Documentos

Se utilizaron para los experimentos un total de 37,153 disposiciones: 2,162 de ellas relativas al Instituto Politécnico Nacional (IPN) y el resto de aplicación general. El conjunto total de disposiciones correspondió a 29 textos pertenecientes al IPN[9] y el resto a leyes de aplicación federal[10].

Además de la Constitución Política de los Estados Unidos Mexicanos, los tipos de documentos utilizados y la cantidad se especifican a continuación:

- Leyes (245), Ordenanza (1), Estatuto (1), Códigos (9) y Reglamentos (31).

Los títulos de los documentos utilizados se encuentran en el Anexo A.

Se conformó un conjunto de documentos en formatos comerciales (PDF y HTML) el cual se convirtió a un formato común. Se escribieron *scripts* para la conversión de los formatos originales de los documentos a texto plano[11], los cuales posteriormente, se procesaron de forma semiautomática para convertir de texto plano a XML. Esto último con la finalidad de facilitar en la implementación del modelo la visualización de los resultados.

Para obtener las referencias entre artículos se analizaron referencias extraídas de forma manual de artículos pertenecientes a 5 documentos. A partir de éstas, se definieron patrones para su obtención automática en los 283 documentos restantes. De esta forma se encontró que del total de artículos, hasta un 28% de ellos hacían referencia a otro artículo (**Tabla 1**).

Tabla 1. Número de referencias entre artículos encontradas en la colección de textos.

Número de artículos		**1117**	**2162**	**8987**	**37153**
Referencias entre artículos:	Interdocumento	36	50	157	1398
	Intradocumento	135	205	1474	8901
Referencias en total:		171 (15%)	255 (12%)	1631 (18%)	10299 (28%)

[9] Los textos pertenecientes al IPN están disponibles en línea: http://www.abogadogeneral.ipn.mx
[10] El conjunto de leyes federales está disponible en: http://www.diputados.gob.mx/LeyesBiblio/index.htm
[11] El conjunto de documentos en texto plano está disponible en línea: http://iarp.cic.ipn.mx/~monroy/

3.2.1.2. Preguntas/Respuestas

Para evaluar el método, se utilizó un conjunto de 40 preguntas expresadas en lenguaje natural, cada una con su respectiva respuesta, constituida por uno o más artículos (máximo 5) pertenecientes a una ley y sus reglamentos correspondientes. En total se utilizaron 70 artículos distintos para responder al conjunto de 40 preguntas. El conjunto de artículos-respuesta se obtuvo de la Ley Federal de los Trabajadores al Servicio del Estado Reglamentaria del Apartado "B" del Artículo 123 Constitucional (LFTSE) y de 10 documentos más, todos relativos al IPN, los cuales fueron:

- Ley Orgánica (LO),
- Reglamento de Academias (RA),
- Reglamento de Titulación Profesional (RTP),
- Reglamento de las Condiciones Interiores de Trabajo del Personal Académico (RCITPA),
- Reglamento de las Condiciones Generales de Trabajo del Personal No Docente (RCITPND),
- Reglamento del Decanato (RD),
- Reglamento de Estudios Escolarizados para los Niveles Medio Superior y Superior (REE),
- Reglamento Interno (RI),
- Reglamento de Estudios de Posgrado, y el Reglamento de Servicio Social (RP)

Las preguntas se clasificaron de dos formas distintas: la primera clasificación se hizo en función de la necesidad de información planteada, y en la segunda clasificación las preguntas se dividieron en 4 tipos en función de características de sus artículos-respuesta.
Para la primera clasificación las clases fueron las siguientes:

Si/no. Son preguntas que se referían a la posibilidad de realizar o no cierta acción, generalmente se expresan como: ¿Es posible…? ¿Procede…? ¿Se puede…? 26 preguntas, Ejemplo (9).

Procedimiento. Preguntas que demandaban como respuesta los requisitos, o la manera oficial o aceptada, para realizar cierta tarea, cubrir algún puesto, cargo, etc. ¿Cómo se lleva

a cabo…? ¿Cuál es el procedimiento…? ¿Cuáles son los requisitos…? 7 preguntas, Ejemplo (10).

Definición. Solicitaban información que permitiera comprender algún concepto. ¿Qué es…? ¿Qué significa…? 7 preguntas, Ejemplo (11).

(9) ¿Es procedente que un alumno solicite mención honorífica a nivel licenciatura si escogió su titulación por la opción de escolaridad?

Respuesta: Artículo 13 y 43 del Reglamento de Titulación Profesional.

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

Respuesta: Artículo 28 de la Ley Orgánica y, 206, 207, 209 y 213 del Reglamento Interno.

(11) ¿Qué debe entenderse por tener estudios de posgrado?

Respuesta: Artículo 3 y 4 del Reglamento de Estudios de Posgrado.

En la segunda clasificación, las preguntas se dividieron en 4 tipos en función de características de sus artículos-respuesta.

- Tipo 1. Preguntas para las cuales sus artículos-respuesta contenían referencias entre sí, por ejemplo, el artículo 175 del Reglamento Interno hacía referencia al artículo 174 del mismo documento y ambos respondían las preguntas 9 y 15. La misma característica se encontró en las preguntas 1, 36, 5, 25, 10, 23 y 19.
- Tipo 2. Los artículos-respuesta contenían referencias a artículos que no respondían las preguntas. Por ejemplo, para la pregunta 16 el artículo-respuesta 208 del Reglamento Interno hacía referencia al artículo 27 de la Ley Orgánica; sin embargo, este último no pertenecía a la respuesta. Las preguntas 17, 21, 39, 8, y 7 presentaron el mismo rasgo.
- Tipo 3. Los artículos-respuesta no contenían una referencia y compartían pocos términos en común con la solicitud. Preguntas: 35, 11, 26, 37, 18, 3, 22, 33, 6, 20, 12, 34, 4, 13, 28 y 14.

- Tipo 4. Los artículos-respuesta no contenían referencias y compartían con las preguntas, a diferencia de los artículos-repuesta del bloque anterior, un mayor número de términos en común. Preguntas: 40, 2, 24, 29, 30, 31, 32, 27 y 38.

El análisis de resultados se realizó considerando ambas clasificaciones. El conjunto de preguntas/respuestas se encuentra en el Anexo B.

3.2.2. Mecanismo de evaluación

Cada uno de los enfoques se evaluó con base en su capacidad para recuperar los artículos que respondieran a las preguntas de prueba. En todos los experimentos, la lista de artículos generada por cada enfoque se calificó como: **Correcta**, si contenía todos los artículos que respondían la pregunta; **Parcial**, en el caso de no contener todos los artículos-respuesta y finalmente **Incorrecta** en caso de no recuperar los artículos requeridos. También se consideraron las posiciones de los artículos-respuesta recuperados. Si bien existen otras formas de evaluación (como las ampliamente utilizadas *Precision* y *Recall*) en el presente trabajo se eligió la evaluación descrita a fin de realizar un análisis detallado de los resultados experimentales.

3.2.3. Experimento I.

El primer experimento se realizó a fin de explorar si incorporar las referencias en la representación y recuperación de los textos legales podría mejorar la búsqueda de disposiciones relacionadas; esto se comparó con respecto al caso de no incluirlas y con uno de los modelos más ampliamente utilizados: el Modelo del Espacio Vectorial.

3.2.3.1. Representación: Grafo de artículos con referencias

En este experimento, para la etapa de representación se generó un grafo con los artículos de un conjunto preestablecido de documentos. Específicamente, los artículos de las disposiciones generales del corpus se representaron mediante el conjunto de vértices $A = \{a_1, a_2, \dots, a_n\}$ de un grafo ponderado no dirigido G, y a través del conjunto de aristas S su similitud y/o las referencias intradocumento e interdocumentos encontradas. Entre los vértices a_i y a_j se estableció una arista con valor asociado p_{ij}, de acuerdo a lo siguiente:

- En el caso en que cualquiera de los artículos representados por los vértices a_i, a_j

incluyera una referencia al otro y además $s(a'_i, a'_j) = 0$: $p_{ij} = 1/100$. Mientras que, para $s(a'_i, a'_j) > 0$: $p_{ij} = 1/50 * s(a'_i, a'_j)$

- En el caso en que los artículos no contuvieran una referencia, y $s(a'_i, a'_j) > 0$: $p_{ij} = 1/(100 * s(a'_i, a'_j))$

3.2.3.2. Recuperación: Algoritmo de la ruta mínima

Primero, cada pregunta, se lematizó y convirtió a minúsculas. A las palabras resultantes se les asoció un número entero en orden progresivo y creciente de acuerdo a su aparición. Se separaron los términos pares de los impares para formar dos conjuntos de términos $q_a = \{2,4, \dots\}$ y $q_b = \{1,3, \dots\}$. Luego, los términos de cada conjunto se ponderaron para su comparación con los artículos a partir también de la semejanza coseno, Ejemplo **(12)**.

(12) Ejemplo división de solicitud y ponderación de términos.

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

llevar cabo procedimiento elección representante alumno ante consejo técnico consultivo escolar			
q_a		q_b	
i	$tfidf_i$	i	$tfidf_i$
cabo	0.310	llevar	0.233
elección	0.352	procedimiento	0.200
alumno	0.174	representante	0.226
consejo	0.207	ante	0.213
consultivo	0.231	técnico	0.171
		escolar	0.165

En el grafo *G* se representaron las dos partes de la pregunta q_a y q_b mediante dos nuevos vértices a_{n+1} y a_{n+2}. Se generó una arista entre dos vértices a_k y a_{n+1} (así como entre a_k y a_{n+2}) sí $s(a'_k, a'_{n+1}) > 0$: $p_{k(n+1)} = \frac{1}{100 * s(a'_k, a'_{n+1})}$. Donde a'_k y a'_{n+1} son los vectores con los valores de ponderación de los términos en a_k y q_a respectivamente.

Finalmente, las rutas entre los vértices correspondientes a la solicitud se obtuvieron mediante el algoritmo de la ruta mínima. El proceso consistió en ejecutar el algoritmo entre los vértices a_{n+1} y a_{n+2} y eliminar del grafo los vértices de la ruta obtenida, continuando de la misma manera hasta que se obtuvieron todas las rutas posibles. Se eligió este algoritmo debido a que las aristas reflejan el grado de similitud entre artículos de forma inversa a la tradicional, es decir, para un par artículo altamente similares o que compartan una referencia, el valor de la arista que los une es menor que en el caso contrario; por ello, al emplear el algoritmo de la ruta mínima, consideramos que se recuperarían precisamente artículos altamente semejantes entre sí o que contuvieran una referencia. Esto último es lo que a su vez consideramos podría ser de utilidad para localizar los artículos relacionados entre sí (de manera implícita o explícita) que respondían al conjunto de preguntas empleado.

3.2.3.3. Corpus: 2,162 artículos. Preguntas/Respuestas: 21.

Como primer experimento exploratorio para determinar la viabilidad del modelo, se utilizaron solamente 2,162 artículos pertenecientes a 29 documentos relativos a la legislación y normatividad del IPN. Debido a la misma razón sólo se utilizaron 21 preguntas/respuestas con diferentes características. Preguntas 1 a la 21.

3.2.3.4. Comparación MEV, grafo sin referencias.

Para verificar el impacto de las referencias en la recuperación de artículos, se realizó una implementación del método propuesto y su evaluación, así como de un grafo sin incluir las referencias entre artículos.

Los resultados de la evaluación de los grafos con y sin referencias, GCR y GSR, respectivamente, fueron comparados con el MEV, implementado para la misma tarea. Este se eligió debido a su amplia difusión y uso en el área de recuperación de información, además de que sólo considera la semejanza entre la solicitud y los documentos, y no entre éstos.

Es importante destacar que para la implementación del MEV, en lugar de la colección de documentos que generalmente emplean, se usó el mismo conjunto de artículos que para la construcción del grafo, así como la misma medida de semejanza y método de ponderación.

3.2.4. Experimento II

El análisis de los resultados del primer experimento mostró prometedora la idea de emplear las referencias entre artículos; por ello, a fin de aprovecharlas de formas distintas se diseñó el presente experimento, básicamente para responder la siguiente pregunta:

¿Se pueden utilizar algoritmos típicos del análisis de referencias para mejorar la localización de información legal con respecto a modelos de recuperación de información del estado del arte más robustos que el MEV?

3.2.4.1. Representación: Grafo de artículos con referencias

De forma similar al experimento I, para este experimento se generó un grafo para representar las disposiciones de un conjunto predefinido de documentos. También, a partir del valor de similitud y las referencias entre artículos, se estableció una arista entre los vértices a_i y a_j con valor asociado p_{ij}; sin embargo, se introdujeron algunas diferencias con respecto al experimento I, de acuerdo a lo siguiente:

- Si los artículos correspondientes a los vértices a_i, a_j no contenían una referencia entre sí y $s(a_i', a_j') > 0$, entonces $p_{ij} = s(a_i', a_j')$.
- En el caso de que los artículos representados por los vértices a_i, a_j contuvieran una referencia entre sí y $s(a_i', a_j') = 0$, entonces $p_{ij} = 1$. Para el caso $s(a_i', a_j') > 0$, entonces $p_{ij} = 2 * s(a_i', a_j')$.

3.2.4.2. Recuperación: Combinación algoritmos de recorrido de grafos.

Para la recuperación de artículos a partir del grafo se siguió el proceso que se describe a continuación. Dada una pregunta q: primero sus términos se dividieron en dos conjuntos, los cuales después se representaron en el grafo como si se tratara de dos nuevos artículos; después, se recorrió el grafo entre los dos nuevos vértices para obtener una lista ordenada de artículos. La **Figura 13** se ilustra el proceso seguido para la recuperación de los artículos a partir de una pregunta dada.

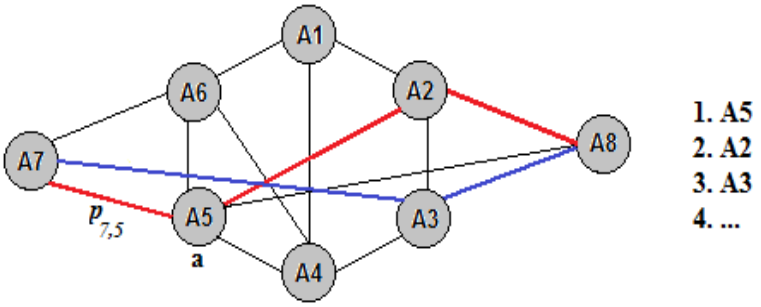


Figura 13. Proceso de representación de la solicitud y recuperación de artículos

A continuación se describe el proceso de división de la pregunta y de recuperación de artículos: primero, cada pregunta se procesó de manera similar a los artículos, es decir, se lematizó y convirtió a minúsculas, se eliminaron signos de ortografía, números, así como los términos que no aparecían en el conjunto de artículos. Posteriormente, a los términos restantes se les asoció un número comenzado con 1 para el primer término 2, para el segundo, etc. Después, se separaron los términos pares de los impares para formar dos conjuntos de términos q_a = {2, 4, …} y q_b = {1, 3, …}. Para la comparación de q_a y q_b con los artículos se utilizó nuevamente la medida de similitud coseno y la ponderación *tfidf*.

Como anteriormente se mencionó, en el grafo se representaron las dos partes de la pregunta q_a y q_b mediante dos nuevos vértices a_{n+1} y a_{n+2}. Se generó una arista a partir del valor de similitud entre dos vértices a_k y a_{n+1}, uno correspondiente a un artículo y el otro a una parte de la solicitud de información. Sí $s(a'_k, a'_{n+1}) > 0$, entonces $p_{k(n+1)} = s(a'_k, a'_{n+1})$, donde a'_k y a'_{n+1} son los vectores con los valores de ponderación de los términos en a_k y q_a respectivamente.

Las rutas entre los vértices a_{n+1} y a_{n+2} se obtuvieron mediante una combinación lineal de los valores asociados a las aristas y valores asociados a los vértices. Cada ruta se comenzó en el vértice a_{n+1}. El avance se realizó eligiendo el nodo que maximizará el valor de la expresión (17), hasta alcanzar el nodo a_{n+2}. El proceso se repitió, sin considerar los vértices previamente recuperados, hasta obtener todas las rutas posibles o un máximo de 75 vértices.

$$w_1 * p_{xy} + w_2 * v(a_y) \qquad (17)$$

El valor asociado a cada vértice se obtuvo a partir del algoritmo de análisis de referencias PageRank en su versión para grafos ponderados no dirigidos (Mihalcea, 2004):

$$v(a_i) = (1-d) + d * \sum_{a_j \in E(a_i)} p_{ji} \frac{v(a_j)}{\sum_{a_k \in S(a_j)} p_{kj}} \quad (18)$$

donde a_i corresponde a un vértice, d un valor fijo predefinido, a_j un vértice adyacente a a_i y p_{ji} el valor asociado a la arista entre los vértices a_i y a_j.

Los valores w_1 y w_2 se obtuvieron de manera experimental a partir de evaluar la implementación del modelo solamente con las 21 preguntas del primer experimento. Se eligieron para el resto de experimentos (II y III) los valores que devolvieron los mejores resultados $w_1 = 0.95$ y $w_2 = 0.05$. Para d se utilizó uno de los valores recomendados en la literatura $d = 0.85$.

Con el objetivo de estudiar el impacto del uso del algoritmo de referencias, se realizaron tres experimentos principales (**Tabla 2**). A continuación se describe cada uno de ellos:

Tabla 2. Descripción experimentos.

Nombre del experimento	Descripción	Condiciones Expresión (17)
SSim.	El recorrido del grafo se realizó solo considerando el valor de similitud entre artículos.	$w_1 = 1, w_2 = 0$
SPR.	El recorrido se realizó solamente con base en el valor de PageRank.	$w_1 = 0, w_2 = 1$
Sem. & PR.	En este caso se utilizó la combinación lineal de los valores de similitud y PR con los valores reportados en la sección precedente.	$w_1 = 0.95, w_2 = 0.05$

3.2.4.3. Corpus: 1,117, 2,162 y 8,987 artículos. Preguntas/Respuestas: 40.

Para los tres experimentos descritos en la **Tabla 2** se utilizó una colección de disposiciones (Colección A) compuesta por 1,117 artículos pertenecientes a los 11 documentos que se utilizaron para responder el conjunto de preguntas de prueba.

Con la finalidad de estudiar el desempeño de la combinación de la similitud y PR al incrementar el espacio búsqueda, se realizaron dos experimentos secundarios con dos colecciones de artículos más: una colección (B) conformada por 2,162 artículos (los artículos de la colección A más los de 18 documentos pertenecientes al IPN) y una colección que denominamos colección C compuesta por un total de 8,987 artículos, los artículos de las dos colecciones anteriores (A y B) y los de 13 textos con normas de aplicación general.

3.2.4.4. Comparación Lucene, JIRS.

Las mismas colecciones A, B y C se indexaron con Lucene y JIRS. Para la fase de indexado Lucene se utilizó sin cambio alguno, mientras que JIRS se utilizó con las siguientes opciones: `Indexing -language Spanish -stemmer no`

Ambos se utilizaron para recuperar una lista de artículos. Los resultados de cada experimento se compararon con los correspondientes al modelo propuesto. Para la fase de recuperación de Lucene, su salida se limitó a un máximo de 75 artículos; en el caso de JIRS además de limitar su salida a un máximo de 75 pasajes (en nuestro caso, los artículos lematizados), se utilizaron las siguientes opciones:

```
SearchPassages -language Spanish -stemmer no -repeat no -model distance -naddlines 3 -disfactor 0.1 -reformulation yes -searchengine rw -term_weighter tf-idf -combination union -npassages 75.
```

3.2.5. Experimentos III

En este experimento se exploró la capacidad del modelo propuesto al incrementar significativamente el espacio de búsqueda de disposiciones. Debido a que en el segundo experimento una observación fue que los mejores resultados se obtenían con una colección de artículos reducida, se pensó en aplicar el modelo propuesto en dos pasos, primero en la

recuperación de documentos y luego generar un grafo de artículos a partir de los artículos sólo de los textos obtenidos en el primer paso.

3.2.5.1. Representación: Grafo de documentos. Grafo de artículos.

Para representar la colección de documentos, se generó un primer grafo siguiendo el mismo proceso que en el experimento I. En el caso de los artículos de los documentos recuperados, se generó un segundo grafo bajo el mismo procedimiento descrito para el experimento II.

3.2.5.2. Recuperación. Documentos: Dijkstra. Artículos, Similitud y PageRank.

Para la recuperación de documentos se utilizó el mismo mecanismo empleado para la recuperación de artículos del experimento I, es decir, dada una pregunta esta se preproceso y agregó al grafo como si se tratara de dos nuevos documentos. Posteriormente, se ejecutó el algoritmo de la ruta mínima entre los nodos correspondientes a la solicitud de información, eliminando luego los nodos de la ruta y repitiendo el proceso hasta el número deseado de documentos o hasta que el grafo se volviera inconexo. Para la recuperación de artículos se siguió el mismo procedimiento que para el experimento II.

3.2.5.3. Corpus. 37,153 artículos. 40 preguntas/respuestas

Con la finalidad de evaluar la capacidad del modelo con colecciones de mayor tamaño (se buscaban de 1 a 5 artículos-respuesta entre 37,153) se utilizó una amplia parte de la Legislación Federal y textos legislativos-normativos del Instituto Politécnico Nacional.

3.2.5.4. Comparación. Lucene, JIRS.

Debido a que Lucene y JIRS son hasta ahora los enfoques que mejores resultados han obtenido para la recuperación de diferentes clases de documentos, de forma similar al experimento II, los resultados del presente experimento se compararon con los proporcionados por tales sistemas de recuperación de información.

En el capítulo siguiente se proporcionan los resultados de los experimentos antes descritos, así como su análisis.

4. Resultados y análisis

En esta sección se presentan los resultados y un análisis descriptivo de cada uno de los experimentos.[12]

4.1. Experimentos I

A continuación, en la **Tabla 3**, se muestra la comparación de los resultados arrojados de la evaluación de los GCR, GSR y MEV en función del número de artículos-respuesta recuperados por cada enfoque.

Tabla 3. Resultados GCR, GSR y MEV en función del número de artículos-respuesta recuperados.

Grupo	IP.	Calificación		
		MEV	GSR	GCR
I	8	P	P	**C**
	4	I	P	**C**
	10	P	I	**C**
	20	C	I	**C**
II	1	P	P	**P**
	5	I	P	**P**
	19	I	C	**C**
	15	P	C	**C**
	11	C	C	**C**
III	7	I	P	**P**
	16	I	P	**P**
	9	I	C	**C**
	17	P	C	**C**
	21	I	C	**C**
IV	3	P	C	**C**
	2	C	C	**C**
	6	I	C	**C**
	12	I	C	**C**
	13	I	C	**C**
	14	I	C	**C**
	18	C	C	**C**

[12] En el Anexo C se proporciona un breve análisis enfocado a responder la pregunta ¿la diferencia entre los resultados de los métodos propuestos y los enfoques alternativos es estadísticamente significativa?

En la **Tabla 3 IP** se refiere al número de pregunta, 1-40, **Pos**., a la posición en que se encontró el artículo-respuesta, 1-75, y **Resp**., a la calificación obtenida en función de los artículos-respuesta recuperados (**C**: Correcta, **I**: Incorrecta, **P**: Parcial).

En la **Tabla 4-Tabla 7**: **IP** se refiere al número de pregunta, 1-40, **AR** a los artículos respuesta, **Documento**, al documento al que pertenece el artículo respuesta, **Referencia**, a las referencias explícitas que los artículos contienen. Y para la columna de los nombres de los documentos:

- Ley Orgánica (LO), Reglamento de Academias (RA), Reglamento de Titulación Profesional (RTP), Reglamento de las Condiciones Interiores de Trabajo del Personal Académico (RCITPA), Reglamento de las Condiciones Generales de Trabajo del Personal No-Docente (RCITPDN), Reglamento del Decanato (RD), Reglamento de Estudios Escolarizados para los Niveles Medio Superior y Superior (REE), Reglamento Interno (RI), Reglamento de Estudios de Posgrado, y el Reglamento de Servicio Social (RP).

A continuación, se da un explicación, por la cual, el GCR respondió correctamente las preguntas del grupo I, en relación con el MEV y GSR, cuyas calificaciones fueron incorrecta y parcial. En la **Tabla 4**, se aprecia que no todos los artículos respuesta estaban referenciados. No obstante, un examinación minuciosa del contenido de los artículos, indicó que en el GCR se formaron mayor número de rutas debido a que no sólo compartían palabras contenidas en las respuestas, sino otras definiciones y conceptos que implicaron una mayor semejanza entre el contenido de los artículos. Por ejemplo; la pregunta 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?

se respondía mediante 4 artículos de 3 diferentes documentos (Tabla 4), los artículos 27, 28 y 202; compartían las palabras “Consejos Técnicos Consultivos” y los 4 artículos, incluían la palabra “centros”. Debido a esto, se recuperaron los artículos 27 y 28, los cuales además, compartían las palabras: “El Subdirector Académico fungirá como secretario” con el artículo 264.

Tabla 4. Preguntas/artículos-respuesta: Grupo I.

IP.	AR.	Documento	Referencia	Posición AR.		
				MEV	GSR	**GCR**
8	28	LO	*	-	56	**14**
	264	RI	Envía al artículo 164 de la LO.	5	10	**19**
	27	LO	*	-	-	**56**
	202	RI	*	-	-	**60**
4	33	RTP	*	-	6	**1**
	24	RTP	*	-	-	**32**
10	21	LO	Referido por el art. 182 del RI.	12	-	**30**
	182	RI	Envía al artículo 21 de la LO.		-	**39**
20	21	LO	*	32	-	**22**

Tabla 5. Preguntas/artículos-respuesta: Grupo II.

IP.	AR.	Documento	Referencia	Posición AR.		
				MEV	GSR	**GCR**
1	206	RI	*	1	2	**1**
	209	RI	Envía a los artículos 27 y 28 de la LO.	2	1	**2**
	207	RI	*	-	36	**22**
	28	LO	Referido del art. 209 del RI.	-	44	**28**
	213	RI	*	-	-	**-**
5	168	RI	Envía a los arts. 14 y 21 de la LO.	-	16	**6**
	21	LO	Referido del art. 168 del RI.	-	54	**7**
	14	LO	Referido del art. 168 del RI.	-	-	**-**
15	175	RI	Envía al art. 174 fracc. II del RI.	1	4	**1**
	174	RI	Referido del art. 175 del RI.	-	69	**28**
19	23	RD	Envía al art. 20 del RD.	-	42	**26**
	20	RD	Referido del art. 23 del RD.	-	70	**34**
11	215	RI	*	5	35	**10**

De la **Tabla 5** se destaca que la mayoría de los artículos recuperados tenían palabras contenidas en la pregunta; además de que tenían referencias entre sí. Lo cual fue evidencia de que éstas mejoraron la posición. Por esta razón, las respuestas de GCR estuvieron mejor ubicadas que las del GSR. En contraste, en la **Tabla 6**, se observa que algunos artículos no contenían referencias o las que contenían no estaban relacionadas directamente con la pregunta. Por ejemplo; la pregunta 17 se respondía con los artículos 44 del Reglamento de Estudios Escolarizados y 81 del Reglamento Interno. Ambos artículos hacían referencia a otros que no eran parte de la respuesta. Adicionalmente, se notó que el contenido de los artículos respuesta era demasiado extenso. Con las características mencionadas inferimos que el contenido extenso de los artículos favoreció la semejanza al compartir más palabras con la pregunta. Mientras que las referencias que no tenían relación con la pregunta, generaban rutas adicionales que provocaron que los artículos respuesta se alejaran de las primeras posiciones.

Tabla 6. Preguntas/artículos-respuesta: Grupo III.

IP.	AR.	Documento	Referencia	Posición AR.		
				MEV	GSR	**GCR**
7	264 RI	RI	*	-	8	**8**
	29 LO	LO	*	-	9	**70**
	14 LO	LO	*	-	-	**-**
16	6	RCITPA	*	-	9	**16**
	208	RI	*	-	-	**-**
9	175	RI	Envía al art. 174 fracc. II del RI	-	12	**13**
	174	RI	Referido del art. 175 del RI	-	38	**59**
17	44	REE	Envía al art. 72 del REE	3	2	**20**
	81	RI	Envía al art. 110 del RI	-	34	**57**
21	173	RI	*	-	30	**46**

Finalmente, en la **Tabla 7**, se observó una situación similar a la de la **Tabla 5**: los artículos no tenían una referencia directa entre ellos. Sin embargo, las preguntas se caracterizaron por responderse mediante un solo artículo, con relativamente pocos términos (sobre definiciones y funciones muy específicas); es por esto que la semejanza tomo especial

relevancia, ayudando a que el artículo-respuesta se ubicará entre las primeras posiciones, hecho que se reflejó en los resultados del GSR y GCR. Adicionalmente, debido a que los artículos-respuesta no contenían referencias implico que no se generan rutas adicionales, y como consecuencia la mayoría de los artículos-respuesta se encontró dentro de las 10 primeras posiciones tanto con el GSR como con el GCR.

Tabla 7. Preguntas/artículos-respuesta: Grupo IV.

IP.	AR.	Documento	Posición AR.		
			MEV	GSR	**GCR**
3	43	RTP	8	1	**3**
	13	RTP	70	5	**8**
14	4	RP	-	2	**3**
2	181	RI	16	5	**2**
6	4	LO	-	22	**23**
12	11	RSS	13	3	**2**
13	80	RCITPA	59	67	**66**
18	200	RI	20	2	**5**

4.2. Experimento II

Con la finalidad de facilitar la comparación de los resultados de los primeros 3 experimentos entre sí, y a su vez con los obtenidos con Lucene y JIRS, los de todos se resumen en las mismas tablas. En la **Tabla 8** se reporta el número de preguntas cuya lista de artículos fue evaluada como **C**orrecta/**I**ncorrecta/**P**arcial, mientras que en la **Tabla 9** se reporta el número de preguntas cuyos artículos-respuesta se encontraron en las primeras *n* posiciones (con n = 10, 30,…, 75). La misma tabla incluye la columna **Ac.**, la cual corresponde a la suma de preguntas respondidas en el intervalo [1, *n*].

Tabla 8. Resultados SSim, SPR, Sem. & PR, Lucene y JIRS. experimento II. Colección A.

Respuesta	**Lucene**	**JIRS**	**SPR.**	**SSim.**	**Sem. & PR.**
Correcta	31	32	5	10	37
Incorrecta	0	0	25	18	0
Parcial	9	8	10	12	3

De los resultados de la **Tabla 8** se puede observar que los mejores resultados se obtuvieron mediante la combinación del valor de similitud entre nodos y el valor asociado a cada vértice dado por el algoritmo de análisis de citas al responder 6 preguntas correctas más que Lucene y 5 más que JIRS.

Tabla 9. Número de preguntas cuyos artículos-respuesta se encontraron dentro de las primeras *n* posiciones. Experimento II. Colección A.

Posición	Lucene	Ac.	JIRS	Ac.	SPR	Ac.	SSim.	Ac.	Sem. &PR.	Ac.
10	**15**	**15**	**14**	**14**	-	0	9	9	**20**	**20**
20	**5**	**20**	**9**	**23**	-	0	-	9	**2**	**22**
30	**3**	**23**	**4**	**27**	-	0	-	9	**8**	**30**
40	3	26	3	30	1	1	-	9	3	33
50	2	28	-	-	1	2	-	9	-	-
60	-	28	2	32	1	3	1	10	1	34
70	3	31	-	32	2	5	-	10	3	37
75	-	31	-	32	-	5	-	10	-	37

Con respecto a la posición de los artículos-respuesta (**Tabla 9**) es interesante que los artículos-respuesta para la mayoría de preguntas en el caso en que sólo se empleó la similitud (SSim) se recuperaron dentro de las primeras 10 posiciones, en tanto que para el caso SPR se obtuvieron en posiciones mayores a 30. Para el caso de la combinación, los artículos-respuesta para 30 de las 37 preguntas respondidas correctamente se encontraron dentro de las primeras 30 posiciones, 3 preguntas más que JIRS y 7 más que Lucene. En la **Tabla 10** se resumen los resultados de los experimentos secundarios. Nuevamente, con el objetivo de facilitar su comparación, se presentan en la misma tabla los resultados de la implementación del modelo propuesto, así como los de Lucene y JIRS con ambas colecciones de documentos (B y C).

Tabla 10. Resultados Sem. & PR., Lucene y JIRS. Experimento II. Colecciones B y C.

Método/Respuesta	Colección B(C)		
	Correcta	Incorrecta	Parcial
Sem. & PR.	35(33)	0(2)	5(5)
Lucene	31(30)	0(1)	9(9)
JIRS	31(30)	0(0)	9(10)

De la **Tabla 10** se puede observar que Sem. & PR., respondió correctamente más preguntas que Lucene y JIRS con ambas colecciones de documentos; sin embargo, para la colección C no respondió 2 preguntas, en comparación con Lucene que no respondió sólo una pregunta y JIRS cuyas respuestas solo fueron correctas o parciales. Es apreciable el hecho que al pasar de 1,117 artículos (porcentaje de respuestas correctas: 93%) a 8,987 artículos (porcentaje preguntas correctas: 83%) sólo se registró una disminución de un 10% en los aciertos de Sem. & PR.. A continuación se analizan detalladamente los resultados de los experimentos principales por cada bloque en que se clasificaron las preguntas, precisamente en función de características de las mismas preguntas/artículos-respuesta, así como del mecanismo de recuperación de cada enfoque. Primero, se presenta una tabla con los resultados a cada pregunta y posteriormente un análisis por cada enfoque.

En las **Tabla 11**-**Tabla 14** se incluye: el identificador de cada pregunta, la calificación obtenida por cada enfoque: **P**arcial/**I**ncorrecta; y en el caso de las respuestas correctas se reporta el intervalo, [1,*n*], con n = 10,…, 75, en que se encontraron los artículos-respuesta. Después se proporciona una pregunta y sus correspondientes artículos-respuesta tal como aparecen en los textos originales y tras el preprocesamiento. A partir de la pregunta ilustrada se realiza el análisis de los resultados. Las observaciones realizadas son aplicables para el resto de preguntas/respuestas del bloque, con excepción en donde se indique.

Tabla 11. Evaluación de la respuesta a preguntas del tipo I.

Pregunta	**Calificación**				
	SPR.	**SSim.**	**Sem. &PR.**	**Lucene**	**JIRS**
1	I	P	**P**	P	P
5	P	I	**10**	P	40
9	P	I	**10**	P	P
15	P	P	**10**	20	20
25	P	P	**10**	P	P
36	40	I	**10**	20	20
10	I	I	**20**	30	60
23	P	P	**30**	40	40
19	P	I	**30**	P	P

Como se observa en la **Tabla 11**, los mejores resultados correspondieron a la implementación Semejanza y PageRank (Sem. & PR), puesto que respondió 8 de las 9 preguntas, mientras JIRS 5 y Lucene sólo 4. Es importante destacar que para los casos en los que se recuperaron los artículos-respuesta, éstos se encontraron dentro de las primeras 30 posiciones en comparación con los recuperados por JIRS o Lucene, que se obtuvieron en el intervalo [20,60]. Otra observación importante es el hecho de que mientras SPR., y SSim sólo respondieron 1 y 0 preguntas respectivamente, en la combinación (Sem. & PR), se encontró respuesta dentro de las primeras 10 posiciones para 5 de 8 preguntas.

(13) Ejemplo pregunta/respuesta tipo I.

Texto original de la pregunta y artículos-respuesta.	**Pregunta y artículos-respuesta en la forma utilizada en los experimentos**
10. ¿Procede que una persona que en dos ocasiones ha sido **subdirector** **de una** **escuela**, **centro**, **o** **unidad** **investigación** 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?**	proceder persona ocasión ser **subdirector** **escuela** **centro** **unidad** **investigación** participar proceso elección terna designación **subdirector** técnico administrativo académico caso ejercer **otro** **vez** **cargo**
Artículo 21 Ley Orgánica del IPN. Los directores de **escuelas**, **centros** **y** **unidades** de enseñanza y de **investigación**, deberán ser mexicanos y reunir además los requisitos que establezcan las normas internas que rijan en el Instituto. Durarán en su **cargo** tres años y podrán ser designados, por una sola **vez**, para **otro** período.	director **escuela** **centro** **unidad** enseñanza **investigación** deber mexicano reunir además requisito establecer norma interno regir instituto durar **cargo** año poder designar **vez** **otro** período
Artículo 182 Reglamento Interno del IPN. Los directores y **subdirectores** de **escuelas**, **centros** y **unidades** de enseñanza y de **investigación** cesarán en sus funciones por cualesquiera de las siguientes causas: I. Por renuncia; II. Al término de la duración de su encargo, conforme a lo dispuesto por el ***artículo 21 de la Ley Orgánica***, o III. Por disposición del director general, cuando así lo requiera el interés institucional o la buena marcha de la **escuela**, **centro** o **unidad**.	director **subdirector** **escuela** **centro** **unidad** enseñanza **investigación** cesar función cualesquiera siguiente causa renuncia término duración encargo conforme dispuesto ley orgánico disposición director general cuando así requerir interés institucional marcha **escuela** **centro** **unidad.**

PageRank. Debido a que no todos los artículos-respuesta de cada pregunta contenían referencias, PR sólo fue capaz de responder una pregunta correctamente; sin embargo, aún las referencias ayudaron en la recuperación dado que 6 preguntas se respondieron parcialmente y sólo una de forma incorrecta.

Similitud. Se encontró que las preguntas y sus artículos-respuesta compartían términos en común, los que podrían haber facilitado su recuperación. Además, lematizar ayudó a aumentar el número de términos en común entre la pregunta y sus artículos respuesta, y en consecuencia, su similitud. Como se puede observar en la pregunta del ejemplo **(13)**, una de las preguntas contenía las palabras, escuela, centro y unidad, mientras que en los artículos-respuesta aparecían las mismas palabras (incluso repetidas) pero en plural. Así, el proceso de lematización convirtió las palabras en esencialmente las mismas, contribuyendo con esto a la similitud. No obstante lo anterior, la cantidad de información contenida en cada artículo-respuesta era muy extensa en relación con la de las preguntas de los demás bloques, lo que dificultó su recuperación (todas las preguntas del bloque se respondían a partir de 2 hasta 5 artículos). Por lo tanto, sólo 4 preguntas se respondieron de forma parcial, 5 de forma incorrecta y 0 correctas.

Combinación. Con respecto a la combinación de PR y la similitud, se encontró que ambos contribuyeron a mejorar el proceso de recuperación. Así, el combinar las bondades de la semejanza y el PR, produjo que los artículos que tenían una referencia y semejanza con la pregunta, aunque fuera baja, se pudieran contestar correctamente debido a que generaron nuevas y más eficientes rutas de recuperación.

Como se puede observar en la **Tabla 11**, en los casos en que los resultados con PR fueron relativamente buenos en comparación con los de sólo la similitud, la combinación mejoró los resultados de los dos enfoques individuales. De igual manera sucedió en la situación contraria. Aún si los resultados de PR., y la similitud individuales no eran buenos, su combinación mejoró la respuesta. Por ejemplo, para la pregunta ilustrada SSim., y SPR., no recuperaron los artículos-respuesta. Esto ocurrió debido a que por una parte, sólo un artículo contenía una referencia, por lo que PR no tuvo impacto; por otro lado, la respuesta estaba conformada por artículos que, a pesar de que contenían términos en común con la solicitud, eran relativamente comunes en los textos, por lo que su similitud no fue suficiente

para lograr recuperar los artículos. Sin embargo, en la combinación, que la similitud fuera baja pero no nula y que un artículo contuviera al menos una referencia fue de ayuda para lograr recuperar los artículos-respuesta. Cabe mencionar, sin embargo, que los artículos-respuesta fueron encontrados algo alejados de las primeras posiciones. Esto se ilustra también en los ejemplos (14), (15) y (16), los cuales incluyen los resultados detallados para 3 preguntas del Bloque I. Los ejemplos incluyen una Pregunta, sus Artículos-Respuesta, el o los artículos a los cuales los artículos-respuesta hacían Referencia y la posición en que se obtuvieron con SPR., SSim., y Sem. & PR.

(14) ¿Qué debe entenderse por voto de calidad del Maestro Decano en el proceso de elección de terna para la designación de director?

Artículos-Respuesta	**Referencia**	**Posición**		
		SPR.	**SSim.**	**Sem. & PR.**
29 LO	21 LO	18	-	**4**
13 RD	24 LO, 149 RI	-	-	**6**
179 RI	29 LO, 207 RI	2	-	**8**
153 RI	179 RI	38	1	**1**
193 RI	295 RI	31	-	**30**

(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?

174 RI	-	14	-	**9**
175 RI	174 RI	-	1	**1**

(16) ¿Qué debe entenderse por tener estudios de posgrado?

7 REE	79 RI	-	3	**3**
79 RI	81 RI	11	2	**2**
81 RI	110 RI	-	-	**4**
110 RI	106 RI	23	-	**5**

Lucene. No obstante que las preguntas y sus artículos-respuesta contenían términos en común que podrían haber facilitado a Lucene su recuperación, la cantidad de información requerida afectó también su desempeño por lo que sólo respondió 4 preguntas correctamente y 5 de manera parcial.

Los artículos-respuesta para las preguntas respondidas correctamente se encontraron en todos los casos más alejados de las primeras posiciones en comparación que con Sem. & PR.

JIRS. De manera similar a Lucene, a pesar de que los artículos-respuesta compartían términos en común e incluso n-gramas con la solicitud, JIRS solo respondió 5 preguntas correctamente y 4 de forma parcial. Esto debido a que se encontró que la mayoría de n-gramas que se formaban a partir de los términos de la pregunta, aparecían en un alto número de artículos, lo que en general dificultó el proceso de recuperación Por ejemplo, para la pregunta ilustrada. Los bi-gramas, tri-gramas y tetra-gramas que se formaron de <**escuela centro unidad investigación**> resultaron relativamente comunes en los textos, sólo **<subdirector escuela centro unidad investigación>** resulto menos frecuente; por ello, a pesar de recuperar los artículos-respuesta, éstos se recuperaron en posiciones alejadas del inicio (<=60), en comparación con Sem. & PR., el cual para la misma pregunta los recuperó dentro de las primeras 20 posiciones. En general los resultados del método propuesto resultaron mejores que los obtenidos con JIRS y Lucene.

Ahora se proporcionan los resultados y análisis de las preguntas en función de los rasgos de sus artículos-respuesta.

Tabla 12. Evaluación de la respuesta a preguntas tipo II.

Pregunta	**Calificación**				
	SPR.	**SSim.**	**Sem. &PR.**	**Lucene**	**JIRS**
17	I	P	**30**	20	20
21	50	I	**40**	20	10
39	I	P	**40**	20	30
8	P	P	**60**	40	40
7	P	I	**70**	P	P
16	I	P	**P**	P	P

(17) Ejemplo pregunta/respuesta tipo II.

Texto original de la pregunta y artículos-respuesta.	Pregunta y artículos-respuesta en la forma utilizada en los experimentos
17. ¿Un **alumno** que **adeuda** cinco materias **puede** seguir fungiendo como consejero representante de los **alumnos**?	**alumno** **adeudar** materia **poder** seguir fungir consejero representante **alumno**
Artículo 44. Reglamento de Estudios Escolarizados. Los **alumnos** que al inicio del periodo de reinscripciones **adeuden** cuatro asignaturas o más, causarán baja y por lo tanto no podrán solicitar su reinscripción, salvo que obtengan resolución favorable de la Comisión de Situación Escolar del Consejo Técnico Consultivo Escolar, en los términos previstos por el *artículo 72 del presente Reglamento*.	**alumno** inicio periodo reinscripciones **adeudar** asignatura causar bajo tanto poder solicitar reinscripción salvo obtener resolución favorable comisión situación escolar consejo técnico consultivo escolar término previsto presente reglamento
Artículo 81. Reglamento Interno. La calidad de **alumno** de la modalidad escolarizada se pierde por: I. La conclusión del plan de estudios; II. Baja voluntaria, temporal o definitiva; III. **Adeuda**r el número de asignaturas o sus equivalentes que fijen los reglamentos aplicables, y IV. Resolución definitiva dictada por la instancia institucional competente, en los casos previstos por el *artículo 110, fracciones V y VI, del presente Reglamento*. Los egresados que estén en proceso de obtener su título profesional o grado académico podrán continuar accediendo a los servicios educativos que correspondan, en los términos del reglamento respectivo.	calidad **alumno** modalidad escolarizada perder conclusión plan estudio bajo voluntaria temporal definitivo **adeudar** número asignatura equivalente fijar reglamento aplicable resolución definitivo dictar instancia institucional competente caso previsto fracción presente reglamento egresado estar proceso obtener título profesional grado académico poder continuar acceder servicio educativo corresponder término reglamento respectivo

PageRank. No todos los artículos-respuesta de cada pregunta contenían referencias, y los que sí las contenían, sólo hacían referencia a un artículo (sólo para la pregunta 7 uno de sus tres artículos-respuesta hacía referencia a dos artículos); por ello, artículos con más referencias se recuperaron en lugar de los artículos-respuesta y en consecuencia 3 de 6 preguntas se respondieron incorrectamente, mientras que sólo 1 se respondió correctamente y 2 de forma parcial.

Similitud. Las preguntas compartían pocos términos con sus artículos-respuesta y/o los términos de las preguntas aparecían en un alto número de artículos; por lo cual, sólo usando similitud no fue posible recuperar los artículos-respuesta. Por ejemplo, la pregunta 17 sólo compartía las palabras alumno y alumnos con los artículos-respuesta, ver ejemplo (17).

No obstante lo anterior, haber lematizado aumentó la similitud entre los artículos-respuesta y las solicitudes; por ello, se respondieron al menos 4 preguntas de forma parcial mientras que sólo 2 se respondieron incorrectamente. Por ejemplo, la pregunta 17 contenía, entre otras, las palabras *adeuda* y *puede* mientras que los artículos contenían *adeuden*, *adeudar*, y *podrán*. Lematizar produjo *adeudar* y *poder* tanto en la pregunta como en los artículos-respuesta, debido a lo cual aumentó su similitud entre sí permitiendo la recuperación de al menos uno de sus dos artículos-respuesta.

Combinación. De manera similar a las preguntas del Tipo I, la combinación de la similitud y PR mejoró sus resultados individuales. El hecho de que los artículos-respuesta contuvieran referencias, y que la similitud entre los artículos-respuesta y la pregunta, si bien baja, no fuera nula, permitieron responder la mayor parte de preguntas. Sin embargo, la baja similitud entre preguntas y sus artículos-respuesta, que no todos los artículos-respuesta a cada pregunta contuvieran referencias y que los que contenían sólo hacían referencia a un artículo, provocó que los artículos-respuesta se encontraran alejados de las primeras posiciones, en comparación con Lucene y JIRS.

Lucene. Se encontró que en los artículos-respuesta de 4 preguntas aparecían los términos de la pregunta en más de dos ocasiones (esto favoreció el mecanismo de recuperación de Lucene) lo que le permitió responder correctamente 4 de las 6 preguntas y las otras dos de forma parcial. Por esta misma razón la posición de los artículos respuesta fue mejor que con respecto a Sem. & PR., y similar a JIRS.

JIRS. En las mismas preguntas que Lucene respondió correctamente se encontró que los n-gramas que contenían aparecían más de una vez en los artículos-respuesta; por ello, JIRS obtuvo resultados parecidos a Lucene.

A continuación se proporcionan los resultados y análisis para las preguntas del tercer tipo.

Tabla 13. Evaluación de la respuesta a preguntas del tipo III.

Pregunta	Calificación				
	SPR.	**SSim.**	**Sem. &PR.**	**Lucene**	**JIRS**
35	I	I	**P**	P	P
3	I	I	**10**	20	20
11	I	I	**10**	10	10
18	I	60	**10**	20	20
22	I	I	**10**	20	30
26	I	I	**10**	10	10
37	P	P	**10**	10	10
33	70	P	**20**	10	20
4	I	I	**30**	50	P
6	I	I	**30**	10	10
12	I	P	**30**	30	20
20	I	I	**30**	10	10
34	P	I	**30**	40	20
13	I	I	**40**	40	30
14	I	P	**70**	P	60
28	I	I	**70**	30	30

Como puede observarse de la **Tabla 13** la combinación de PR con la similitud mejoró sus resultados individuales. Es relevante el hecho de que a pesar de que en el caso en el que la respuesta dada por SPR y SSim fue incorrecta, con Sem. & PR., fue correcta y además que en 12 de 15 casos los artículos-respuesta hayan sido obtenidos dentro de las primeras 30 posiciones, logrando incluso obtener resultados ligeramente mejores que Lucene y JIRS.

(18) Ejemplo pregunta/respuesta tipo III.

Texto original de la pregunta y artículos-respuesta.	Pregunta y artículos-respuesta en la forma utilizada en los experimentos
P20. ¿Una persona **puede** ser **director** de una **escuela centro** o **unidad** de investigación por tercera ocasión?	persona **poder director escuela centro unidad investigación** ocasión
Artículo 21. LO. Los **directores** de **escuelas**, **centros** y **unidades** de enseñanza y de investigación, deberán ser mexicanos y reunir además los requisitos que establezcan las normas internas que rijan en el Instituto. Durarán en su cargo tres años y **podrán** ser designados, por una sola vez, para otro período.	**director escuela centro unidad** enseñanza **investigación** deber mexicano reunir además requisito establecer norma interno regir instituto durar cargo año **poder** designar vez otro período

PageRank. Debido a que los artículos-respuesta no contenían referencias a otros artículos, su valor de PR resultó menor con respecto a artículos-respuesta con referencias como los de los bloques anteriores; de aquí que se tuviera un bajo desempeño. Este método sólo logró responder una pregunta en comparación con las 13 incorrectas, dos parciales y sólo una correcta.

Similitud. Para esta clase de preguntas sólo la similitud entre la solicitud y los artículos-respuesta contribuyó a su recuperación; sin embargo, los pocos términos que compartían la solicitud y los artículos-respuesta en general dificultaron el proceso de recuperación. Para el ejemplo, el artículo 2 del REE se recuperó en la primera posición: "Artículo 2 del Reglamento de Estudios Escolarizados. Es alumno la **persona** inscrita en cualquier programa académico que se imparta en las **escuelas**, **centros** o **unidades** de enseñanza".

Por otra parte, artículos con más términos que los artículos-respuesta y en consecuencia con más términos de la solicitud, fueron favorecidos y recuperados antes que los artículos-respuesta. Siguiendo con el mismo ejemplo, el Artículo 176 del RI se recuperó en la segunda posición:

> Son facultades y obligaciones de los subdirectores de **escuelas, centros y unidades** de enseñanza y de **investigación**: I. Auxiliar al **director** de la **escuela, centro o unidad** en el ejercicio de sus funciones; II. Dirigir y coordinar las

> actividades de la subdirección a su cargo; III. Desempeñar las comisiones que el **director** de la **escuela, centro o unidad** les encomiende; IV. Acordar con el **director** de la **escuela, centro o unidad**; V. Despachar los asuntos de su competencia; VI. Presentar al **director** de la **escuela, centro o unidad** las sugerencias de carácter académico, técnico y administrativo que consideren pertinentes para mejorar los servicios educativos a su cargo; VII. Integrarse a los programas de desarrollo directivo del Instituto, y VIII. Las demás que se requieran para cumplir con las anteriores, las que se deriven de la Ley Orgánica, del presente Reglamento y de otros ordenamientos jurídicos y administrativos internos aplicables.

Es importe destacar el hecho que el lematizar ayudo a mejorar ligeramente el proceso de recuperación, con respecto al caso en el que solo se empleó PR. Esto debido a que lematizar incrementó la similitud entre la solicitud y los artículos respuesta. Como se puede observar en la pregunta ilustrada, las palabras *podrán* y *puede*, *director* y *directores* después de lematizar se volvieron las mismas palabras: *poder* y *director*, respectivamente, incrementado con ello la similitud entre la solicitud y los artículos-respuesta. Sin embargo, la mejora no fue suficiente para lograr responder las preguntas, como se puede observar de los resultados, ya que sólo 1 pregunta se respondió correctamente, 11 de forma incorrecta y sólo 4 se respondieron parcialmente.

Combinación. Resultó interesante que la combinación mejoró notablemente los resultados individuales obtenidos con la similitud y PR, puesto que se lograron responder de esta forma 15 preguntas correctamente, en comparación con sólo 1 que se respondió con la similitud y PR por separado. En este caso PR ayudó en la recuperación debido precisamente a que los artículos-respuesta no contenían referencias. PageRank, al favorecer a artículos con referencias, alejó a artículos que a pesar de ser similares a la pregunta no la respondían, lo que su vez, provocó que se mejorara la posición de los artículos-respuesta.

Lucene. De forma similar que para las preguntas del bloque anterior, se encontró que los términos de la pregunta aparecían repetidos en los artículos respuesta, situación que favoreció a Lucene y que le permitió responder 14 de las 16 preguntas de este bloque.

JIRS. A pesar de que las preguntas de este bloque y sus artículos respuesta compartían pocos términos en común, el hecho de que los términos que contenían las preguntas contenían n-gramas que también aparecían en los artículos-respuesta, ayudó a responder la mayor parte de las preguntas de este tipo.

Finalmente se muestran los resultados y análisis para las preguntas del cuarto tipo.

Tabla 14. Evaluación de la respuesta a preguntas del tipo IV.

Pregunta	Calificación				
	SPR.	SSim.	Sem. &PR.	Lucene	JIRS
2	I	10	**10**	10	10
24	I	10	**10**	10	10
27	**70**	10	**10**	10	10
29	I	10	**10**	10	10
30	I	10	**10**	10	10
31	I	10	**10**	10	10
32	I	10	**10**	10	10
38	I	10	**10**	10	**20**

(19) Ejemplo pregunta/respuesta tipo IV.

Texto original de la pregunta y artículos-respuesta.	Pregunta y artículos-respuesta en la forma utilizada en los experimentos
P24 ¿Es procedente que un **trabajador** solicite que la media hora, que tiene asignada para **tomar alimentos** y/o **descanso** a **mitad** de la **jornada**, se le otorgue a la entrada o a la salida del trabajo?	procedente **trabajador** solicitar medio hora tener asignar **tomar alimento descanso mitad jornada** otorgar entrada salida trabajo
Artículo 49. RCITPA. Los **trabajadores** que presten sus servicios por **jornada** continua tendrán derecho a los treinta minutos diarios, programados a **mitad** de la **jornada**, para la **toma de alimentos** o **descanso**. Ese tiempo será considerado como efectivamente laborado.	**trabajador** prestar servicio **jornada** continuo derecho treinta minuto diario programar **mitad jornada** **tomar alimento descanso** tiempo considerado efectivamente laborar.

PageRank. Debido a que los artículos-repuesta no contenían referencias, el enfoque de PR fue el menos favorecido. PR sólo recuperó los artículos-respuesta en dos casos. Esencialmente, los artículos-respuesta fueron recuperados debido a que otros artículos hacían referencia a ellos, lo que incrementó su valor de PR.

Similitud. Para esta clase de preguntas la similitud tuvo un papel primordial, puesto que a partir de ésta es que se recuperaron los artículos-respuesta. Adicionalmente, ésta se benefició por el proceso de lematización al incrementarse con éste el número de términos similares entre la solicitud y los artículos-respuesta.

Combinación. En la combinación de la similitud y PR, este último no benefició o afectó al proceso de recuperación, lo que se puede observar a partir de que con sólo la similitud y su combinación con PR se obtuvieron esencialmente los mismos resultados: con ambos los artículos-respuesta para las 9 preguntas se obtuvieron dentro de las primeras 10 posiciones.

Lucene. Puesto que la similitud es la que ayudó en el proceso de recuperación y dado que Lucene se basa en la similitud coseno, la cual es también utilizada en la implementación del método propuesto, con ambos se obtuvieron los mismos resultados.

JIRS. El número de términos en común entre los artículos-respuesta y la solicitud que favoreció la recuperación por similitud, ayudó también al mecanismo de recuperación de JIRS. Además, debido al hecho de que los términos en común formaron n-gramas, JIRS obtuvo los mismos resultados que Lucene y Sem. & PR. La excepción en JIRS fue la pregunta 38, la cual se respondía con 3 artículos, dos de ellos se encontraron en posiciones mayores a 10. Esto es porque a pesar de que los artículos-respuesta compartían términos en común con la solicitud, éstos no aparecían juntos, lo que no favoreció a JIRS.

4.3. Experimento III

En esta sección se muestran los resultados del enfoque de recuperación de documentos y artículos empleando el modelo del grafo ponderado no dirigido, el algoritmo de la ruta mínima y el algoritmo de análisis de referencias PageRank. En la **Tabla 15**, se resumen los resultados del tercer experimento en función únicamente de los artículos-respuesta recuperados.

Tabla 15. Resultados globales para el conjunto de 37,153 artículos.

Respuesta	Lucene	JIRS	Sem. & PR.
Correcta	27	27	**37**
Incorrecta	1	2	**0**
Parcial	12	11	**3**

Es en este experimento el enfoque de recuperación mediante el grafo obtuvo los mejores resultados, dejando sin responder únicamente 3 preguntas de las 40 utilizadas en la evaluación, y en tales casos la respuesta fue parcial. Mientras que en el experimento anterior la implementación del modelo propuesto sólo superó con 5 y 6 respuestas correctas más a JIRS y a Lucene, respectivamente, en el presente experimentó se logró alcanzar una diferencia de 10 preguntas. Vale la pena recordar que mientras en el experimento anterior sólo se utilizaron 8,987 artículos, en éste se utilizaron 37,153. De manera preliminar podemos establecer que la reducción del espacio de búsqueda sobre documentos fue mejor que sobre los artículos mismos; esto con base en que precisamente este último mecanismo es el que emplean Lucene y JIRS. Adicionalmente, como se verá a continuación, la posición de los artículos-respuesta fue en general mejor con la implementación del modelo propuesto que con Lucene y JIRS.

El análisis del experimento II es aplicable a la recuperación de artículos de este experimento; debido a esto, el siguiente análisis se centra en los casos para los cuales no se encontró respuesta en los experimentos I, II y III, y en la recuperación de documentos.

En los encabezados de las siguientes tablas: **TP** corresponde al tipo de pregunta (**P**: Procedimiento, **S**: Sí/No, **D**: Definición). **NP**: 1-40, al número de pregunta. **PD**: 1-10, a la posición en la que se recuperó el documento que contenía al artículo-respuesta. **AR.Doc**. a los artículos-respuesta y documento al que pertenecían. **Referencia**. A el(los) artículo(s) a los cuales hacían referencia los artículos-respuesta. Finalmente, **Posición AR**. se refiere a la posición de los artículos-respuesta por cada enfoque: Lucene, JIRS y Sem. & PR.

Tabla 16. Resultados del experimento III. Preguntas del tipo I.

TP.	NP.	PD.	AR.Doc.	Referencia	Posición AR.		
					Lucene	JIRS	Sem. & PR.
P	1	5	28LO	*	38	42	**3**
		4	206 RI	*	1	1	**1**
			207 RI	*	67	-	**5**
			209 RI	27, 28 LO	2	2	**2**
			213 RI	*	-	-	**-**
P	36	1	12 RA	*	1	1	**4**
			13 RA	12 RA	18	24	**5**
S	15	2	174 RI	*	36	46	**9**
			175 RI	174 RI	1	1	**1**
S	5	2	168 RI	14,21 LO	5	13	**7**
		1	14 LO	13 LO, 2554 CC	-	39	**5**
			21 LO	*	1	8	**4**
D	25	1	7 REE	79 RI	4	3	**3**
		3	79 RI	81 RI	3	2	**2**
			81 RI	110 RI	-	-	**4**
			110 RI	106 RI	-	-	**5**
S	9	4	174 RI	*	-	-	**2**
			175 RI	174 RI	12	39	**1**
S	10	1	21 LO	*	37	-	**20**
		2	182 RI	21 LO	23	41	**18**
D	23	2	29 LO	21 LO	7	8	**4**
		1	13 RD	24 LO, 149 RI	5	5	**6**
		5	179 RI	29 LO, 207 RI	4	6	**8**
			153 RI	179 RI	1	1	**1**
			193 RI	295 RI	-	-	**30**
S	19	1	20 RD	*	-	-	**24**
			23 RD	20 RD	23	22	**22**

Como se puede observar en la Tabla 16 los documentos que contenían los artículos-respuesta para todas las preguntas se obtuvieron dentro de las primeras 5 posiciones. Por otro lado, mientras que con Sem. & PR., se respondieron 8 de 9 preguntas, Lucene y JIRS

sólo respondieron 3 preguntas. En general, los artículos-respuesta se encontraron en mejores posiciones con el modelo propuesto que con JIRS y Lucene.

(20) Ejemplo de pregunta/respuesta del tipo I, tercer experimento.

Texto original de la pregunta y artículos-respuesta.	Pregunta y artículos-respuesta en la forma utilizada en los experimentos
1. ¿Cómo se **lleva** a **cabo** el procedimiento de **elección** de **representantes alumnos ante** el **Consejo Técnico Consultivo Escolar**?	**llevar cabo** procedimiento **elección representante alumno ante consejo técnico consultivo escolar**
Artículo 213 del Reglamento Interno. Para ser candidato a consejero alumno se requiere: I. Ser **alumno** en situación escolar regular y estar cursando estudios entre el tercero y el penúltimo semestres, para los niveles medio superior y superior, y II. Ser **alumno** de posgrado y haber acreditado más del 50 por ciento de las asignaturas que integran el plan de estudios correspondiente.	candidato consejero **alumno** requerir **alumno** situación escolar regular estar cursar estudio entre penúltimo semestre nivel medio superior superior **alumno** posgrado haber acreditar ciento asignatura integrar plan estudio correspondiente
Artículo 206 del Reglamento Interno. 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.	**elección representante** tanto personal académico **alumno ante consejo técnico consultivo escolar llevar cabo** cada año mes inmediato posterior inicio ciclo escolar término respectivo convocatoria propio **consejo** expedir

La pregunta ilustrada en el ejemplo fue la única que no se respondió en todos los experimentos con el modelo propuesto, con Lucene y con JIRS. La razón de esto se determinó a partir del análisis de los términos de la solicitud y el contenido de sus artículos-respuesta, tal como se muestra en el ejemplo Mientras que el artículo 206 del RI contenía 9 de los 10 términos de la solicitud, el artículo 213 del mismo documento sólo contenía uno, esto dificultó enormemente su recuperación.

Tabla 17. Resultados del experimento III. Preguntas del tipo II.

TP	NP	PD	AR-Doc.	Referencia	Posición AR.		
					Lucene	JIRS	Sem. & PR.
S	16	7	208 RI	27 LO	-	-	**-**
		10	6 RCITPA	*	-	21	**59**
S	17	2	81 RI	110 RI	24	27	**23**
		1	44 REE	72 REE	7	21	**9**
P	21	2	173 RI	179 RI	30	6	**31**
S	39	6	8 LFTSE	5 LFTSE	63	-	**35**
			70 LFTSE	*	6	-	**3**
S	8	1	27 LO	*	27	25	**54**
			28 LO	*	18	24	**19**
		2	202 RI	*	30	38	**48**
			264 RI	164 RI	15	21	**17**
S	7	2	14 LO	13 LO, 2554 CC	31	9	**35**
			29 LO	*	7	11	**7**
		1	264 RI	164 RI	-	-	**68**

La recuperación de los artículos-respuesta para las preguntas del segundo tipo resultó de mayor dificultad que las del tipo I, debido a que si bien sólo una de las preguntas no se respondió correctamente, los artículos-respuesta se encontraron en su mayoría (10 de 13) en el intervalo 20-70. También los documentos resultaron más difíciles de localizar; no obstante que todos se recuperaron dentro de las primeras diez posiciones, sólo siete de ellos se encontraron dentro de las primeras 2 posiciones y 3 después de la quinta posición.

Por otro lado, en lo que respecta a la posición de los artículos respuesta, los resultados que obtuvieron Lucene y JIRS fueron ligeramente mejores que los de la implementación del modelo; sin embargo, de los 14 artículos-respuesta que se requerían para responder a todas las preguntas del bloque, Lucene no logró recuperar 4, JIRS 3 y nuestra implementación sólo 1.

A continuación se analiza la pregunta del tipo II que no se logró responder en todos los experimentos, I, II y III, así como también con Lucene y JIRS.

(21) Ejemplo pregunta/respuesta sin procesar, después del preprocesamiento.

Texto original de la pregunta y artículos-respuesta.	Pregunta y artículos-respuesta en la forma utilizada en los experimentos
P16. ¿Tienen derecho de votar o no los profesores interinos adscritos a alguna escuela?	tener derecho votar profesor interino adscritos escuela
Artículo 208 del Reglamento Interno del Instituto Politécnico Nacional. El personal académico de cada programa académico, o equivalente elegirá en forma directa y por mayoría de votos a sus respectivos representantes, en los términos del artículo 27 de la Ley Orgánica.	personal académico cada programa académico equivalente elegir forma directo mayoría voto respectivo representante término ley orgánico
Artículo 6 del Reglamento de las Condiciones Interiores de Trabajo del Personal Académico del IPN. Para la debida interpretación y aplicación del presente Reglamento de las Condiciones Interiores de Trabajo del Personal Académico del Instituto Politécnico Nacional, en el curso de este instrumento, se denominará: I. SEP: A la Secretaría de Educación Pública. II. IPN: Al Instituto Politécnico Nacional. XXV… 592 términos en total.	313 términos.

Como se ilustra en el ejemplo **(21)** el artículo 208, al no tener términos en común con la solicitud, no fue recuperado en todos los experimentos (I, II y III). Si bien contenía una referencia, al haber tenido una similitud nula con respecto a la solicitud, el valor de PR asociado por sí solo no ayudó en su recuperación. Por otra parte, la recuperación del artículo 6 se dificultó debido a su extensión y a los pocos términos en común con respecto a la solicitud. Precisamente por esto último no se incluyó su contenido. Lo mismo sucedió en todos los experimentos (I-III).

Tabla 18. Resultados del experimento III. Preguntas del tipo III.

TP	NP	Pregunta		Posición artículos respuesta		
		PD	Artículos respuesta-Documento	Lucene	JIRS	Sem. & PR.
P	35	4	70 RCITPA	3	4	**5**
			77 RCITPA	-	-	**-**
			51 RCITPA	-	-	**-**
S	11	3	215 RI	22	28	**3**
D	26	1	31 REE	3	5	**3**
D	37	1	96 RP	6	4	**5**
			104 RP	3	5	**1**
			105 RP	2	3	**3**
S	18	4	200 RI	46	-	**7**
S	3	1	13 RTP	16	22	**8**
			14 RTP	9	9	**3**
S	22	1	63 RCITPA	11	9	**8**
P	33	1	14 RA	2	2	**10**
			15 RA	8	23	**11**
			16 RA	4	17	**3**
S	6	9	4 LO	9	1	**22**
S	20	1	21 LO	5	6	**23**
S	12	2	5 RSS	-	52	**22**
			11 RSS	1	1	**2**
P	34	1	23 RP	37	18	**25**
			26 RP	1	1	**26**
S	4	2	24 RTP	-	-	**29**
			33 RTP	26	35	**4**
S	13	1	80 RCITPA	56	49	**38**
D	28	5	17 RCITPA	7	27	**1**
			19 RCITPA	-	-	**3**
D	14	1	3 RP	5	5	**1**
			4 RP	-	56	**70**

Es apreciable el hecho de que la recuperación de documentos devolvió aquellos que contenían todos los artículos-respuesta para todas las preguntas del tipo III, además 9 de 16 en la posición 1, 6 de 16 dentro de las primeras 5 posiciones, y sólo 1 en la posición 9. A pesar de esto, al igual que con los tipos anteriores de preguntas, una sola de ellas no se logró responder correctamente en todos los experimentos.

(22) Ejemplo. Pregunta/respuesta sin procesar, después del preprocesamiento.

Texto original de la pregunta y artículos-respuesta.	Pregunta y artículos-respuesta en la forma utilizada en los experimentos
¿Qué **requisitos** se **deben** cubrir para **obtener** una **promoción**?	**requisito deber** cubrir **obtener promoción**
Artículo 70 del Reglamento sobre las Condiciones Interiores de Trabajo del Personal Académico del IPN. **requisitos** … **obtener** … **promoción** … **deberá obtener** … **promoción** … **promoción** … **requisitos** 142 términos más.	**requisito** … **obtener promoción** … **deber obtener** … **promoción** … **promoción** … **requisito** … 68 términos más.
Artículo 77 del Reglamento sobre las Condiciones Interiores de Trabajo del Personal Académico del IPN. **promoción** … **deberá ...** 88 términos más.	**promoción** … **deber** … 45 términos más.
Artículo 51 del Reglamento sobre las Condiciones Interiores de Trabajo del Personal Académico del IPN. 201 términos. 0 en común con la solicitud.	96 términos. 0 en común con la solicitud.

Debido a la extensión de los artículos no se incluyó su contenido completo. Como puede apreciarse en el ejemplo, para el caso del artículo-respuesta 51, éste no contenía términos en común con respecto a la solicitud, mientras el artículo 77 sólo contenía 2. Es por esto que tales artículos no fueron recuperados en todos los experimentos (I, II y III). En lo que

respecta al artículo 70, éste sí se logró recuperar, pues contenía todos los términos de la solicitud a excepción de uno; tal situación favoreció a incrementar la similitud entre ellos y en consecuencia no sólo se recuperó sino que se encontró entre las primeras 10 posiciones. Lo mismo ocurrió con Sem. & PR., Lucene y JIRS.

Tabla 19. Resultados del experimento III. Preguntas del tipo IV.

TP.	ND.	Pregunta		Posición artículos respuesta		
		PD	Artículos respuesta-Documento	Lucene	JIRS	Sem. & PR.
P	2	9	181 RI	1	17	**1**
S	24	1	49 RCITPND	2	2	**1**
D	27	2	26 RCITPA	1	2	**1**
			27 RCITPA	13	17	**3**
S	29	2	33 REE	1	1	**1**
S	30	1	30 REE	1	1	**1**
S	31	1	56 RP	2	1	**2**
			59 RP	1	2	**1**
S	32	1	139 RCITPA	1	2	**1**
			140 RCITPA	2	1	**3**
S	38	9	51 REE	11	29	**8**
		3	85 RI	3	2	**2**
			86 RI	1	1	**4**
S	40	5	30 LFTSE	3	13	**2**

Al igual que en los experimentos anteriores, las preguntas del cuarto tipo se respondieron correctamente en todos los experimentos, tanto con el modelo propuesto como con JIRS y Lucene. La cantidad de términos en común entre los artículos-respuesta y las solicitudes facilitaron su recuperación; aunado a ello, la lematización aumentó aún más los términos en común, y con ello la similitud entre ellos. Por la misma razón, en la recuperación de documentos se lograron obtener todos los que contenían los artículos-respuesta para todas las preguntas dentro de las primeras 10 posiciones, con 7 de ellos en las primeras 3 posiciones, uno en la quinta y solo 2 en la novena posición. Debido a la relativa sencillez de las preguntas, se obtuvo la respuesta correcta para todas en este experimento como en el anterior, lo que refuerza el análisis ya realizado.

Finalmente, se muestra una comparación (informal) entre los resultados del sistema de consulta de información jurídica INFOJUS (del Instituto de Investigaciones Jurídicas de la Universidad Nacional Autónoma de México) y los de la implementación del modelo propuesto. Para satisfacer la necesidad de información sobre cuáles son las disposiciones relevantes en la legislación federal sobre ciencia y tecnología. Para ello se formuló la solicitud de información simple: *ciencia AND tecnología* en INFOJUS, y *ciencia y tecnología* en la interfaz de la implementación del modelo propuesto en su versión del experimento III. En cuanto al término *informal*, éste se refiere a que la comparación no se puede justificar completamente, puesto que INFOJUS y nuestra implementación manejan diferentes colecciones de documentos. INFOJUS se compone de 279 textos sobre legislación federal, mientras que la implementación del modelo propuesto utiliza 288 documentos legislativos-normativos: 261 sobre legislación federal y 26 relativos a normatividad del IPN, ver **Tabla 20**.

Tabla 20. Colecciones de documentos: implementación del modelo propuesto y el sistema de consulta de información jurídica INFOJUS.

Tipo de documentos	**Implementación**	**INFOJUS**[13]
Constitución Política de los Estados Unidos Mexicanos		
Ley	245	268
Estatuto	1	1
Código	9	8
Reglamento	31	0
Ordenanza	1	1
Total	**288**	**279**

Con la implementación del modelo propuesto para la solicitud *ciencia y tecnología* se obtuvo el artículo 3 de la Ley de Ciencia y Tecnología como primera opción, mismo que se reproduce a continuación:

Artículo 3 de la Ley de Ciencia y Tecnología.

[13] Consulta realizada el 22-11-2012.

Para los efectos de esta Ley, el Sistema Nacional de Ciencia, Tecnología e Innovación se integra por:

I. La política de Estado en materia de ciencia, tecnología e innovación que defina el Consejo General;
II. El Programa Especial de Ciencia, Tecnología e Innovación, así como los programas sectoriales y regionales, en lo correspondiente a ciencia, tecnología e innovación;
III. Los principios orientadores e instrumentos legales, administrativos y económicos de apoyo a la investigación científica, el desarrollo tecnológico y la innovación que establecen la presente Ley y otros ordenamientos;
IV. Las dependencias y entidades de la Administración Pública Federal que realicen actividades de investigación científica, desarrollo tecnológico e innovación o de apoyo a las mismas, así como las instituciones de los sectores social y privado y gobiernos de las entidades federativas, a través de los procedimientos de concertación, coordinación, participación y vinculación conforme a ésta y otras leyes aplicables, y
V. La Red Nacional de Grupos y Centros de Investigación y las actividades de investigación científica de las universidades e instituciones de educación superior, conforme a sus disposiciones aplicables.

El resto de resultados 8 de 10 también correspondieron a artículos pertenecientes a la Ley de Ciencia y Tecnología, como se ilustra en la **Figura 14** y sólo dos de ellos a normatividad del IPN. En dicha figura se encuentran los resultados para la solicitud *ciencia y tecnología*. Cada artículo está acompañado (en la figura está oculto) de su correspondiente texto, además también el nombre del documento al que pertenece es un *hipervínculo,* el cual se incluyó para permitir la consulta de los artículos transitorios u otra información relevante sobre el documento al que pertenece un artículo en particular.

Una observación importante es que el artículo 10 de la Ley de Ciencia y Tecnología hace referencia al artículo 41 de la misma Ley. Al considerar las referencias entre artículos la implementación del modelo propuesto es capaz de recuperar ambos artículos facilitando de esa forma su consulta.

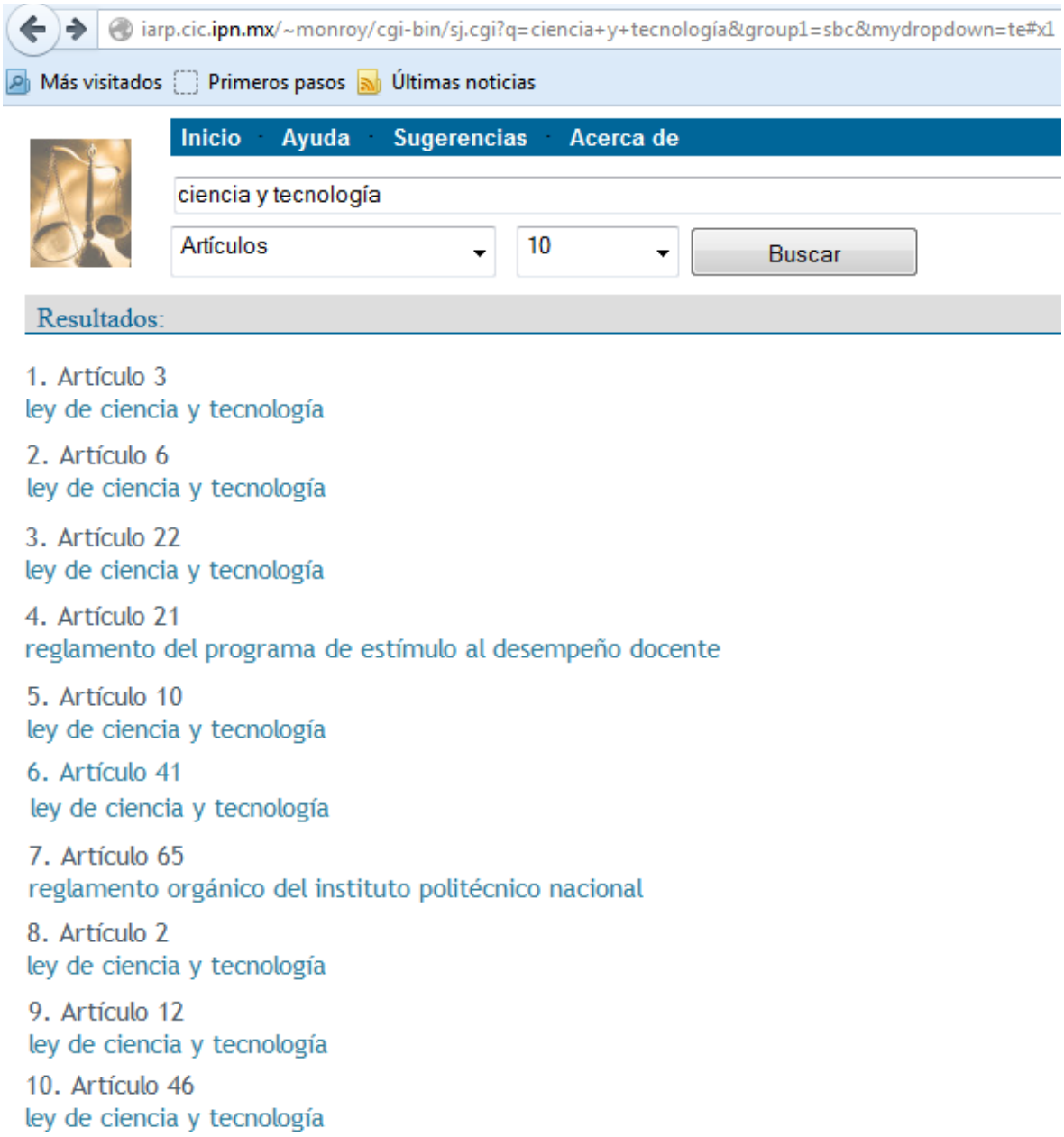


Figura 14. Resultados de la implementación del modelo propuesto en su versión del experimento III para la solicitud *ciencia y tecnología*.

En las figuras siguientes se proporcionan los resultados devueltos por INFOJUS para la solicitud *ciencia AND tecnologia*. Como se puede observar en la **Figura 15** el primer resultado que devuelve INFOJUS es el Artículo tercero de la Ley de Bioseguridad de

Organismos Genéticamente Modificados, parte del contenido del mismo se reproduce a continuación:

Para los efectos de esta ley, se entiende por:

I. Accidente: la liberación involuntaria de organismos genéticamente modificados durante su utilización y que pueda suponer, con base en criterios técnicos, posibles riesgos para la salud humana o para el medio ambiente y la diversidad biológica.

II. Actividades: la utilización confinada, la liberación experimental, la liberación en programa piloto, la liberación comercial, la comercialización, la importación y la exportación de organismos genéticamente modificados, conforme a esta ley.

III. Autorización: es el acto administrativo mediante el cual la secretaria de salud, en el ámbito de su competencia conforme a esta ley, autoriza organismos genéticamente modificados determinados expresamente en este ordenamiento, a efecto de que se pueda realizar su comercialización e importación para su comercialización, así como su utilización con finalidades de salud pública o de biorremediación.

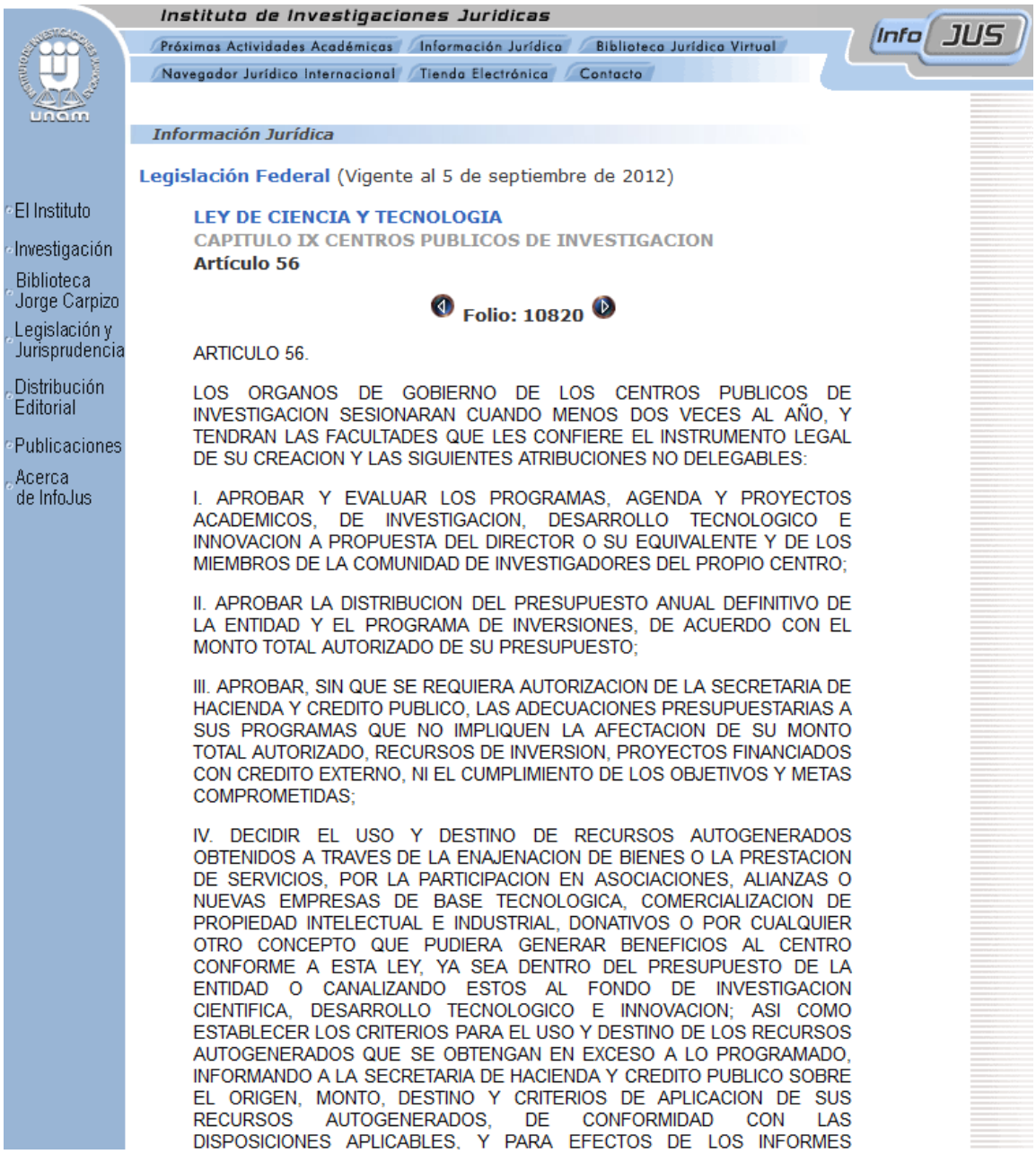
IV. Biorremediación: el proceso en el que se utilizan microrganismos genéticamente modificados para la degradación o desintegración de contaminantes que afecten recursos y/o elementos naturales, a efecto de convertirlos en componentes más sencillos y menos dañinos o no dañinos al ambiente.

V. Bioseguridad: las acciones y medidas de evaluación, monitoreo, control y prevención que se deben asumir en la realización de actividades con organismos genéticamente modificados, con el objeto de prevenir, evitar o reducir los posibles riesgos que dichas actividades pudieran ocasionar a la salud humana o al medio ambiente y la diversidad biológica, incluyendo los aspectos de inocuidad de dichos organismos que se destinen para uso o consumo humano…

XXXVI Zonas restringidas: los centros de origen, los centros de diversidad genética y las áreas naturales protegidas, dentro de los cuales se restrinja la realización de actividades con organismos genéticamente modificados, en los términos de esta ley.

Debido a su extensión no se incluye todo el contenido del artículo. Esto mismo es posible sea la causa que tal artículo haya sido seleccionado por INFOJUS como primera opción, dejando uno de los artículos (y el único de los 10 resultados) de la Ley de Ciencia y Tecnología en el cuarto lugar. En la **Figura 16** y **Figura 17** se encuentra el resultado que devuelve INFOJUS al elegir la opción *Ver documento* con las palabras resaltadas: en resumen y en texto completo, respectivamente, para el artículo 56 de la Ley de Ciencia y Tecnología.

Figura 15. Primeros 10 resultados devueltos por el sistema de consulta de información jurídica INFOJUS para la solicitud *ciencia AND tecnología.*



Instituto de Investigaciones Jurídicas

Próximas Actividades Académicas | Información Jurídica | Biblioteca Jurídica Virtual | Navegador Jurídico Internacional | Tienda Electrónica | Contacto

InfoJUS

Información Jurídica

El Instituto
Investigación
Biblioteca Jorge Carpizo
Legislación y Jurisprudencia
Distribución Editorial
Publicaciones
Acerca de InfoJus

Legislación Federal (Vigente al 5 de septiembre de 2012)

LEY DE CIENCIA Y TECNOLOGIA

CAPITULO IX CENTROS PUBLICOS DE INVESTIGACION

Artículo 56

Folio: 10820

ARTICULO 56.

LOS ORGANOS DE GOBIERNO DE LOS CENTROS PUBLICOS DE INVESTIGACION SESIONARAN CUANDO MENOS DOS VECES AL AÑO, Y TENDRAN LAS FACULTADES QUE LES CONFIERE EL INSTRUMENTO LEGAL DE SU CREACION Y LAS SIGUIENTES ATRIBUCIONES NO DELEGABLES:

I. APROBAR Y EVALUAR LOS PROGRAMAS, AGENDA Y PROYECTOS ACADEMICOS, DE INVESTIGACION, DESARROLLO TECNOLOGICO E INNOVACION A PROPUESTA DEL DIRECTOR O SU EQUIVALENTE Y DE LOS MIEMBROS DE LA COMUNIDAD DE INVESTIGADORES DEL PROPIO CENTRO;

II. APROBAR LA DISTRIBUCION DEL PRESUPUESTO ANUAL DEFINITIVO DE LA ENTIDAD Y EL PROGRAMA DE INVERSIONES, DE ACUERDO CON EL MONTO TOTAL AUTORIZADO DE SU PRESUPUESTO;

III. APROBAR, SIN QUE SE REQUIERA AUTORIZACION DE LA SECRETARIA DE HACIENDA Y CREDITO PUBLICO, LAS ADECUACIONES PRESUPUESTARIAS A SUS PROGRAMAS QUE NO IMPLIQUEN LA AFECTACION DE SU MONTO TOTAL AUTORIZADO, RECURSOS DE INVERSION, PROYECTOS FINANCIADOS CON CREDITO EXTERNO, NI EL CUMPLIMIENTO DE LOS OBJETIVOS Y METAS COMPROMETIDAS;

IV. DECIDIR EL USO Y DESTINO DE RECURSOS AUTOGENERADOS OBTENIDOS A TRAVES DE LA ENAJENACION DE BIENES O LA PRESTACION DE SERVICIOS, POR LA PARTICIPACION EN ASOCIACIONES, ALIANZAS O NUEVAS EMPRESAS DE BASE TECNOLOGICA, COMERCIALIZACION DE PROPIEDAD INTELECTUAL E INDUSTRIAL, DONATIVOS O POR CUALQUIER OTRO CONCEPTO QUE PUDIERA GENERAR BENEFICIOS AL CENTRO CONFORME A ESTA LEY, YA SEA DENTRO DEL PRESUPUESTO DE LA ENTIDAD O CANALIZANDO ESTOS AL FONDO DE INVESTIGACION CIENTIFICA, DESARROLLO TECNOLOGICO E INNOVACION; ASI COMO ESTABLECER LOS CRITERIOS PARA EL USO Y DESTINO DE LOS RECURSOS AUTOGENERADOS QUE SE OBTENGAN EN EXCESO A LO PROGRAMADO, INFORMANDO A LA SECRETARIA DE HACIENDA Y CREDITO PUBLICO SOBRE EL ORIGEN, MONTO, DESTINO Y CRITERIOS DE APLICACION DE SUS RECURSOS AUTOGENERADOS, DE CONFORMIDAD CON LAS DISPOSICIONES APLICABLES, Y PARA EFECTOS DE LOS INFORMES

Figura 16. Parte del artículo 56 de la Ley de Ciencia y Tecnología devuelto por INFOJUS para la solicitud *ciencia AND tecnologia.*

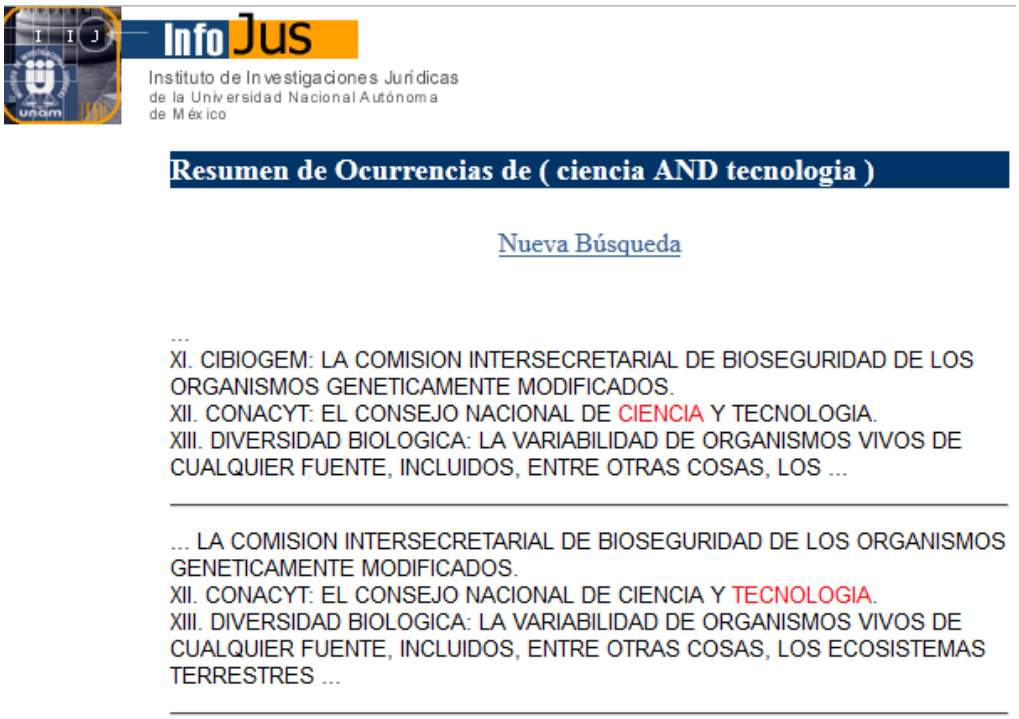


Figura 17. Resultado devuelto por INFOJUS con palabras resaltadas para la solicitud *ciencia AND tecnologia*.

A partir de la comparación trivial antes mostrada podría valer la pena el hacer una comparación formal entre ambos enfoques. Entre las contribuciones del presente trabajo se encuentra el mismo corpus empleado en los experimentos, el cual está a disposición en línea para su uso bajo las condiciones indicadas en el presente trabajo[14].

[14] http://iarp.cic.ipn.mx/~monroy/ versión más reciente [01-10-2012].

Conclusiones

En este trabajo hemos descrito un modelo para la recuperación de información para disminuir el problema de síntesis del conocimiento en el dominio legal, tomando en cuenta la organización tradicional de los textos legislativos-normativos y las referencias que los artículos de sus disposiciones generales contienen. El modelo combina métodos del área de Procesamiento de Lenguaje Natural, Teoría de Grafos y Recuperación de Información con una representación formal de los documentos legales para abordar el problema de síntesis del conocimiento.

Como se mostró, el incluir las referencias mejoró los resultados de los modelos tradicionales de recuperación de información como el Modelo Vectorial y Lucene, a pesar de también emplear la misma medida de similitud (al menos en parte, en el caso de Lucene). El hecho de que los grafos consideren no solamente la similitud entre la solicitud y los documentos, sino también entre estos últimos, mejoró los resultados con respecto al modelo vectorial. Si bien en el primer experimento las referencias incluidas representaron solamente el 11.7% del total de artículos, es apreciable el impacto de éstas en la recuperación. Las referencias intra e interdocumentos favorecieron que la respuesta no sólo fuera calificada como correcta, sino que los artículos respuesta estuvieran mejor posicionados. Para el caso de preguntas sencillas y específicas, la recuperación tanto por semejanza como por las referencias, fue favorecida igualmente.

Los experimentos diseñados para evaluar el desempeño del modelo propuesto, mostraron que las mejores calificaciones correspondieron a la combinación de las referencias y similitud entre artículos; tanto para situaciones en que la información requerida pertenecía diversos documentos como a uno solo, en contraste con los resultados arrojados por las implementaciones del modelo propuesto, cuando sólo se consideraron las referencias y la similitud por separado. De lo anterior, se observó que existe una relación entre la similitud y las referencias. Para cada pregunta en particular, el compromiso entre estas variables, dio como resultado la calificación y posición encontradas. Se destaca el hecho de que este tipo de combinación no había sido empleada para una tarea específica, como en el presente trabajo, lo cual implica una ventaja sobre los tradicionales sistemas de recuperación de información, tales como Lucene y JIRS

Consideramos que el presente trabajo abre una línea de investigación para explorar diferentes medidas de semejanza, el uso algoritmos de citas u otras formas de combinar resultados en recuperación de información. También es posible extrapolar el modelo a documentos con una estructura similar, por ejemplo casos legales.

Referencias

Aceves-Pérez, R. M. (2008). Búsqueda de Respuestas en Fuentes Documentales Multilingües. Tesis de Doctorado, INAOE, Puebla, México.

Alvite-Díez, M. L. (2003). Tendencias en la investigación sobre recuperación de información jurídica. Rev. Esp. Doc. Cient. 26 (2), 191-212.

Aslam, J. A. & Montague, M. (2001). Models for Metasearch. In Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information retrieval (SIGIR '01). ACM, New York, NY, USA, 276-284.

Baeza-Yates, R. (1999). Modern Information Retrieval. Addison Wesley Longman Publishing Co. Inc.

Barmakian, D. (2000). Better Search Engines for Law. Law Library Journal, 92(4), 399-438.

Bartell, B. T., Cottrell, G. W. & Belew, R. K. (1994). Automatic Combination of Multiple Ranked Retrieval Systems. W. Bruce Croft and C. J. van Rijsbergen, editors. SIGIR 94: Proceedings of the Seventeenth Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval, Dublin, Ireland, Springer-Verlag, 73-81.

Brin, S. & Page, L. (1998). The Anatomy of a Large-Scale Hypertextual Web Search Engine. Comput. Netw. ISDN Syst., 30(1-7), 107-117.

Brunzel, M. y M. Spiliopoulou. 2007. Domain Relevance on Term Weighting. Lecture Notes in Computer Science 4592, Springer, 2007, 427-432.

Buscaldi, D., Rosso, P., Gómez-Soriano, J. & Sanchis, E. (2010). Answering Questions with an n-gram based Passage Retrieval Engine. Journal of Intelligent Information Systems, 34(2), 113-134.

Callan, J. (1994). Passage-level Evidence in Document Retrieval. In Proceedings of the 17th ACM SIGIR Conference on Research and Development in Information Retrieval, 302-310.

Callan, J., Lu, Z., and Croft, W. (1995). Searching Distributed Collections with Inference Networks. In Proceedings of the 18th ACM SIGIR Conference on Research and Development in Information Retrieval, 21-28.

Caputo, A., Basile, P. & Semeraro, G. (2009). Boosting a Semantic Search Engine by Named Entities. In Proceedings of the 18th International Symposium on Foundations of Intelligent Systems (ISMIS '09), Jan R., Zbigniew W. R., Petr B. & Tapio E. (Eds.). Springer-Verlag, Berlin, Heidelberg, 241-250.

Chuang, H-Y., Liu, H., Chen, F-A., Kao, C-Y. & Hsu, D. F. (2004). Combination Methods in Microarray Analysis. In 7th International Symposium on Parallel Architectures, Algorithms, and Networks (I-SPAN 2004), IEEE Computer Society Press, 625-630.

Conitzer, V. (2006). Computational Aspects of Preference Aggregation. Tesis de Doctorado. Carnegie Mellon University. Pittsburgh, PA., USA.

Croft, W. & Harper, D. (1979). Using Probabilistic Models of Document Retrieval without Relevance Information. Journal of Documentation, 35:285-295.

Croft, W. and Thompson, R. (1987). I_R: A New Approach to the Design of Document Retrieval Systems. Journal of the American Society for Information Science, 38(6):389-404.

Daniels, J. J. & Rissland, E. L. (1997). Finding Legally Relevant Passages in Case Opinions. In Proceedings of the 6th International Conference on Artificial Intelligence and Law (ICAIL '97). ACM, New York, NY, USA, 39-46.

Davenport A. & Kalagnanam, J. (2004). A Computational Study of the Kemeny Rule for Preference Aggregation. In Proceedings of the 19th National Conference on Artificial Intelligence (AAAI'04), Anthony G. Cohn (Ed.). AAAI Press, 697-702.

Deerwester, S., Dumais, S., Furnas, G., Landauer, T., and Harshman, R. (1990). Indexing by Latent Semantic Analysis. Journal of the American Society for Information Science, 41:391-407.

Dwork, C., Kumar, R., Naor, M. & Sivakumar, D. (2001). Rank Aggregation Methods for the Web. In Proceedings of the 10th International Conference on World Wide Web (WWW '01). ACM, New York, NY, USA, 613-622

Erkan, G. & Radev, R. D. (2004). LexRank: Graph-based Lexical Centrality as Salience in Text Summarization. J. Artif. Int. Res. 22 (1), 457-479.

Fagin, R. (1996). Combining Fuzzy Information from Multiple Systems. In Proceedings of the 15th ACM Conference on Principles of Database Systems (PODS), 216-226.

Fagin, R. (1998). Fuzzy Queries in Multimedia Database Systems. In Proceedings of the 17th ACM Conference on Principles of Database Systems (PODS), 1-10.

Ferreira, A. & Atkinson, J. (2009). Disminución de la sobrecarga de información en la World Wide Web a partir de interacciones dialógicas hombre-computador. Rev. Signos. 42 (69), 9-27.

Fisher, H. & Elchesen, D. (1972). Effectiveness of Combining Title Words and Index Terms in Machine Retrieval Searches. Nature, 238:109-110.

Fitchett, T. (1997). The Road to the Virtual Library: The Center for Electronic Text in the Law Builds DIANA. Journal of Information, Law and Technology, 1997 (3).

Fox, E. and France, R. (1987). Architecture of an Expert System for Composite Document Analysis, Representation, and Retrieval. Journal of Approximate Reasoning, 1:151-175.

Francesconi, E. (2006). The "*Norme in Rete*" Project: Standards and Tools for Italian Legislation. International Journal of Legal Information, 34(2), 358-376.

Fuhr, N. (1992). Probabilistic Models in Information Retrieval. Computer Journal, 35(3):243-255.

Fuller, M., Kaszkiel, M., Ng, C.L., Vines, P., Wilkinson, R. & Zobel, J. (1998). MDS TREC6 report. Donna K. Harman, editor. The Sixth Text REtrieval Conference (TREC-6), Gaithersberg, MD., National Institute of Standards and Technology, 241-257. NIST Special Publication, 500-240.

Fürnkranz, J. & Hüllermeier E. Eds. (2011). Preference Learning. Springer Berlin Heidelberg.

Gargano, M. L. & Prasad, K. M. (2006). Rank Agregation for Metasearch Engines. Technical Report, PACE University, NY, USA.

Geist, A. (2009). Using Citation Analysis Techniques for Computer-Assisted Legal Research in Continental Jurisdictions. Tesis de maestría, Universidad de Edimburgo, Edimburgo, Reino Unido.

Golubitsky, M. & Dellnitz M. (1999). Linear Algebra and Differential Equations Using MATLAB. Brooks Cole Publishing Company.

Hiemstra, D. (2009). Information Retrieval Models. In Information Retrieval: Searching in the 21st Century (Eds. A. Göker and J. Davies), John Wiley & Sons, Ltd, Chichester, UK.

Huerta-Ochoa Carla. (2001). Artículos Transitorios y Derogación. Boletín Mexicano de Derecho Comparado, 34(102), 811-840.

Hull, D., Pedersen, J., and Schutze, H. (1996). Method Combination for Document Filtering. In Proceedings of the 19th ACM SIGIR Conference on Research and Development in Information Retrieval, 279-287.

Johnsonbaugh, R. (2005). Matemáticas Discretas. Sexta Edición. Pearson-Prentice-Hall.

Kantor, P. B. (1995). Decision Level Data Fusion for Routing of Documents in the TREC3 Context: A Best Case Analysis of Worst Case Results. D.K. Harman, editor. The Third Text REtrieval Conference (TREC-3), Gaithersberg, MD., National Institute of Standards and Technology. NIST Special Publication, 500-226.

Katzer, J., McGill, M., Tessier, J., Frakes, W. & DasGupta, P. (1982). A Study of the Overlap Among Document Representations. Information Technology: Research and Development, 1(4):261-274.

Kerrigan, S. & Law, K. H. (2003). Logic-based Regulation Compliance-Assistance. In Proceedings of the 9th International Conference on Artificial Intelligence and Law (ICAIL '03). ACM, New York, NY, USA, 126-135.

Kleinberg, J. M. (1999). Authoritative Sources in a Hyperlinked Environment. J. ACM, 46(5), 604-632.

Knaus, D., Mittendorf, E. & Schauble, P. (1995). Improving a Basic Retrieval Method by Links and Passage Level Evidence. D.K. Harman, editor. The Third Text REtrieval Conference (TREC-3), Gaithersberg, MD., National Institute of Standards and Technology. NIST Special Publication, 500-226.

Lapata, M. (2006). Automatic Evaluation of Information Ordering: Kendall's Tau. Comput. Linguist. 32, (4), 471-484.

Larkey, L. and Croft, W. (1996). Combining classifiers in text categorization. In Proceedings of the 19th ACM SIGIR Conference on Research and Development in Information Retrieval, 289-297.

Lau, G. T., Law, K. H. & Wiederhold, G. (2006). A Relatedness Analysis of Government Regulations Using Domain Knowledge and Structural Organization. Information Retrieval, 9(6), 657-680.

Ledeneva Y. Automatic Language-Independent Detection of Multiword Descriptions for Text Summarization. Tesis de Doctorado. CIC-IPN, D.F., México.

Lee, J. (1995). Combining Multiple Evidence from Different Properties of Weighting Schemes. In Proceedings of the 18th ACMSIGIR Conference on Research and Development in Information Retrieval, 180-188.

Lee, J. (1997). Analyses of Multiple Evidence Combination. In Proceedings of the 20th ACM SIGIR Conference on Research and Development in Information Retrieval, 267-276.

Lewis, D. and Hayes, P. (1994). Special Issue on Text Categorization. ACM Transactions on Information Systems, 12(3).

Lin, S. (2010). Rank Aggregation Methods. WIREs Comp Stat, 2(5), 555-570.

Luhn, H. P. (1957). A Statistical Approach to Mechanized Encoding and Searching of Literary Information. IBM J. Res. Dev., 1 (4), 309-317.

Manning, D. C., Raghavan, P. & Schtze, H. (2008). Introduction to Information Retrieval, Cambridge University Press, New York, NY.

McGill, M., Koll, M., & Noreault, T. (1979). An Evaluation of Factors Affecting Document Ranking by Information Retrieval Systems. Final report for grant NSF-IST-78-10454 to the National Science Foundation, Syracuse University.

Mercatali, P., Romano, F., Boschi, L. & Spinicci, E. (2005). Automatic Translation from Textual Representations of Laws to Formal Models through UML. In Proceedings of the 2005 conference on Legal Knowledge and Information Systems: JURIX 2005: The Eighteenth Annual Conference, M. F. Moens & P. Spyns (Eds.). IOS Press, Amsterdam, The Netherlands, The Netherlands, 71-80.

Mihalcea, R. (2004). Graph-based Ranking Algorithms for Sentence Extraction, Applied to Text Summarization. In Proceedings of the ACL 2004 on Interactive poster and demonstration sessions (ACLdemo '04). Association for Computational Linguistics, Stroudsburg, PA, USA, Article 20.

Mihalcea, R., Tarau, P. & Figa, E. (2004). PageRank on Semantic Networks, with Application to Word Sense Disambiguation. In Proceedings of the 20th international conference on Computational Linguistics (COLING '04). Association for Computational Linguistics, Stroudsburg, PA, USA, Article 1126.

Mitchell, T. (1997). Machine Learning. McGraw-Hill, New York.

Moens M.F. (2001). Innovative Techniques for Legal Text Retrieval. Artificial Intelligence and Law, 9(1), 29-57.

Moens, M. F. (2006). Improving Access to Legal Information: How Drafting Systems Help. En A. Lodder & A. Oskamp (Eds.), Information Technology and Lawyers (119-136). Springer Netherlands.

Moens, M. F., Uyttendaele, C. & Dumortier, J. (1997). Abstracting of Legal Cases: the SALOMON Experience. In Proceedings of the 6th international conference on Artificial intelligence and law (ICAIL '97). ACM, New York, NY, USA, 114-122.

Monroy, A., Calvo, H. & Gelbukh, A. (2008). Using Graphs for Shallow Question Answering on Legal Documents. En A. Gelbukh & E. Morales (Eds.), MICAI 2008: Advances in Artificial Intelligence. Lecture Notes in Computer Science (165-173). Springer Berlin/Heidelberg.

Monroy, A., Calvo, H. & Gelbukh, A. (2009). NLP for Shallow Question Answering of Legal Documents Using Graphs. En A. Gelbukh (Ed.), Computational Linguistics and Intelligent Text Processing. Lecture Notes in Computer Science (498-508). Springer Berlin/Heidelberg.

O'Connor, J. (1975). Retrieval of Answer-Sentences and Answer Figures from Papers by Text Searching. Information Processing and Management, 11(5/7):155-164.

O'Connor, J. (1980). Answer-Passage Retrieval by Text Searching. Journal of the American Society for Information Science, 31(4):227-239.

Osborn, J. & Sterling, L. (1999). JUSTICE: a Judicial Search Tool Using Intelligent Concept Extraction. In Proceedings of the 7th International Conference on Artificial Intelligence and Law (ICAIL '99). ACM, New York, NY, USA, 173-181.

Pearl, J. (1988). Probabilistic reasoning in intelligent systems: Networks of Plausible Inference. Morgan Kaufmann, San Mateo.

Peñas, A., Forner, P., Sutcliffe, R., Rodrigo, A., Forăscu, C., Alegria, I., Giampiccolo, D., Moreau, N. & Osenova, P. (2009). Overview of ResPubliQA 2009: Question Answering Evaluation Over European Legislation. In Proceedings of the 10th Cross-Language Evaluation Forum Conference on Multilingual Information Access Evaluation: Text Retrieval Experiments (CLEF'09), C. Peters, G. M. Di Nunzio, M. Kurimo, T. Mandl & D. Mostefa (Eds.). Springer-Verlag, Berlin, Heidelberg, 174-196.

Porter, M. F. (1997). An Algorithm for Suffix Stripping. In Readings in Information Retrieval, Karen Sparck Jones & Peter Willett (Eds.). Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 313-316.

Post, D. G. & Eisen, M. B. (2000). How Long is the Coastline of the Law? Thoughts on the fractal nature of legal systems. The Journal of Legal Studies, 29(1), 545-584.

Ríos-Estavillo, J. J. (1997). Derecho e informática en México. México: Instituto de Investigaciones Jurídicas (Universidad Nacional Autónoma de México).

Roberts, F. S. (2008). Computer Science and Decision Theory. Annals of Operations Research, 163 (1), 209-253.

Robertson, S. (1977). The Probability Ranking Principle in Information Retrieval. Journal of Documentation, 33:294-304.

Rosso, P., Correa, S. & Buscaldi, D. (2011). Passage Retrieval in Legal Texts. The Journal of Logic and Algebraic Programming, 80(3-5), 139-153.

Salton, G. & Buckley, C. (1988). Term-weighting Approaches in Automatic Text Retrieval. Information Processing and Management, 24:513-523.

Salton, G., Allan, J., & Buckley, C. (1993). Approaches to Passage Retrieval in Full Text Information Systems. In Proceedings of the 17th ACM SIGIR Conference on Research and Development in Information Retrieval, 49-56.

Salton, G., Wong, A., & Yang, C. (1975). A Vector Space Model for Automatic Indexing. Communications of the ACM, 18:613-620.

Schneider, R. (2002). n-grams of Seeds: A Hybrid System for Corpus-based Text Summarization. In proceedings of LREC, Third International Conference on Language Resources and Evaluation, Las Palmas de Gran Canaria, España, 29-31.

Schweighofer, E., Rauber, A. & Dittenbach, M. (2001). Automatic Text Representation, Classification and Labeling in European Law. In Proceedings of the 8th International Conference on Artificial Intelligence and Law (ICAIL '01), R. P. Loui (Ed.). ACM, New York, NY, USA, 78-87.

Selberg, E. & Etzioni, O. (1996). Multi-service Search and Comparison Using the MetaCrawler. In Proceedings of the 4th International World Wide Web Conference, 195-208.

Shaw, J.A. & Fox, E.A. (1995). Combination of Multiple Searches. D.K. Harman, editor. The Third Text REtrieval Conference (TREC-3), Gaithersberg, MD., National Institute of Standards and Technology. NIST Special Publication, 500-226.

Sjöberg, C. M. (1998). Critical Factors in Legal Document Management: A Study of Standardised Markup Languages. Stockholm: Jure AB.

Strzalkowski, T., Lin, F. & Pérez-Carballo, J. (1998). Natural Language Information Retrieval TREC6 Report. Donna K. Harman, editor. The Sixth Text REtrieval Conference (TREC-6), Gaithersberg, MD., National Institute of Standards and Technology. 347-366. NIST Special Publication, 500-240.

Truchon, M. & Gordon, S. (2009). Statistical Comparison of Aggregation Rules for Votes. Mathematical Social Sciences, 57 (2), 199-212.

Tumer, K. and Ghosh, J. (1999). Linear and Order Statistics Combiners for Pattern Classification. In Sharkey, A., editor, Combining Artificial Neural Networks, 127-162. Springer-Verlag.

Turtle, H. (1995). Text Retrieval in the Legal World. Artificial Intelligence and Law, 3(1), 5-54.

Turtle, H. and Croft, W. (1992). A Comparison of Text Retrieval Models. Computer Journal, 35(3):279-290.

Ukkonen, A. (2004). Data Mining Techniques for Discovering Partial Orders. Tesis de Maestría. Helsinki University of Technology. Helsinki, Finlandia.

van Engers, T., van Gog, T. & Jacobs, A. (2005). How Technology Can Help Reducing the Legal Burden. In Proceedings of the 2005 conference on Legal Knowledge and Information Systems: JURIX 2005: The Eighteenth Annual Conference, M. F. Moens & P. Spyns (Eds.). IOS Press, Amsterdam, The Netherlands, The Netherlands, 101-102.

Van Rijsbergen, C. (1986). A non-classical Logic for Information Retrieval. Computer Journal, 29:481-485.

Vogt, C. and Cottrell, G. (1998). Predicting the Performance of Linearly Combined IR Systems. In Proceedings of the 21st ACM SIGIR Conference on Research and Development in Information Retrieval, 190-196.

Vogt, C. C., Cottrell, G. W., Belew, R.K. & Bartell, B. T. (1997). Using Relevance to Train a Linear Mixture of Experts. In Harman Donna K. Harman, editor. The Fifth Text REtrieval Conference (TREC-5), Gaithersberg, MD., National Institute of Standards and Technology, 503-515. NIST Special Publication, 500-238.

Voorhees, E., Gupta, N., and Johnson-Laird, B. (1995). Learning Collection Fusion Strategies. In Proceedings of the 18th ACM SIGIR Conference on Research and Development in Information Retrieval, 172-179.

Wilkinson, R. (1994). Effective Retrieval of Structured Documents. In Proceedings of the 17th ACM SIGIR Conference on Research and Development in Information Retrieval, 311-317.

Yang, J-M., Chen, Y-F., Shen, T-W., Kristal, B. S. & Hsu, D. F. (2005). Consensus Scoring Criteria for Improving Enrichment in Virtual Screening. Journal of Chemical Information and Modeling, 45(4), 1134-1146.

Anexo A

#. Número de documento

Arts. Número de artículos que contiene el documento.

Documento. Título del documento

#	Arts.	Documento
1	34	Ley orgánica del Instituto Politécnico Nacional
2	34	Reglamento de academias del Instituto Politécnico Nacional
3	72	Reglamento del archivo histórico del Instituto Politécnico Nacional
4	34	Reglamento para la aprobación de planes y programas de estudio en el Instituto Politécnico Nacional
5	46	Reglamento de titulación profesional del Instituto Politécnico Nacional
6	39	Reglamento de becas de estudio apoyos económicos y licencias con goce de sueldo del personal académico del Instituto Politécnico Nacional
7	33	Reglamento de becas estímulos y otros medios de apoyo para alumnos del Instituto Politécnico Nacional
8	54	Reglamento del consejo general consultivo del Instituto Politécnico Nacional
9	153	Reglamento de las condiciones interiores de trabajo del personal académico del Instituto Politécnico Nacional
10	144	Reglamento de las condiciones generales de trabajo del personal no docente del Instituto Politécnico Nacional
11	28	Reglamento interior de la comisión de operación y fomento de actividades académicas del Instituto Politécnico Nacional
12	26	Reglamento del decanato del Instituto Politécnico Nacional
13	42	Reglamento de diplomados del Instituto Politécnico Nacional
14	39	Reglamento de distinciones al mérito politécnico del Instituto Politécnico Nacional
15	28	Reglamento de evaluación del Instituto Politécnico Nacional
16	78	Reglamento de estudios escolarizados para los niveles medio superior y superior del Instituto Politécnico Nacional
17	295	Reglamento interno del Instituto Politécnico Nacional
18	87	Reglamento de incorporación reconocimiento de validez oficial equivalencia y revalidación de estudios del Instituto Politécnico Nacional
19	77	Reglamento de integración social del Instituto Politécnico Nacional
20	87	Reglamento orgánico del Instituto Politécnico Nacional
21	61	Reglamento para la operación administración y uso de la red institucional de cómputo y telecomunicaciones del Instituto Politécnico Nacional
22	23	Reglamento de planeación del Instituto Politécnico Nacional
23	204	Reglamento de promoción docente del Instituto Politécnico Nacional
24	46	Reglamento del programa de estímulo al desempeño docente
25	107	Reglamento de estudios de posgrado del Instituto Politécnico Nacional

26	34	Reglamento de prácticas y visitas escolares del Instituto Politécnico Nacional
27	57	Reglamento del Sistema de Becas por Exclusividad (SIBE)
28	25	Reglamento de servicio social del Instituto Politécnico Nacional
29	31	Reglamento de becas del consejo nacional de ciencia y tecnología
30	136	Constitución Política de los Estados Unidos Mexicanos
31	73	Ley reglamentaria del artículo 5 constitucional relativo al ejercicio de las profesiones en el Distrito Federal
32	102	Reglamento de la ley reglamentaria del artículo 5 constitucional relativo al ejercicio de las profesiones en el Distrito Federal
33	92	Reglamento de las condiciones generales de trabajo del personal de la Secretaría de Educación Pública
34	175	Ley federal de los trabajadores al servicio del estado reglamentaria del apartado b del artículo 123 constitucional
35	254	Ley del Instituto de Seguridad y Servicios Sociales de los Trabajadores del Estado
36	1073	Ley federal del trabajo
37	94	Ley federal de responsabilidades de los servidores públicos
38	3154	Código civil para el distrito federal
39	514	Código penal federal
40	630	Código federal de procedimientos penales
41	581	Código federal de procedimientos civiles
42	91	Reglamento del sistema nacional de investigadores
43	110	Ley de fiscalización y rendición de cuentas de la federación
44	27	Ley de fomento para la lectura y el libro
45	96	Ley de fondos de aseguramiento agropecuario y rural
46	191	Ley de desarrollo rural sustentable
47	328	Ley de instituciones de crédito
48	42	Ley de inversión extranjera
49	17	Ley de la casa de moneda de México
50	21	Ley de la comisión nacional bancaria y de valores
51	15	Ley de la comisión nacional de hidrocarburos
52	76	Ley de la comisión nacional de los derechos humanos
53	19	Ley de la comisión nacional para el desarrollo de los pueblos indígenas
54	13	Ley de la comisión reguladora de energía
55	55	Ley de la policía federal
56	279	Ley de la propiedad industrial
57	50	Ley de los derechos de las personas adultas mayores
58	342	Ley de concursos mercantiles
59	58	Ley de los institutos nacionales de salud
60	146	Ley de los sistemas de ahorro para el retiro
61	37	Ley de nacionalidad

62	328	Ley de navegación y comercio marítimos
63	110	Ley de obras públicas y servicios relacionados con las mismas
64	26	Ley de organizaciones ganaderas
65	73	Ley de Petróleos Mexicanos
66	45	Ley de planeación
67	135	Ley de premios estímulos y recompensas civiles
68	31	Ley de promoción y desarrollo de los bioenergéticos
69	93	Ley de protección al ahorro bancario
70	9	Ley de protección al comercio y la inversión de normas extranjeras que contravengan el derecho internacional
71	118	Ley de protección y defensa al usuario de servicios financieros
72	83	Ley de caminos puentes y autotransporte federal
73	69	Ley de puertos
74	69	Ley de recompensas de la armada de México
75	31	Ley de responsabilidad civil por daños nucleares
76	67	Ley de seguridad nacional
77	44	Ley de sistemas de pagos
78	97	Ley de sociedades de inversión
79	109	Ley de comercio exterior
80	43	Ley de sociedades de solidaridad social
81	22	Ley de transparencia y de fomento a la competencia en el crédito garantizado
82	145	Ley de uniones de crédito
83	593	Ley de vías generales de comunicación
84	97	Ley de vivienda
85	68	Ley del banco de México
86	5	Ley del consejo supremo de la defensa nacional
87	18	Ley del diario oficial de la federación y gacetas gubernamentales
88	13	Ley del impuesto a los depósitos en efectivo
89	61	Ley del Impuesto al Valor Agregado
90	19	Ley del Impuesto Empresarial a Tasa Única
91	64	Ley del impuesto especial sobre producción y servicios
92	261	Ley del impuesto sobre la renta
93	64	Ley de coordinación fiscal
94	229	Ley del Instituto de Seguridad Social para las Fuerzas Armadas Mexicanas
95	3101	Código civil federal
96	84	Ley del Instituto del Fondo Nacional de la Vivienda para los Trabajadores
97	33	Ley del Instituto del Fondo Nacional para el Consumo de los Trabajadores
98	17	Ley del Instituto Mexicano de la Juventud
99	35	Ley del Instituto Nacional de las Mujeres
100	423	Ley del Mercado de Valores

101	27	Ley del Registro Público Vehicular
102	371	Ley del Seguro Social
103	41	Ley del servicio de administración tributaria
104	98	Ley del servicio de inspección fiscal
105	141	Ley del servicio de tesorería de la federación
106	73	Ley del servicio exterior mexicano
107	65	Ley del servicio militar
108	68	Ley del servicio postal mexicano
109	80	Ley del servicio profesional de carrera en la administración pública federal
110	47	Ley del servicio público de energía eléctrica
111	7	Ley del sistema de horario en los Estados Unidos Mexicanos
112	126	Ley del Sistema Nacional de Información Estadística y Geográfica
113	21	Ley en favor de los veteranos de la Revolución como servidores del estado
114	46	Ley federal contra la delincuencia organizada
115	98	Ley federal de armas de fuego y explosivos
116	58	Ley federal de cinematografía
117	50	Ley federal de competencia económica
118	23	Ley federal de correduría pública
119	39	Ley federal de defensoría pública
120	714	Ley federal de derechos
121	70	Ley federal de extinción de dominio reglamentaria del artículo 22 de la constitución política de los Estados Unidos Mexicanos
122	32	Ley federal de fomento a las actividades realizadas por organizaciones de la sociedad civil
123	191	Ley federal de instituciones de fianzas
124	17	Ley federal de juegos y sorteos
125	68	Ley federal de las entidades paraestatales
126	24	Ley federal de los derechos del contribuyente
127	1560	Código de comercio
128	137	Ley federal de metrología y normalización
129	118	Ley federal de presupuesto y responsabilidad hacendaria
130	118	Ley federal de procedimiento administrativo
131	100	Ley federal de procedimiento contencioso administrativo
132	42	Ley federal de producción certificación y comercio de semillas
133	177	Ley federal de protección al consumidor
134	69	Ley federal de protección de datos personales en posesión de los particulares
135	126	Ley federal de radio y televisión
136	35	Ley federal de responsabilidad patrimonial del estado
137	927	Código de justicia militar
138	52	Ley federal de responsabilidades administrativas de los servidores públicos

139	175	Ley federal de sanidad animal
140	96	Ley federal de sanidad vegetal
141	44	Ley federal de seguridad privada
142	79	Ley federal de telecomunicaciones
143	64	Ley federal de transparencia y acceso a la información pública gubernamental
144	48	Ley federal de variedades vegetales
145	244	Ley federal del derecho de autor
146	14	Ley federal del impuesto sobre automóviles nuevos
147	65	Ley federal del mar
148	394	Código federal de instituciones y procedimientos electorales
149	23	Ley federal para el control de precursores químicos productos químicos esenciales y máquinas para elaborar cápsulas tabletas y o comprimidos
150	75	Ley federal para el control de sustancias químicas susceptibles de desvío para la fabricación de armas químicas
151	45	Ley federal para el fomento de la microindustria y la actividad artesanal
152	59	Ley federal para la administración de bienes asegurados decomisados y abandonados
153	97	Ley federal para la administración y enajenación de bienes del sector público
154	85	Ley federal para prevenir y eliminar la discriminación
155	12	Ley federal para prevenir y sancionar la tortura
156	57	Ley federal sobre monumentos y zonas arqueológicos artísticos e históricos
157	59	Ley general de acceso de las mujeres a una vida libre de violencia
158	60	Ley general de asentamientos humanos
159	16	Ley general de bibliotecas
160	152	Ley general de bienes nacionales
161	57	Ley general de contabilidad gubernamental
162	140	Ley general de cultura física y deporte
163	25	Ley general de derechos lingüísticos de los pueblos indígenas
164	172	Ley general de desarrollo forestal sustentable
165	85	Ley general de desarrollo social
166	32	Ley general de deuda pública
167	85	Ley general de educación
168	203	Ley general de instituciones y sociedades mutualistas de seguros
169	33	Ley general de la infraestructura física educativa
170	36	Ley general de las personas con discapacidad
171	198	Ley general de organizaciones y actividades auxiliares de crédito
172	150	Ley general de pesca y acuacultura sustentables
173	158	Ley general de población
174	40	Ley general de protección civil

175	596	Ley general de salud
176	122	Ley general de sociedades cooperativas
177	266	Ley general de sociedades mercantiles
178	460	Ley general de títulos y operaciones de crédito
179	73	Ley general de turismo
180	135	Ley general de vida silvestre
181	262	Ley general del equilibrio ecológico y la protección al ambiente
182	111	Ley general del Sistema de Medios de Impugnación en Materia Electoral
183	152	Ley general del Sistema Nacional de Seguridad Pública
184	55	Ley general para el control del tabaco
185	49	Ley general para la igualdad entre mujeres y hombres
186	125	Ley general para la prevención y gestión integral de los residuos
187	26	Ley orgánica de la Universidad Autónoma Agraria Antonio Narro
188	63	Ley minera
189	24	Ley monetaria de los Estados Unidos Mexicanos
190	61	Ley orgánica de la Administración Pública Federal
191	94	Ley orgánica de la Armada de México
192	61	Ley orgánica de la Financiera Rural
193	14	Ley orgánica de la Lotería Nacional para la Asistencia Pública
194	28	Ley orgánica de la Procuraduría de la Defensa del Contribuyente
195	86	Ley orgánica de la Procuraduría General de la República
196	37	Ley orgánica de la Universidad Autónoma Metropolitana
197	18	Ley orgánica de la Universidad Nacional Autónoma de México
198	30	Ley orgánica de los Tribunales Agrarios
199	45	Ley orgánica de los Tribunales Militares
200	37	Ley orgánica de Nacional Financiera
201	37	Ley orgánica de Sociedad Hipotecaria Federal
202	39	Ley orgánica del Banco del Ahorro Nacional y Servicios Financieros
203	36	Ley orgánica del Banco Nacional de Comercio Exterior
204	35	Ley orgánica del Banco Nacional de Obras y Servicios Públicos
205	58	Ley orgánica del Banco Nacional del Ejército Fuerza Aérea y Armada
206	136	Ley orgánica del Congreso General de los Estados Unidos Mexicanos
207	68	Ley de ciencia y tecnología
208	211	Ley orgánica del Ejército y Fuerza Aérea Mexicanos
209	20	Ley orgánica del Instituto Nacional de Antropología e Historia
210	124	Ley de bioseguridad de organismos genéticamente modificados
211	255	Ley orgánica del Poder Judicial de la Federación
212	16	Ley orgánica del seminario de cultura mexicana
213	55	Ley orgánica del tribunal federal de justicia fiscal y administrativa
214	22	Ley de capitalización del procampo

215	31	Ley para el aprovechamiento de energías renovables y el financiamiento de la transición energética
216	33	Ley para el aprovechamiento sustentable de la energía
217	26	Ley para el desarrollo de la competitividad de la micro pequeña y mediana empresa
218	13	Ley para el diálogo la conciliación y la paz digna en Chiapas
219	128	Ley para el tratamiento de menores infractores para el Distrito Federal en materia común y para toda la República en materia federal
220	20	Ley para el uso y protección de la denominación y del emblema de la Cruz Roja
221	25	Ley para la comprobación ajuste y cómputo de servicios de la Armada de México
222	37	Ley para la comprobación ajuste y cómputo de servicios en el Ejército y Fuerza Aérea Mexicanos
223	27	Ley para la coordinación de la educación superior
224	7	Ley para la depuración de depósitos en efectivo y valores
225	9	Ley para la depuración y liquidación de cuentas de la Hacienda Pública Federal
226	56	Ley para la protección de los derechos de niñas niños y adolescentes
227	12	Ley para la reforma del estado
228	76	Ley para la transparencia y ordenamiento de los servicios financieros
229	20	Ley para prevenir y sancionar la trata de personas
230	132	Ley para regular las actividades de las sociedades cooperativas de ahorro y préstamo
231	67	Ley para regular las agrupaciones financieras
232	76	Ley para regular las sociedades de información crediticia
233	10	Ley que aprueba la adhesión de México al convenio constitutivo del banco de desarrollo del caribe y su ejecución
234	150	Ley de desarrollo sustentable de la caña de azúcar
235	9	Ley que crea el fideicomiso que administrará el fondo de apoyo social para ex-trabajadores migratorios mexicanos
236	15	Ley que crea el fideicomiso que administrará el fondo para el fortalecimiento de sociedades y cooperativas de ahorro y préstamo y de apoyo a sus ahorradores
237	8	Ley que crea el fondo de fomento a la industria y de garantía de valores mobiliarios
238	14	Ley que crea el fondo de garantía y fomento a la agricultura ganadería y avicultura
239	16	Ley que crea el Instituto Nacional de Bellas Artes y Literatura
240	41	Ley que crea la Agencia de Noticias del Estado Mexicano
241	17	Ley que crea la Agencia Espacial Mexicana
242	10	Ley que crea la Comisión Federal de Electricidad

243	16	Ley que crea la Universidad Autónoma de Chapingo
244	7	Ley que crea la Universidad del Ejército y Fuerza Aérea
245	16	Ley que declara reservas mineras nacionales los yacimientos de uranio torio y las demás substancias de las cuales se obtengan isótopos hendibles que puedan producir energía nuclear
246	12	Ley que establece bases para la ejecución en México por el Poder Ejecutivo Federal del convenio constitutivo del Banco Interamericano de Desarrollo
247	11	Ley que establece bases para la ejecución en México por el Poder Ejecutivo Federal del convenio constitutivo de la Asociación Internacional de Fomento
248	16	Ley que establece el Instituto Federal de Capacitación del Magisterio
249	19	Ley que establece las normas mínimas sobre readaptación social de sentenciados
250	15	Ley reglamentaria de la fracción V del artículo 76 de la Constitución General de la República
251	25	Ley reglamentaria de la fracción VI del artículo 76 de la Constitución Política de los Estados Unidos Mexicanos
252	24	Ley reglamentaria de la fracción XIII Bis del apartado B del artículo 123 de la Constitución Política de los Estados Unidos Mexicanos
253	2	Ley reglamentaria de la fracción XVIII del artículo 73 constitucional en lo que se refiere a la facultad del Congreso para dictar reglas para determinar el valor relativo de la moneda extranjera
254	73	Ley reglamentaria de las fracciones I y II del artículo 105 de la Constitución Política de los Estados Unidos Mexicanos
255	21	Ley reglamentaria del artículo 27 constitucional en el ramo del petróleo
256	52	Ley reglamentaria del artículo 27 constitucional en materia nuclear
257	14	Ley de energía para el campo
258	62	Ley reglamentaria del servicio ferroviario
259	36	Ley sobre delitos de imprenta
260	94	Ley de aviación civil
261	62	Ley sobre el escudo la bandera y el himno nacionales
262	9	Ley sobre elaboración y venta de café tostado
263	14	Ley sobre la aprobación de tratados internacionales en materia económica
264	11	Ley sobre la celebración de tratados
265	5	Ley tendiente a procurar el rápido despacho de asuntos pendientes en el tribunal fiscal de la federación
266	1868	Ordenanza general de la Armada
267	15	Ley de contribución de mejoras por obras públicas federales de infraestructura hidráulica
268	22	Ley de ayuda alimentaria para los trabajadores
269	37	Ley de extradición internacional
270	80	Ley de disciplina para el personal de la Armada de México
271	360	Código fiscal de la federación

272	53	Ley de disciplina del Ejército y Fuerza Aérea Mexicanos
273	15	Estatuto de las Islas Marías
274	238	Ley aduanera
275	200	Ley agraria
276	96	Ley de adquisiciones arrendamientos y servicios del sector público
277	85	Ley de aeropuertos
278	190	Ley de aguas nacionales
279	218	Ley de ahorro y crédito popular
280	45	Ley de cámaras empresariales y sus confederaciones
281	7	Ley de Amnistía
282	4	Ley de Amnistía situación en Chiapas
283	247	Ley de amparo reglamentaria de los artículos 103 y 107 de la Constitución Política de los Estados Unidos Mexicanos
284	56	Ley de ascensos de la Armada de México
285	76	Ley de ascensos y recompensas del Ejército y Fuerza Aérea Mexicanos
286	68	Ley de asistencia social
287	19	Ley de asociaciones agrícolas
288	37	Ley de asociaciones religiosas y culto público

Anexo B

Conjunto de preguntas y sus correspondientes artículos-respuesta.

P1 ¿Cómo se lleva a cabo el procedimiento de elección de representantes alumnos ante el Consejo Técnico Consultivo Escolar?

Artículo 28 de la Ley Orgánica y 206, 207, 209 y 213 del Reglamento Interno del Instituto Politécnico Nacional.

P2 ¿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 del Instituto Politécnico Nacional.

P3 ¿Es procedente que un alumno solicite mención honorífica a nivel licenciatura si escogió su titulación por la opción de escolaridad?

Artículo 13 y 14 del Reglamento de titulación profesional del IPN.

P4 ¿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 título 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?

Artículo 24 y 33 del Reglamento de Titulación Profesional del IPN.

P5 ¿La esposa del actual director de una escuela unidad o centro de investigación puede ser candidata a dicho cargo?

Artículo 168 del Reglamento Interno del IPN, y artículos 14 y 21 de la Ley Orgánica del Instituto Politécnico Nacional.

P6 ¿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 4 de la Ley Orgánica del Instituto Politécnico Nacional

P7 ¿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?
Artículos 14 y 29 de la Ley Orgánica del Instituto Politécnico Nacional, y 264 del Reglamento Interno.

P8 ¿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 del Instituto Politécnico Nacional, y 202 y 264 del Reglamento Interno.
264 Reglamento Interno del Instituto Politécnico Nacional

P9 ¿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 del Instituto Politécnico Nacional.

P10 ¿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.
Artículo 21 de la Ley Orgánica, y 182 del Reglamento Interno del Instituto Politécnico Nacional.

P11 ¿Puede un mismo individuo por segundo año consecutivo ser electo consejero representante del personal académico?
Artículo 215 del Reglamento Interno del IPN.

P12 ¿Pueden prestar el servicio social alumnos que hayan acreditado el 70% de créditos pero adeuden materias?
Artículos 5 y 11 del Reglamento de Servicio Social del IPN.

P13 ¿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.

P14 ¿Qué debe entenderse por tener estudios de posgrado?
Artículos 3 y 4 del Reglamento de Estudios de Posgrado del IPN.

P15 ¿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 174 del Reglamento Interno del IPN

P16 ¿Tienen derecho de votar o no los profesores interinos adscritos a alguna escuela?
Artículo 208 del Reglamento Interno y 6 del Reglamento de las Condiciones Interiores de Trabajo del Personal Académico del Instituto Politécnico Nacional

P17 ¿Un alumno que adeuda cinco materias puede seguir fungiendo como consejero representante de los alumnos?
Artículo 81 del Reglamento Interno y 44 del Reglamento de Estudios Escolarizados para los Niveles Medio Superior y Superior del Instituto Politécnico Nacional.

P18 ¿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 del Instituto Politécnico Nacional.

P19 ¿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?
Artículos 20 y 23 del Reglamento del Decanato del Instituto Politécnico Nacional.

P20 ¿Una persona puede ser director de una escuela centro o unidad de investigación por tercera ocasión?
Artículo 21 de la Ley Orgánica del Instituto Politécnico Nacional.

P21 ¿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 del Instituto Politécnico Nacional

P22 ¿Es procedente cambiar a un trabajador académico de horario sin su consentimiento?
Artículo 63 del Reglamento de las Condiciones Interiores de Trabajo del Personal Académico del Instituto Politécnico Nacional

P23 ¿Qué debe entenderse por Voto de calidad del Maestro Decano en el proceso de elección de terna para la designación de Director?
Artículo 29 de la Ley Orgánica, y 13, 179, 153 y 193 del Reglamento Interno del Instituto Politécnico Nacional.

P24 ¿Es procedente que un trabajador solicite que la media hora, que tiene asignada para tomar alimentos y/o descanso a mitad de la jornada, se le otorgue a la entrada o a la salida del trabajo?

Artículo 49 Reglamento de las Condiciones Generales de Trabajo del Personal No Docente del Instituto Politécnico Nacional.

P25 ¿Qué se debe entender por alumno irregular?

Artículo 7 del Reglamento de Estudios Escolarizados para los Niveles Medio Superior y Superior del Instituto Politécnico Nacional, y 79, 81 y 110 del Reglamento Interno del Instituto Politécnico Nacional.

P26 ¿Qué debe entenderse por examen ordinario?

Artículo 31 del Reglamento de Estudios Escolarizados para los Niveles Medio Superior y Superior del Instituto Politécnico Nacional.

P27 ¿Qué debe entenderse por concurso de oposición de cátedra?

Artículos 26 y 27 del Reglamento de las Condiciones Interiores de Trabajo del Personal Académico del Instituto Politécnico Nacional.

P28 ¿Qué debe entenderse por profesor de carrera?

Artículo 17 y 19 del Reglamento de las Condiciones Interiores de Trabajo del Personal Académico del Instituto Politécnico Nacional.

P29 ¿Se puede presentar examen extraordinario con la finalidad de mejorar el promedio obtenido en la evaluación ordinaria?

Artículo 31 del Reglamento de Estudios Escolarizados para los Niveles Medio Superior y Superior del Instituto Politécnico Nacional

P30 ¿Se puede solicitar la revisión de los exámenes ordinarios?

Artículo 30 del Reglamento de Estudios Escolarizados para los Niveles Medio Superior y Superior del Instituto Politécnico Nacional.

P31 ¿Se puede cambiar el tema de tesis de maestría registrado?

Artículo 56 y 59 del Reglamento de Estudios de Posgrado del IPN.

P32 ¿Puede el personal académico cambiar su lugar de adscripción?
Artículos 139 y 140 del Reglamento de las Condiciones Interiores de Trabajo del Personal Académico del Instituto Politécnico Nacional

P33 ¿Cuál es el procedimiento para la elección de los presidentes de las academias?
Artículos 14, 15 y 16 del Reglamento de Academias del Instituto Politécnico Nacional.

P34 ¿Cuáles son los requisitos para iniciar un nuevo programa de posgrado?
Artículos 23 y 26 del Reglamento de Estudios de Posgrado del IPN.

P35 ¿Qué requisitos se deben cubrir para obtener una promoción?
Artículos 70, 77 y 51 del Reglamento de las Condiciones Interiores de Trabajo del Personal Académico del IPN

P36 ¿Cuáles son los requisitos para ser presidente de una academia?
Artículos 12 y 13 del Reglamento de Academias del Instituto Politécnico Nacional

P37 ¿Que debe entenderse por profesor invitado?
Artículos 96, 104 y 105 del Reglamento de Estudios de Posgrado del Instituto Politécnico

P38 ¿Se puede cambiar de carrera por una impartida en otra escuela?
Artículo 51 del Reglamento de Estudios Escolarizados para los Niveles Medio Superior y Superior del Instituto Politécnico Nacional, y 85 y 86 del Reglamento Interno del IPN

P39 ¿Tiene derecho al pago de alguna gratificación, el personal que ha ocupado un puesto de confianza por varios años, y que concluye su relación laboral por renuncia?
Artículo 8 de la Ley Federal de los Trabajadores al Servicio del Estado Reglamentaria del Apartado B del Artículo 123 Constitucional

P40 ¿Un trabajador después de disfrutar de una licencia sin goce de sueldo por seis meses, tiene derecho a vacaciones?

Artículo 30 de la Ley Federal de los Trabajadores al Servicio del Estado Reglamentaria del Apartado B del Artículo 123 Constitucional

Anexo C.

Para el análisis estadístico se evaluó la respuesta a cada pregunta en función del número de artículos-respuesta obtenidos por cada enfoque así como su posición, de la siguiente forma:

1) Calificación en función de la posición de los artículos-respuesta recuperados.

$$C_x = \frac{\sum_{i \in A} \frac{1}{i}}{\sum_{j=1}^{NAR} \frac{1}{j}}$$

donde, NAR es el número total de artículos-respuesta esperados y A contiene el conjunto de posiciones de los artículos-respuesta recuperados. Por ejemplo, si la pregunta *x* se respondía con 5 artículos y solo dos de ellos se recuperaron en las posiciones 3 y 5, entonces:

$$C_x = \frac{1 + \frac{1}{2}}{1 + \frac{1}{2} + \frac{1}{3} + \frac{1}{4} + \frac{1}{5}} = \frac{0.533}{2.283} = 0.657$$

2) Calificación en función del número de artículos-respuesta recuperados.

$$C_y = \frac{|A|}{NAR}$$

donde, |A| es el número de artículos-recuperados. Para el ejemplo anterior |A| = 2 y NAR = 5, así:

$$C_y = \frac{2}{5}$$

Calificación total:

$$C_t = w_1 C_x + w_2 C_y$$

Con, $w_1 = w_2 = 0.5$

$$C_t = 0.5 * 0.657 + 0.5 * 0.40 = 0.528$$

En la Tabla 1 y la Tabla 5, los encabezados corresponden a N.P. Número de pregunta. N.A-R. Número de Artículos-Respuesta. P.A-R.R. Posición Artículos-Respuesta Recuperados. MEV. Modelo del Espacio Vectorial. GSR. Grafo Sin Referencias. GCR. Grafo Con Referencias. C. Calificación total.

Tabla 1. Calificación MEV, GSR y GCR. Experimento I.

N.P.	N.A-R.	MEV		GSR		GCR	
		P.A-R.R.	C	P.A-R.R.	C	P.A-R.R.	C
1	5	1,2	0.528	1, 2, 36, 44	0.740	1, 2, 22, 28	0.746
2	1	4	0.625	5	0.600	2	0.750
3	2	2	0.417	1, 5	0.900	3, 8	0.653
4	2	-	0.000	6	0.306	1, 32	0.844
5	3	-	0.000	16, 54	0.355	6, 7	0.418
6	1	-	0.000	22	0.523	23	0.522
7	3	-	0.000	8, 9	0.398	8, 70	0.371
8	4	5	0.173	10, 56	0.278	14, 19, 56, 60	0.538
9	2	-	0.000	12, 38	0.537	13, 59	0.531
10	2	12	0.278	-	0.000	30, 39	0.520
11	1	5	0.600	35	0.514	10	0.550
12	2	-	0.000	49	0.257	2	0.417
13	1	-	0.000	67	0.507	66	0.508
14	2	-	0.000	2	0.417	3	0.361
15	2	1	0.583	4, 69	0.588	1, 28	0.845
16	2	-	0.000	9	0.287	16	0.271
17	2	3	0.361	2, 34	0.676	20,57	0.523
18	1	1	1.000	2	0.750	5	0.600
19	2	-	0.000	42, 70	0.513	26,34	0.523
20	1	32	0.516	-	0.000	22	0.523
21	1	-	0.000	30	0.517	46	0.511
Media		**0.242**		**0.460**		**0.549**	
Varianza		**0.091**		**0.051**		**0.023**	

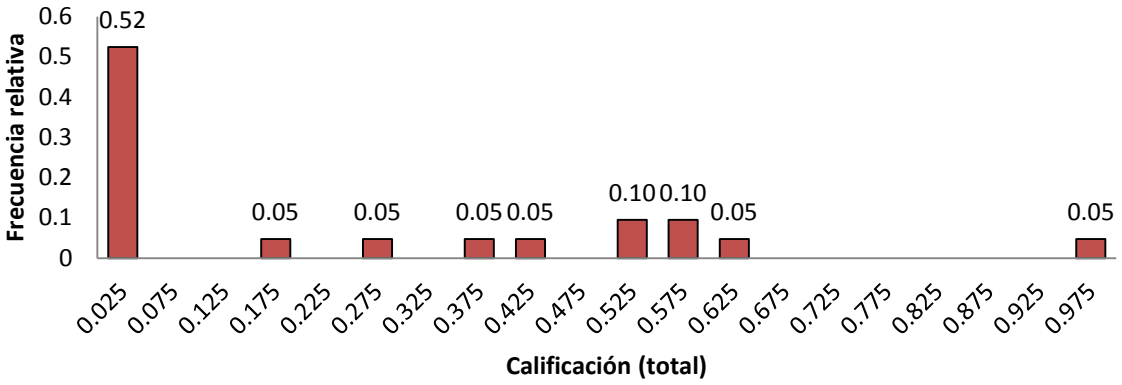


Gráfico 1. Histograma de frecuencia relativa. Calificación MEV.

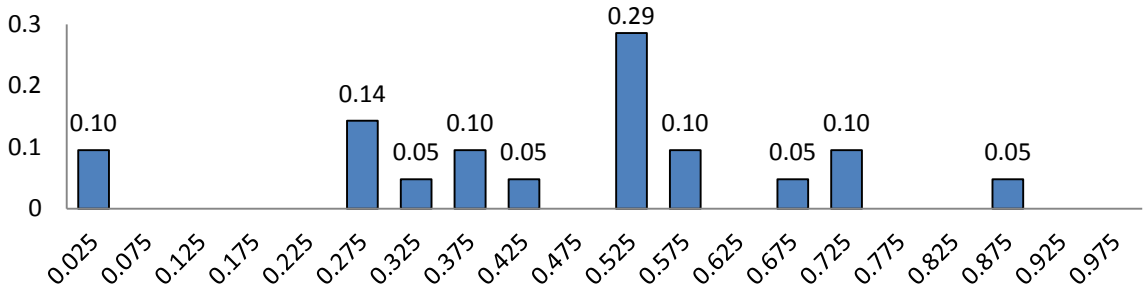


Gráfico 2. Histograma de frecuencia relativa. Calificación GSR.

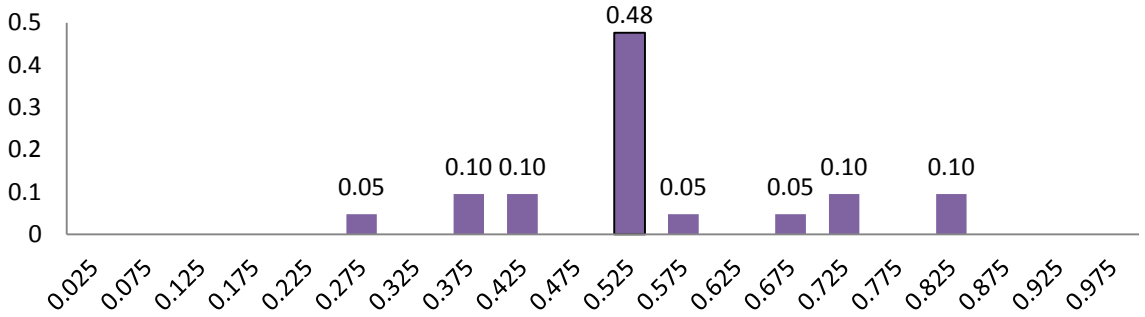


Gráfico 3. Histograma de frecuencia relativa. Calificación GCR.

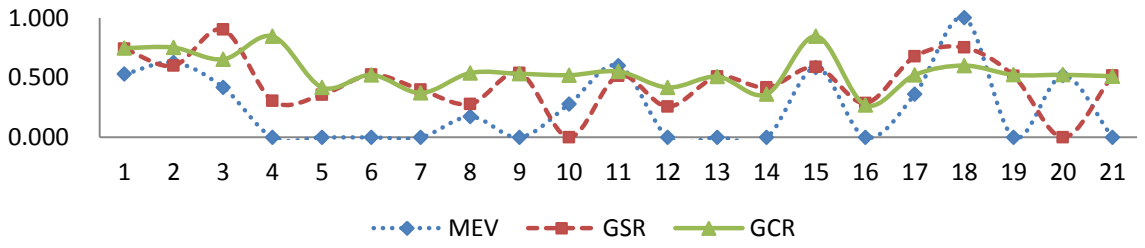


Grafico 4. Calificación MEV, GSR, GCR.

Con el objetivo de determinar si la diferencia entre los resultados con la implementación del modelo propuesto y los enfoques alternativos era estadísticamente significativa, se realizó una prueba-t a partir de calificar la respuesta a cada pregunta con base en la posición y el número de artículos-respuesta recuperados por cada enfoque. La prueba-t realizada se describe a continuación.

Sean μ_1 y μ_2 la calificación promedio del Modelo del Espacio Vectorial y del Grafo con Referencias, respectivamente, obtenidas en el experimento I.

1.- Partimos de que $H_o: \mu_1 = \mu_2$ *o* $\mu_D = \mu_1 - \mu_2 = 0$, *y*

2.- $H_1 = \mu_1 \neq \mu_2$ *o* $\mu_D = \mu_1 - \mu_2 \neq 0$

3.- Con: $\alpha = 0.05$

4.- Región crítica: $t < -2.086$ *y* $t > 2.086$, *donde* $t = \frac{\bar{d} - d_0}{s_D/\sqrt{n}}$ *con v = 20 grados de libertad.*

5.- Cálculos: La media y la desviación estándar para la diferencia d_i resultaron en:

$$\bar{d} = 0.218 \text{ y } s_D = 0.304$$

Por consiguiente $t = 3.2859$. Además:

$$P = P(|T| > 3.2859) \approx 0.0037$$

Esto indica que hay efectivamente una diferencia estadísticamente significativa entre el desempeño del MEV y el GSR.

Tabla 2. Prueba t para medias de dos muestras emparejadas. MEV y GSR.

Valor crítico de t (una cola)	**1.725**
P(T<=t) una cola	**0.001847**
Estadístico t	**-3.2859**
P(T<=t) dos colas	**0.0037**

Bajo las mismas condiciones (1, 2, 3 y 4) los cálculos para el caso del MEV y el GCR indican que la diferencia es aún mayor en comparación que con el GSR.

Tabla 3. Prueba t para medias de dos muestras emparejadas. MEV y GCR.

P(T<=t) una cola	**0.000013**
Estadístico t	**-5.4314**
P(T<=t) dos colas	**0.0000257**

Bajo las mismas condiciones (1-4) los cálculos para el caso del GSR y el GCR indicaron que la diferencia entre éstos era menor que en los casos anteriores.

Tabla 4. Prueba t para medias de dos muestras emparejadas. GSR y GCR.

P(T<=t) una cola	**0.039790**
Estadístico t	**-1.8471**
P(T<=t) dos colas	**0.0796**

Cómo se indicó en el análisis del Capítulo 4 la diferencia entre el GSR y el GCR resulto efectivamente menor que entre éstos y el MEV. Debido a lo cual la investigación se encamino al estudio de un mecanismo de recuperación más eficiente que aprovechara las referencias entre los artículos.

El mismo análisis se realizó con los resultados del Experimento III.

Tabla 5. Calificación (total) por pregunta. Lucene, Grafo y JIRS. Experimento III.

N.P.	N.A-R.	Lucene		Grafo		JIRS	
		P.A-R.R.	C	P.A-R.R.	C	P.A-R.R.	C
1	5	1,2,38,67	0.737	1,2,3,5	0.845	1,2,42	0.634
2	1	1	1.000	1	1.000	17	0.529
3	2	9,16	0.558	3,8	0.653	9,22	0.552
4	2	26	0.263	4,29	0.595	35	0.260
5	3	1,5	0.661	4,5,7	0.662	8,13,39	0.562
6	1	9	0.556	22	0.523	1	1.000
7	3	7,31	0.381	7,35,68	0.551	9,11	0.388
8	4	15,18,27,30	0.546	17,19,48,54	0.536	21,24,25,38	0.537
9	2	12	0.278	1,2	1.000	39	0.259
10	2	23,37	0.524	18,20	0.535	41	0.258
11	1	22	0.523	3	0.667	28	0.518
12	2	1	0.583	2,22	0.530	1,52	0.840
13	1	56	0.509	38	0.513	49	0.510
14	2	5	0.317	1,70	0.838	5,56	0.573
15	2	1,36	0.843	1,9	0.870	1,46	0.841
16	2	-	0.000	59	0.256	21	0.266
17	2	7,24	0.562	9,23	0.552	21,27	0.528
18	1	46	0.511	7	0.571	-	0.000
19	2	23	0.264	22,24	0.529	22	0.265
20	1	5	0.600	23	0.522	6	0.583

Tabla 5. Cont.

N.P.	N.A-R.	Lucene		Grafo		JIRS	
		P.A-R.R.	C	P.A-R.R.	C	P.A-R.R.	C
21	1	30	0.517	31	0.516	6	0.583
22	1	11	0.545	8	0.563	9	0.556
23	5	1,4,5,7	0.749	1,4,6,8,30	0.845	1,5,6,8	0.727
24	1	2	0.750	1	1.000	2	0.750
25	4	3,4	0.390	2,3,4,5	0.808	2,3	0.450
26	1	3	0.667	3	0.667	5	0.600
27	2	1,13	0.859	1,3	0.944	2,17	0.686
28	2	7	0.298	1,3	0.944	27	0.262
29	1	1	1.000	1	1.000	1	1.000
30	1	1	1.000	1	1.000	1	1.000
31	2	1,2	1.000	1,2	1.000	1,2	1.000
32	2	1,2	1.000	1,3	0.944	1,2	1.000
33	3	2,4,8	0.739	3,10,11	0.643	2,17,23	0.664
34	2	1,37	0.842	25,26	0.526	1,18	0.852
35	3	3	0.258	5	0.221	4	0.235
36	2	1,18	0.852	4,5	0.650	1,24	0.847
37	3	2,3,6	0.773	1,3,5	0.918	3,4,5	0.714
38	3	1,3,11	0.888	2,4,8	0.739	1,2,29	0.918
39	2	6,63	0.561	3,35	0.621	-	0.000
40	1	3	0.667	2	0.750	13	0.538
Media		0.614		0.701		0.582	
Varianza		0.0607		0.0431		0.0728	

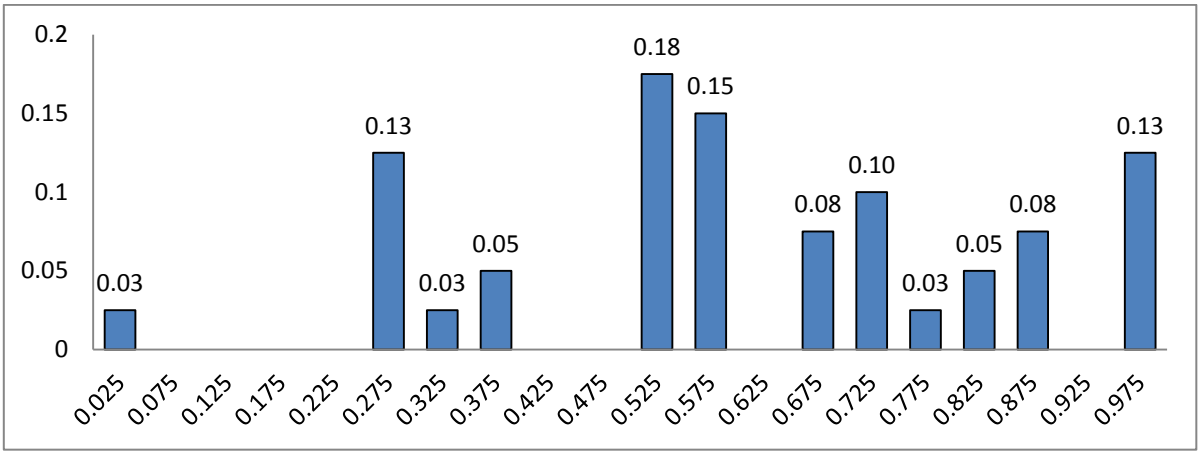


Gráfico 5. Frecuencia relativa calificación (total) Lucene. Experimento III.

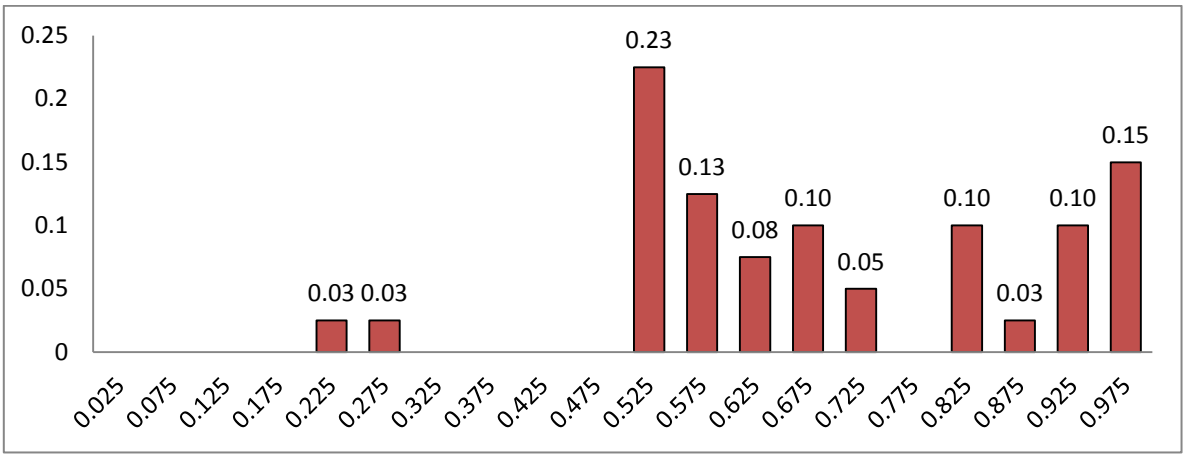


Gráfico 6. Histograma de frecuencia relativa calificación (total) Grafo. Experimento III.

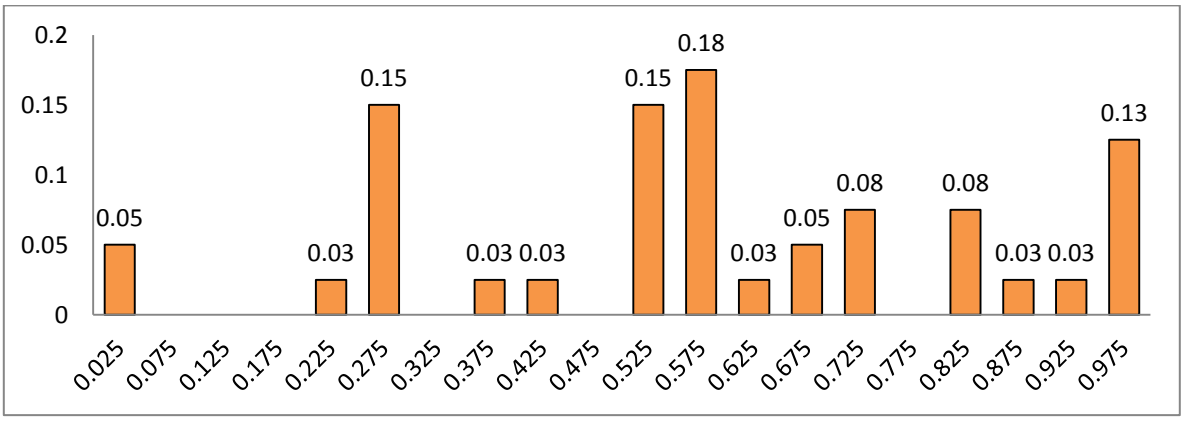


Gráfico 7. Frecuencia relativa calificación (total) JIRS. Experimento III.

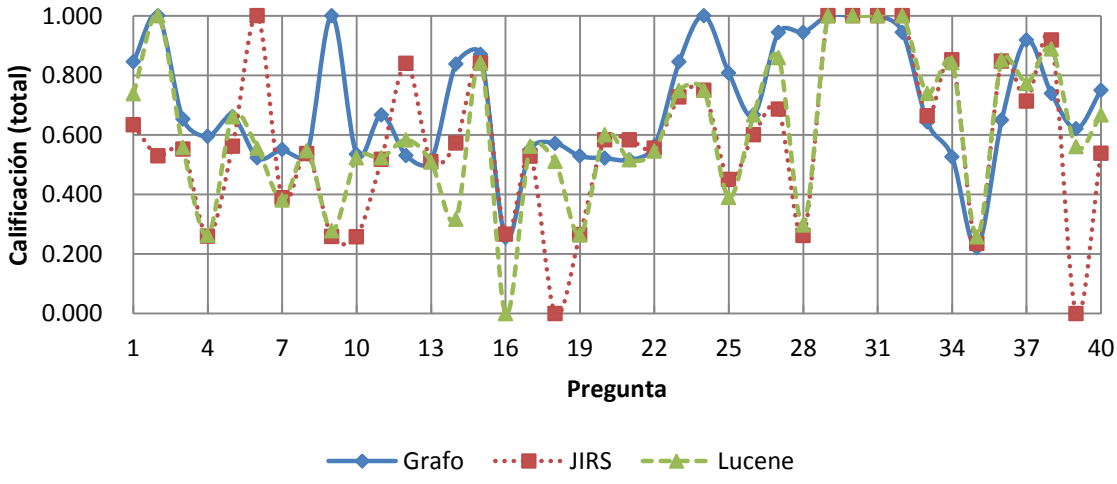


Gráfico 8. Calificación (total) por pregunta. Experimento III.

1.- Partimos de que $H_o: \mu_1 = \mu_2$ *o* $\mu_D = \mu_1 - \mu_2 = 0$, *y*

2.- $H_1 = \mu_1 \neq \mu_2$ *o* $\mu_D = \mu_1 - \mu_2 \neq 0$

3.- Con: $\alpha = 0.05$

4.- Región crítica: $t < -2.023$ *y* $t > 2.023$, *donde* $t = \frac{\bar{d} - d_0}{s_D/\sqrt{n}}$ *con* $v = 39$ *grados de libertad.*

Tabla 6. Prueba t para medias de dos muestras emparejadas: Lucene y Grafo.

Valor crítico de t (una cola)	1.685
Estadístico t	-2.63
P(T<=t) dos colas	0.012
P(T<=t) una cola	0.006

Tabla 7. Prueba t para medias de dos muestras emparejadas: JIRS y Grafo.

Estadístico t	-2.860
P(T<=t) dos colas	0.007
P(T<=t) una cola	0.003

Tabla 8. Prueba t para medias de dos muestras emparejadas: JIRS y Lucene.

Estadístico t	-1.102
P(T<=t) dos colas	**0.277**
P(T<=t) una cola	**0.139**