PROCEDIMIENTO PARA LA CREACIÓN DE ÍNDICES SEMÁNTICOS BASADOS EN ONTOLOGIAS DE DOMINIO.

ANEXOS

PROYECTO DE GRADO

DIGNORY JIMENA PEREZ URBANO
DIANA MARIBEL PEZO ARTEAGA

DIRECTOR: Magíster. Miguel Ángel Niño
ASESOR: PhD Carlos Cobos Lozada

Universidad del Cauca
Facultad de Ingeniería Electrónica y Telecomunicaciones
Departamento de Sistemas
Popayán 2010

TABLA DE CONTENIDO

ÍNDICE DE TABLAS

INDICES FIGURAS

1 ANEXO A

A continuación se exponen las fichas bibliográficas que hacen parte de la fase descriptiva siguiendo la metodología para la creación del Survey.

1.1 FICHAS BIBLIOGRAFICAS INDEXACION SEMANTICA

Ficha Bibliográfica	
Aspectos formales sobre el documento	
Autor : Rada Mihalcea and Dan Moldovan **Titulo:** SEMANTIC INDEXING USING WORDNET SENSES (Indexación semántica usando sentidos de WordNet) **Tipo de material:** Artículo -tesis	
Autor y titulo del documento :	
Resumen y palabras claves (Enfoque)	
Disciplina: señala el área del saber desde donde se define y aborda el objeto de estudio	Indexación Semántica y el uso de un algoritmo de desambiguación WSD.
Paradigma conceptual: indica la postura teórica de y/o metodológica que orienta la investigación.	Los autores deciden implementar la combinación de indexación basada en palabras y basada en sentidos, utilizando WordNet. De esta manera pretenden demostrar la efectividad en la relevancia de resultados ante una consulta hecha por el usuario
Referentes teóricos: describe los autores específicos en los cuales se apoya el autor y los fundamentos disciplinarios y teóricos que apuntalan la investigación	Se refiere a Indexación conceptual, expansión de consulta e indexación semántica. Se enfocan en esta ultima y se refiere a autores de documentos como: Sanderson, 2000; Yarowsk, 1993; Brill, 1992; Krovetz, 1997; Gonzalo, 1998
Conceptos principales: señala el soporte técnico de la tesis, explicaciones, problemas, ideas y conclusiones planteadas para la investigación	• Sistema de recuperación de información booleano que adiciona palabras semánticas a las palabras clásicas basadas en indexación. • Combinación del enfoque basado en palabras y basado en sentidos. • Metodología para construir representaciones semánticas de texto abierto en palabras y en nivel de ubicación. • Adicionan información léxica y semántica a la consulta y a los documentos. • El proceso de desambiguación identifica el significado de las palabras basadas en sentidos de WordNet.
Hipótesis: registra las proposiciones que sirven de guía a la investigación (conjeturas sobre el objeto de estudio) (suposiciones a partir de los cuales se organizó la investigación)	• Se propone un sistema que combine los beneficios de la indexación basada en palabras y en sentidos. • Las palabras y sentidos son indexados en el texto de entrada y la recuperación es realizada usando una o dos de estas fuentes de información. • La clave es el uso de un modulo Word Sense Disambiguation (WSD), el cual realiza una desambiguacion semi-completa pero precisa. • La salida debe ser un nuevo documento en el cual, cada palabra es reemplazada con un nuevo formato.
Tesis: describe las proposiciones que sintetizan las generalizaciones sobre el objeto de estudio ("generalizaciones empíricas" existentes en las	• Optan por un algoritmo de desambiguacion que es semi-completo pero es altamente preciso (cerca de 92% de eficiencia).

cuales se apoyó el trabajo)	• El índice es creado usando las palabras como cadenas léxicas (para asegurar la recuperación basada en la palabra) y las etiquetas semánticas (para la recuperación basada en el sentido). • Se presentan los procedimientos para identificar el sentido correcto de una palabra. • El proceso de indexación toma un grupo de archivos de documentos y produce un nuevo índice. Todos los elementos del documento son indexados. • La información obtenida del modulo WSD es usada para el principal proceso de indexación donde la palabra madre y el lugar están indexados junto al Synset (conjunto de sinónimos) WordNet (si existe).
Tipo de investigación: señala el tipo de investigación (exploratoria: examina un tema o problema poco estudiado; Descriptiva: responde al qué del fenómeno analizado; o Explicativa: responde al porqué del fenómeno analizado)	Exploratoria: examina la indexación semántica con la implementación de un algoritmo de desambiguacion de palabras y la combinación de palabras y sentidos (significados).
Resumen y palabras claves (metodología)	
Tipo de metodología: Describe el tipo de metodología utilizada (Cualitativa: Privilegia la comprensión de los fenómenos y la interacción investigador - investigado; Cuantitativa: Privilegia la cuantificación de los fenómenos estudiados y la no interacción investigador - investigado; o Mixta: Combina procedimientos cualitativos y cuantitativos)	Metodologia Mixta: combina procedimientos cualitativos mostrando los pasos de la investigación y el proceso; asi mismo muestra en valores y porcentajes los resultados del estudio.
Técnicas: Describe el tipo de herramientas utilizadas en la recolección, registro y sistematización de la información por el autor en la investigación	Utilizan documentos bibliográficos guiarse, la base de datos de WordNet para extraer las palabras y los significados, un algoritmo de desambiguacion de palabras y un conjunto de documentos para realizar las pruebas de relevancia.
Resultados:	
Conclusiones: Señala las conclusiones que especifica explícitamente el autor en el documento	La indexación semántica ofrece una mejora en las técnicas actuales de recuperación de información. La clave es utilizar un WSD para grandes colecciones de documentos. Se mostro el método WSD para dominios abiertos, lo cual es rápido y eficiente. Además, usaron un enfoque de indexación híbrida que combina la indexación basada en palabras y la basada en sentidos. El algoritmo WSD presentado es nuevo para la comunidad NLP y demuestra que es muy adecuado para una tarea como la indexación semántica.
Recomendaciones: Registra las recomendaciones que hace el autor en el documento	
Aspectos formales sobre el documento:	
Resumen del documento:	
Palabras claves:	Indexacion Semantica, algoritmo de Desambiguacion, word-based y sense-based approach,
Anexos Referencia los anexos que se consideran relevantes	
Glosas Indica las inconsistencias, omisiones o errores detectados por quien analizó el documento	No se especificó muy bien el momento en que se pasa al proceso real de indexación, pues al principio se describe el proceso de desambiguación y lo que

	sigue no se expresa exactamente como indexación semántica.
Comentarios Contiene las reflexiones, relaciones, inferencias y asociaciones que hace quien analizó el documento. También deben indicarse los aportes, vacíos y limitaciones del documento estudiado en relación con el tema central respectivo y/o el área temática correspondiente.	El proceso de desambiguación está muy bien explicado, a continuación es el proceso de indexación que se lleva a cabo pero no se describe tan profundo como el anterior. En el ejemplo especifican mejor en qué momento se pasa al proceso de indexación.

Tabla 1. Ficha B 1 indexación semántica

Ficha Bibliográfica	
Aspectos formales sobre el documento	
Autor : Paolo Rosso, Antonio Molina, Ferran Pla, Daniel Jimenez, and Vicent Vidal **Titulo:** INFORMATION RETRIEVAL AND TEXT CATEGORIZATION WITH SEMANTIC INDEXING (Recuperación de Información y Clasificación de Texto con indexación semántica) **Tipo de material:** Artículo (Soportado por el Spanish Research Projects CICYT TIC2000-0664-C02 and TIC2003-07158-C04-03.)	
Autor y titulo del documento :	
Resumen y palabras claves (Enfoque)	
Disciplina: señala el área del saber desde donde se define y aborda el objeto de estudio	Indexación Semántica, WSD y modelos especializado de Markov
Paradigma conceptual: indica la postura teórica de y/o metodológica que orienta la investigación.	Se enfocan en la eficiencia de la indexación semántica con WordNet senses y la utilizacion de SHMMs.
Referentes teóricos: describe los autores específicos en los cuales se apoya el autor y los fundamentos disciplinarios y teóricos que apuntalan la investigación	
Conceptos principales: señala el soporte técnico de la tesis, explicaciones, problemas, ideas y conclusiones planteadas para la investigación	• Los documentos han sido etiquetados por sentidos usando un sistema WSD basados en modelos especializados ocultos de Markov (SHMMs). • El SemCor corpus fue usado para aprender los modelos. Este provee las características de entrada: palabras (W), Lemas (L) y etiquetas POS correspondientes (P). • El enfoque HMM no puede incluir diferentes tipos de información lingüística disponible, por lo tanto utilizaron Specialized HMM.
Hipótesis: registra las proposiciones que sirven de guía a la investigación (conjeturas sobre el objeto de estudio) (suposiciones a partir de los cuales se organizó la investigación)	• Según un estudio previo, el modelo de espacio vectorial clásico para IR, da mejores resultados si se utiliza Synsets de WordNet como el espacio en lugar de usar términos de indexación (hasta 29% de mejoría en los resultados experimentales se obtuvo de una colección de prueba manualmente ambigüedad derivada del corpus SemCor). • La redefinición de HMM se puede hacer en dos procesos: • El proceso de selección: redefine el vocabulario de entrada. • El proceso de especialización: redefine las etiquetas de salida.
Tesis: describe las proposiciones que sintetizan las generalizaciones sobre el objeto de estudio ("generalizaciones empíricas" existentes en las cuales se apoyó el trabajo)	• En el sistema de WSD usado, definieron la salida de etiquetas semánticas establecidas por considerar algunos datos estadísticos que fueron extraídos de los corpus anotado. En el corpus SemCor, cada palabra anotada es etiquetada con una clave *sense_key* que tiene la forma *Lemma%lex_sense*. Consideran el campo

	lex_sense de la sense_key asociada a cada Lema, como la etiqueta semántica con el fin de reducir el tamaño del conjunto de etiquetas de salida. Esto no presenta pérdida de información debido a la posibilidad de obtener la clave *sense_key* concatenando el *Lema* a la etiqueta de salida. • Decidieron representar cada documento a través de un vector de synsets correspondiente, en lugar de un vector de términos de referencia. La desambiguación del significado de cada término se obtuvo utilizando SHMMs. • La tarea de IR se llevó a cabo inicialmente empleando la técnica de K-Means agrupación bisectriz del esférico. Su algoritmo intenta unirse a las ventajas del algoritmo de bisección K-Means con las ventajas de una Versión modificada del K-Means esférico.
Tipo de investigación: señala el tipo de investigación (exploratoria: examina un tema o problema poco estudiado; Descriptiva: responde al qué del fenómeno analizado; o Explicativa: responde al porqué del fenómeno analizado)	Descriptiva
Resumen y palabras claves (metodología)	
Tipo de metodología: Describe el tipo de metodología utilizada (Cualitativa: Privilegia la comprensión de los fenómenos y la interacción investigador - investigado; Cuantitativa: Privilegia la cuantificación de los fenómenos estudiados y la no interacción investigador - investigado; o Mixta: Combina procedimientos cualitativos y cuantitativos)	Mixta
Técnicas: Describe el tipo de herramientas utilizadas en la recolección, registro y sistematización de la información por el autor en la investigación	Documentos previos de referencias. Una herramienta para modelar lenguaje estático, recuperación de texto, clasificación y clustering, disponible en www.cs.cmu.edu/~mccallum/bow/ WSD SHMM WordNet Synsets
Resultados:	
Conclusiones: Señala las conclusiones que especifica explícitamente el autor en el documento	• En futuros trabajos, las dos representaciones del vector de cada documento deben ser combinadas, a fin de tener en cuenta con diferentes pesos, los términos y synsets de WordNet, al mismo tiempo. • La introducción de la semántica permitió una pequeña mejora de la precisión: 79% (indexación sentido) vs 77,68% (indexación de plazo).
Recomendaciones: Registra las recomendaciones que hace el autor en el documento	
Aspectos formales sobre el documento:	
Resumen del documento:	
Palabras claves:	Indexación Semántica, WordNet
Anexos Referencia los anexos que se consideran relevantes	Modelo espacio vector y synsets en [1] J. Gonzalo, F. Verdejo, I. Chugur, J. Chigarr¶an. Indexing with WordNet Synsets can improve Text Retrieval. In: Proc. of the Workshop on Usage of WordNet for NLP, 1998

Tabla 2. Ficha B 2 indexación semántica

Ficha Bibliográfica	
Aspectos formales sobre el documento	
Autor : Bo-Yeong Kang **Titulo:** A NOVEL APPROACH TO SEMANTIC INDEXING BASED ON CONCEPT (Una nueva aproximación a la indexación semántica basada en el concepto) **Tipo de material:** Artículo -tesis	
Autor y titulo del documento :	
Resumen y palabras claves (Enfoque)	
Disciplina: señala el área del saber desde donde se define y aborda el objeto de estudio	Indexación Semántica y modelo espacio vector
Paradigma conceptual: indica la postura teórica de y/o metodológica que orienta la investigación.	
Referentes teóricos: describe los autores específicos en los cuales se apoya el autor y los fundamentos disciplinarios y teóricos que apuntalan la investigación	En varios conceptos del documento, especialmente sobre las cadenas léxicas y la estructura de cohesión léxica, se refieren al autor Morris y sus trabajos de 1991. .
Conceptos principales: señala el soporte técnico de la tesis, explicaciones, problemas, ideas y conclusiones planteadas para la investigación	• Un documento se considera como un concepto complejo que consta de varios conceptos, es reconocido como un vector en el concepto de espacio vectorial. • Las cadenas léxicas enlazan los elementos léxicos relacionados en un documento, y que representan la estructura de la cohesión léxica de un documento. Cada cadena léxica es considerada como un concepto que expresa el significado de un documento, además, cada concepto fue extraído por cadenas léxicas. • Model Space Vector Concept. El espacio de concepto es un espacio n-dimensional formado por los ejes n concepto. Cada eje representa un concepto de concepto, y tiene una magnitud definida. En el concepto de espacio, un documento T está representado por la suma de los vectores de concepto n-dimensionales. • Cantidad de información es la cantidad semántica de un texto, concepto o palabra en la información de los documentos. • La proporción (ratio) de información es la relación de cantidad de información de una etiqueta comparativa con la cantidad de información de un texto, palabra o concepto.
Hipótesis: registra las proposiciones que sirven de guía a la investigación (conjeturas sobre el objeto de estudio) (suposiciones a partir de los cuales se organizó la investigación)	• Se decide utilizar métodos basados en los fenómenos lingüísticos para mejorar el rendimiento de la indexación. Se centran en el concepto de espacio vectorial para la extracción y los índices de ponderación. Además, proponen compensar las limitaciones de los métodos basados en la frecuencia de término mediante el empleo de cadenas léxicas. • Plantean un enfoque que cambia el índice de base de la ponderación de término, considerando la semántica y los conceptos de un documento. Así, los conceptos de un documento se pueden entender, y los índices semánticos y sus pesos son derivados de dichos conceptos.
Tesis: describe las proposiciones que sintetizan las generalizaciones sobre el objeto de estudio ("generalizaciones empíricas" existentes en las cuales se apoyó el trabajo)	• El sistema propuesto consta de cuatro componentes principales: o La construcción de las cadenas léxicas. o La ponderación de cadenas y nombres. o Reponderación del término basada en el concepto. o La extracción del índice del término

	semántico. Los dos primeros componentes se basan en la extracción de concepto utilizando cadenas léxicas, y los dos últimos componentes están relacionados con la extracción del índice de término basado en el concepto de espacio vectorial. • Un documento se considera como un concepto complejo que consta de varios conceptos, es reconocido como un vector en el espacio concepto vectorial. • Cada concepto fue extraído por las cadenas léxicas. Los conceptos extraídos y los elementos léxicos se anotaron en el momento de la construcción de cadenas léxicas. Cada cadena fue representada como un vector en el espacio concepto de vector, y el vector de texto está formado por los vectores de concepto. La importancia semántica de los conceptos y las palabras se normalizan de acuerdo con el vector de texto en su conjunto. Se extraen los índices que incluyen su peso semántico. • El índice semántico y el peso son extraídos de acuerdo al valor numérico de la cantidad de información y la proporción de información.
Tipo de investigación: señala el tipo de investigación (exploratoria: examina un tema o problema poco estudiado; Descriptiva: responde al qué del fenómeno analizado; o Explicativa: responde al porqué del fenómeno analizado)	Explicativa
Resumen y palabras claves (metodología)	
Tipo de metodología: Describe el tipo de metodología utilizada (Cualitativa: Privilegia la comprensión de los fenómenos y la interacción investigador - investigado; Cuantitativa: Privilegia la cuantificación de los fenómenos estudiados y la no interacción investigador - investigado; o Mixta: Combina procedimientos cualitativos y cuantitativos)	Mixta: combina la investigación teórica y los resultados ponderados.
Técnicas: Describe el tipo de herramientas utilizadas en la recolección, registro y sistematización de la información por el autor en la investigación	Estudios previos del modelo espacio vectorial Colección de documentos para realizar las pruebas.
Resultados:	
Conclusiones: Señala las conclusiones que especifica explícitamente el autor en el documento	En este trabajo se pretende modificar los métodos de indexación básica mediante la presentación de un nuevo enfoque que utilice un modelo de espacio de conceptos vectorial para la extracción y los índices de ponderación. El experimento para la indexación semántica apoya la validez del enfoque que se presenta, que es capaz de captar la importancia semántica de una palabra en el conjunto del documento. Desde los resultados experimentales, el método propuesto alcanza un nivel de rendimiento comparable a los métodos de ponderación mayor. En un experimento, no comparan este método con el de la frecuencia inversa de documentos (IDF), porque se deja para trabajos futuros.
Recomendaciones: Registra las recomendaciones que hace el autor en el documento	
Aspectos formales sobre el documento:	
Resumen del documento:	Se plantea el método de indexación eficiente basado

	en un concepto de espacio vectorial que es capaz de representar el contenido semántico de un documento. Las dos medidas de información: la cantidad de información y el ratio (proporción) de información, se definen para representar el grado de la relevancia semántica en un documento. Se propone un método para compensar las limitaciones de los métodos basados en la frecuencia de término mediante la explotación de los elementos léxicos relacionados. Además, con la proporción de información, este enfoque es independiente de la longitud del documento.
Palabras claves:	
Anexos Referencia los anexos que se consideran relevantes	Morris and G. Hirst, Lexical cohesion computed by thesaural relations as an indicator of the structure of text, Computational Linguistics 17(1) 1991. M.-F. Moens, Automatic Indexing and Abstracting of Document Texts, Kluwer Academic Publishers, 2000.

Tabla 3. Ficha B 3 indexación semántica

Ficha Bibliográfica	
Aspectos formales sobre el documento	
Autor : Marco Suárez Barón, Kathleen Salinas Valencia **Titulo:** AN APPROACH TO SEMANTIC INDEXING AND INFORMATION RETRIEVAL (Una aproximación a la indexación semántica y a la recuperación de información) 2008 **Tipo de material:** Artículo – tesis	
Autor y titulo del documento :	
Resumen y palabras claves (Enfoque)	
Disciplina: señala el área del saber desde donde se define y aborda el objeto de estudio	Indexación Semántica. Describen otras técnicas de recuperación de información
Paradigma conceptual: indica la postura teorica de y/o metodológica que orienta la investigación.	Se describen las actuales técnicas (overview) de recuperación de información, sus bases y principios para dar una idea general de ellas. Se profundiza en la indexación semántica y similitud semántica teniendo en cuenta el uso de ontologías.
Referentes teóricos: describe los autores específicos en los cuales se apoya el autor y los fundamentos disciplinarios y teóricos que apuntalan la investigación	En la descripción de indexación semántica se apoyan en: [E. Brill. “Some advances in rule-based part of speech tagging”] y [C. Wei, T. H. Cheng, Y. C. Pai. “Semantic enrichment in knowledge repositories: annotating reply semantic relationships between discussion documents”]
Conceptos principales: señala el soporte técnico de la tesis, explicaciones, problemas, ideas y conclusiones planteadas para la investigación	• Modelos de recuperación de información: índices semánticos en temas particulares. • Para la recuperación de información se distinguen características de índices semánticos que son relaciones semánticas entre términos de índices controlados. • El término de indexación ha sido ampliamente utilizado para referirse al proceso de construcción de dichas representaciones. • La indexación puede ser manual o automática, la segunda es menos costosa, provee un contador de cada ocurrencia de una palabra en un documento. • Técnica sofisticada para automatizar la indexación: Latent Semantic Indexing. • Indexación semántica: Usar información semántica para proveer calidad. Los objetos son indexados por los conceptos. Emplea conjunto de relaciones semánticas entre términos de índices determinados por medio de tesauros. • El modelo de espacio vectorial se basa en la

	representación de ambos documentos y consultas de ponderados como vectores en el espacio del índice de términos, cuya dimensionalidad está determinada por el tamaño del vocabulario utilizado en el proceso de indexación. • Un modelo probabilístico es un enfoque novedoso para la indexación de documentos automatizado que se basa en un modelo estadístico para el análisis de clase latente de factor de los datos de recuento.
Hipótesis: registra las proposiciones que sirven de guía a la investigación (conjeturas sobre el objeto de estudio) (suposiciones a partir de los cuales se organizó la investigación)	• Analizar las técnicas más utilizadas de indexación semántica para mostrar la efectividad en la recuperación de información.
Tesis: describe las proposiciones que sintetizan las generalizaciones sobre el objeto de estudio ("generalizaciones empíricas" existentes en las cuales se apoyó el trabajo)	• Un índice semántico es inherentemente multidimensional, desde cualquier combinación de propiedades en un concepto como un elemento de indexación. • Un índice semántico en su conjunto, es altamente adaptable a los patrones de uso. Los conceptos de indexación pueden ser adicionados o quitados como se desee, haciéndolo muy densos y precisos con respecto a los intereses de grupos de individuos. • Dado que el índice es en realidad un conjunto de descripciones parciales de los objetos indexados, mucha información se puede extraer del índice solo, sin acceso a las descripciones individuales de todos. • Se describe el inconveniente de las técnicas de recuperación al intentar combinar las palabras de una consulta con las palabras de los documentos, es decir, optar solo por las palabras claves en los textos.
Tipo de investigación: señala el tipo de investigación (exploratoria: examina un tema o problema poco estudiado; Descriptiva: responde al qué del fenómeno analizado; o Explicativa: responde al porqué del fenómeno analizado)	Exploratoria: examina las técnicas de recuperación de información enfocándose en la Indexación Semántica.
Resumen y palabras claves (metodología)	
Tipo de metodología: Describe el tipo de metodología utilizada (Cualitativa: Privilegia la comprensión de los fenómenos y la interacción investigador - investigado; Cuantitativa: Privilegia la cuantificación de los fenómenos estudiados y la no interacción investigador - investigado; o Mixta: Combina procedimientos cualitativos y cuantitativos)	Cualitativa: Muestra los enfoques de autores en el tema y la propia opinión en la indexación semántica
Técnicas: Describe el tipo de herramientas utilizadas en la recolección, registro y sistematización de la información por el autor en la investigación	Bibliografía de varios autores en técnicas de recuperación de información e indexación semántica además del modelo de espacio vectorial, modelo probabilístico, indexación semántica latente.
Resultados:	
Conclusiones: Señala las conclusiones que especifica explícitamente el autor en el documento	La indexación semántica permite indexar la gran colección de documentos que se obtiene mediante la determinación de las palabras clave y también considerar los grupos de palabras clave equivalente, o de términos relacionados con las palabras clave encontradas. La similitud de medición ha sido estudiada ampliamente en muchas áreas de investigación. Aproximaciones a la similitud semántica se pueden sacar de muchos campos, tales como bases de datos distribuidas, recuperación de información, integración de datos y procesamiento del lenguaje natural. Las técnicas específicas para la evaluación de la similitud semántica que podrían ser aplicables en este contexto son: Utilizar un conjunto de sinónimos para direccionar el uso

	de diferentes términos para describir el mismo concepto. El uso de un contexto léxico de un término conceptual, incluyendo la evaluación del efecto semántico de términos en el mismo contexto. El uso de los enlaces taxonómicos jerárquicos y la ponderación de los valores asignados a estos enlaces en función de la profundidad del vínculo en la taxonomía.
Recomendaciones: Registra las recomendaciones que hace el autor en el documento	
Aspectos formales sobre el documento:	
Resumen del documento:	Se presenta un acercamiento a los índices semánticos de temas particulares así como los más importantes modelos para la recuperación de información. Estos modelos serán construidos dinámicamente a través de anotaciones identificadas en recursos de la web y en las definiciones de anotaciones ontológicas de los términos utilizados. Los índices están proyectados a sus agentes activos que 'conocen' qué tópicos pueden manejar (e.j; encontrar el contenido para...) con base en sus propias ontologías.
Palabras claves:	Indexación Semántica, ontologías, similud semántica, técnicas de recuperación de información, P2P, taxonomía.
Comentarios Contiene las reflexiones, relaciones, inferencias y asociaciones que hace quien analizó el documento. También deben indicarse los aportes, vacíos y limitaciones del documento estudiado en relación con el tema central respectivo y/o el área temática correspondiente.	

Tabla 4. Ficha B 4 indexación semántica

Ficha Bibliográfica	
Aspectos formales sobre el documento	
Autor : Albrecht Schmiedel **Titulo:** SEMANTIC INDEXING BASED ON DESCRIPTION LOGICS (Indexación Semántica Basada en lógicas de descripción) **Tipo de material:** Artículo -tesis	
Autor y titulo del documento :	
Resumen y palabras claves (Enfoque)	
Disciplina: señala el área del saber desde donde se define y aborda el objeto de estudio	Indexación Semántica y lógica de descripción
Paradigma conceptual: indica la postura teórica de y/o metodológica que orienta la investigación.	Se enfoca en una lógica de descripción basada en una aproximación a la indexación
Referentes teóricos: describe los autores específicos en los cuales se apoya el autor y los fundamentos disciplinarios y teóricos que apuntalan la investigación	
Conceptos principales: señala el soporte técnico de la tesis, explicaciones, problemas, ideas y conclusiones planteadas para la investigación	• Clave (Key): en lugar de utilizar valores de los atributos, los elementos de indexación pueden ser conceptos arbitrarios estructurados conforme a lo dispuesto por el lenguaje terminológico, en este caso BACK (Hoppe 1993). • Propiedades de los sistemas basados en lógica de descripción: o La habilidad para manejar algún grado de información parcial en conjunto con un supuesto del mundo abierto. o La habilidad para describir individuos con conceptos complejos y usar estas descripciones como respuesta a las consultas.

Hipótesis: registra las proposiciones que sirven de guía a la investigación (conjeturas sobre el objeto de estudio) (suposiciones a partir de los cuales se organizó la investigación)	• Construir un índice persistente en un gran número de objetos para clasificarlos con respecto al conjunto de conceptos de indexación. • Almacenamiento de la relación resultante entre identificadores de objetos y los conceptos indexados más específicos en un archivo.
Tesis: describe las proposiciones que sintetizan las generalizaciones sobre el objeto de estudio ("generalizaciones empíricas" existentes en las cuales se apoyó el trabajo)	• Las descripciones son términos construidos con: o Operadores de formación de términos: and, all, some, …, las constantes lógicas que provee el lenguaje. o Los conceptos primitivos y roles: introducidos por el usuario. o Conceptos y roles definidos: "dulce sintáctico" para abreviar descripciones posiblemente complejas. • Las descripciones son usadas también para definir reglas que son expresadas como implicaciones entre dos descripciones. • Basandose en el tipo de entrada (de la consulta), el sistema calcula: o Las relaciones de inclusión y disyunción para conceptos, ejemplo: Por cada par de conceptos calcula si uno incluye al otro o si son disyuntos. o Si el conjunto de conceptos es (y no es) una instancia de cada individuo. • Basándose en la relación almacenada y las definiciones originales de conceptos, puede construirse una nueva base de conocimiento que contiene sólo la clasificación de los individuos con respecto a los conceptos indexados.
Tipo de investigación: señala el tipo de investigación (exploratoria: examina un tema o problema poco estudiado; Descriptiva: responde al qué del fenómeno analizado; o Explicativa: responde al porqué del fenómeno analizado)	Exploratoria
Resumen y palabras claves (metodología)	
Tipo de metodología: Describe el tipo de metodología utilizada (Cualitativa: Privilegia la comprensión de los fenómenos y la interacción investigador - investigado; Cuantitativa: Privilegia la cuantificación de los fenómenos estudiados y la no interacción investigador - investigado; o Mixta: Combina procedimientos cualitativos y cuantitativos)	Cualitativa
Técnicas: Describe el tipo de herramientas utilizadas en la recolección, registro y sistematización de la información por el autor en la investigación	
Resultados:	
Conclusiones: Señala las conclusiones que especifica explícitamente el autor en el documento	El mecanismo de indexación semántica depende mucho del razonamiento con descripciones que proveen los sistemas terminológicos. Los elementos de indexación son potencialmente complejas descripciones lógicamente relacionadas por inclusiones y disyunciones.
Recomendaciones: Registra las recomendaciones que hace el autor en el documento	
Aspectos formales sobre el documento:	
Resumen del documento:	Un método para la construcción y mantenimiento de un índice semántico usando un sistema basado en

	lógicas de descripción. Se construye un índice persistente en un gran número de objetos con respecto a un conjunto de conceptos de indexación y almacenamiento de la relación resultante entre ids de objetos y conceptos de índices más específicos, en un archivo. Estos archivos pueden ser actualizados incrementalmente. El índice puede ser usado y basado en la inclusión y el razonamiento disjunto con respecto a conceptos de indexación, las instancias son inmediatamente categorizadas en hits, faltantes o candidatos con respecto a la consulta. Basados solo en el índice, la retroalimentación concerniente a la cardinalidad de la consulta (límites superior e inferior) puede ser provista durante la edición de la consulta.
Palabras claves:	

Tabla 5. Ficha B 5 indexación semántica

Ficha Bibliográfica	
Aspectos formales sobre el documento	
Autor : Thanh Nguyen, Tuoi Phan **Titulo:** THE EFFECT OF SEMANTIC INDEX IN INFORMATION RETRIEVAL DEVELOPMENT (El efecto del índice semántico en el desarrollo de la recuperación de información) **Tipo de material:** Articulo	
Autor y titulo del documento :	
Resumen y palabras claves (Enfoque)	
Disciplina:	Área de la informática
Paradigma conceptual :	
Referentes teóricos:	
Conceptos principales:	Ontología: Una ontología proporciona el vocabulario común de un dominio específico y define más o menos formalmente el sentido de los términos y algunas de sus relaciones
Hipótesis:	Con el fin de mejorar la calidad de recuperación de información y para apoyar a los usuarios en búsquedas de datos, esta investigación se centra en el desarrollo de un índice semántico y su transformación como la función de extracción de los índices y búsqueda de datos con lo cual se pretende mejorar el sistema de recuperación de información semántica. Además este enfoque varios atributos semánticos y de información, para apoyar las consultas de búsqueda por búsqueda de significado. Este modelo SIRS, se puede aplicar no sólo para un sistema de recuperación de información, sino también sistema de bibliotecas, para apoyar a los usuarios a buscar información necesaria, ya sea en documentos, libros en las bibliotecas digitales.
Tesis:	
Tipo de investigación:	Exploratoria
Resumen y palabras claves (metodología)	
Tipo de metodología:	•
Técnicas:	Para la creación del índice normal ellos utilizan herramientas existentes de indexación como Lucene y lémur.
Resultados:	
Conclusiones:	El resultado de este proyecto muestra que el tamaño de los índices no es muy alto y el tiempo de indexación es alto. Por lo tanto las tareas para el próximo trabajo futuro para mejorar la indexación semántica y la búsqueda de algoritmos es reducir el tiempo de procedimiento y para aplicar el SIRS para el idioma vietnamita, especialmente a los

	documentos de VietNam y consultas, y trabajar más en sus experimentos para los algoritmos de fusión y extracción de los índices semánticos.
Recomendaciones:	
Aspectos formales sobre el documento:	
Resumen del documento:	En la Recuperación de Información (IR) de los sistemas, muchos de ellos publicados en investigaciones, proponen soluciones para la indización de documentos para mejorar el rendimiento de la tarea de búsqueda de la consulta. Para recuperar los datos de consultas complejas sobre la información semántica, incluidos los atributos semánticos para índice tradicional es necesario e importante. Este documento presenta una propuesta para elaborar el índice semántico, así como la heurística de su aplicación como la fusión y extracción de manera que apoye a las IR para ser más fuerte y más útil. La experimentación con los documentos de Inglés muestra que el enfoque de los mecanismos es factible y que puede afectar el desarrollo de IR en futuro.
Palabras claves:	Indexación semántica, índice semántico, fusionar índices semánticos, partición, extracto índice semántico.

Tabla 6. Ficha B 6 indexación semántica

Ficha Bibliográfica	
Aspectos formales sobre el documento	
Autor : Conrad T. K. Chang Bruce R. Schatz **Titulo:** PERFORMANCE AND IMPLICATIONS OF SEMANTIC INDEXING IN A DISTRIBUITED ENVIROMENT (Rendimiento y Consecuencias de la indexación semántica en un entorno distribuido) **Tipo de material: Artículo – tesis**	
Autor y titulo del documento :	
Resumen y palabras claves (Enfoque)	
Disciplina: señala el área del saber desde donde se define y aborda el objeto de estudio	Indexación Semántica, computación distribuida
Paradigma conceptual: indica la postura teórica de y/o metodológica que orienta la investigación.	El prototipo está motivado por el deseo de proporcionar una información más eficiente y eficaz con respecto a los sistemas de recuperación de información con los que contamos hoy en día.
Referentes teóricos: describe los autores específicos en los cuales se apoya el autor y los fundamentos disciplinarios y teóricos que apuntalan la investigación	Para la función de similitud este proyecto se basa en una asimetría cluster la cual fue desarrollada por Chen y Lynch. Estos autores demostraron que esta función representa mejor la asociación de conceptos que la función coseno popular.
Conceptos principales: señala el soporte técnico de la tesis, explicaciones, problemas, ideas y conclusiones planteadas para la investigación	• Dominio: Dominio es el punto de conexión entre los diversos servidores de la capa de servicios y el kernel. • Concepto: Concepto representa un pedazo de significados semánticos. Cada concepto es único y puede aparecer en muchos ámbitos y espacios. Sin embargo, no contiene ninguna información computacional. Esta información está contenida en ConceptInCS y clases ConceptInDomain. • Representación: Esta es una clase abstracta que representa la aparición de un concepto. Podría haber muchas subclases, cada de los cuales es un tipo diferente de representación. • El Espacio intermedio: como una entidad, es una colección de información relacionados entre sí espacios donde cada componente espacial codifica el conocimiento de una comunidad o un dominio tema.

	• Un espacio de información es un colección de objetos relacionados entre sí. • Un SOM es un algoritmo de red neuronal introducido por Teuvo Kohonen en 1982. Se utiliza para visualizar e interpretar grandes conjuntos de datos de alta dimensión.
Hipótesis: registra las proposiciones que sirven de guía a la investigación (conjeturas sobre el objeto de estudio) (suposiciones a partir de los cuales se organizó la investigación)	• Su principal objetivo es unificar los recursos de información distribuidos en un modelo coherente. • Está diseñado para proporcionar un mayor sistema de recuperación de información la cual sea eficiente y eficaz en comparación con las técnicas actuales. • El modelo contiene un amplio conjunto de clases y relaciones de los datos para el módulo de indexación semántica. La base de nuestra semántica la indexación se realiza mediante la creación de espacios concepto.
Tesis: describe las proposiciones que sintetizan las generalizaciones sobre el objeto de estudio ("generalizaciones empíricas" existentes en las cuales se apoyó el trabajo)	
Tipo de investigación: señala el tipo de investigación (exploratoria: examina un tema o problema poco estudiado; Descriptiva: responde al qué del fenómeno analizado; o Explicativa: responde al porqué del fenómeno analizado)	Descriptiva
Resumen y palabras claves (metodología)	
Tipo de metodología: Describe el tipo de metodología utilizada (Cualitativa: Privilegia la comprensión de los fenómenos y la interacción investigador - investigado; Cuantitativa: Privilegia la cuantificación de los fenómenos estudiados y la no interacción investigador - investigado; o Mixta: Combina procedimientos cualitativos y cuantitativos)	
Técnicas: Describe el tipo de herramientas utilizadas en la recolección, registro y sistematización de la información por el autor en la investigación	Documentos bibliográficos, algoritmo de red neuronal llamado SOM, cluster asimétrico.
Resultados:	Sus experimentos demuestran que la indexación semántica puede ser implementada en un entrono distribuido y que el rendimiento que se obtiene es significativo.
Conclusiones: Señala las conclusiones que especifica explícitamente el autor en el documento	En este trabajo, se debatió la aplicación de indexación semántica el prototipo de espacio intermedio en un entorno distribuido. Los resultados preliminares de nuestros experimentos han demostrado que la indexación semántica tiene un rendimiento significativo en este el medio ambiente. Utilizando nuestra red de estaciones de trabajo pequeños de sólo cinco máquinas, hemos sido capaces de calcular un concepto tamaño de la comunidad el espacio (aproximadamente 10 km) en cerca de tres horas. En la práctica, es que muchas colecciones que eran demasiado grandes para la indexación semántica en el pasado ahora se puede hacer con regularidad dentro de una comunidad de recursos limitados
Recomendaciones: Registra las recomendaciones que hace el autor en el documento	
Aspectos formales sobre el documento:	

Resumen del documento:	Un prototipo de investigación se presenta para la indexación semántica y recuperación en la recuperación de información. El prototipo está motivado por un deseo de proporcionar una información más eficiente y eficaz, sobre los sistemas de recuperación que contamos hoy día. Una visión general de las capas de arquitectura espacio intermedio se discute. Una modelo de objetos de apoyo a las operaciones semánticas se desarrolla. El modelo contiene un amplio conjunto de clases y relaciones de los datos para el módulo de indexación semántica. La base de nuestra semántica la indexación se realiza mediante la creación de espacios concepto. Un espacio de concepto es un índice de una colección que utiliza estadísticas de documento para capturar las relaciones entre conceptos. Es útil para impulsar la búsqueda de texto, por la sugerencia plazo de términos alternativos semánticamente relacionados con los términos de la consulta. Con los años, hemos tecnología desarrollada concepto genérico para los espacios de cómputo en grandes colecciones a través de muchos temas. Cálculos recientes sobre colecciones escala de la disciplina-se han hecho en la gama alta superordenadores. El presente trabajo describe nuestra aplicación e implicaciones de la computación en una informática distribuida el medio ambiente. Los resultados experimentales con diferentes tamaños de recogida y el número de procesos se presentan para demostrar la viabilidad de este enfoque. También mostramos que las de laboratorio y de la comunidad colecciones ya están fácilmente computable medio de un grupo de ordenadores en un laboratorio a través de un modelo de paso de mensajes. Se concluye que las agrupaciones de PC poco tiempo será capaz de calcular el índice de semántica para cualquier real colecciones.
Palabras claves:	Recuperación de información, concepto de espacio, la indexación semántica, de computación distribuida.

Tabla 7. Ficha B 7 indexación semántica

Ficha Bibliográfica	
Aspectos formales sobre el documento	
Autor : Ming Yi Chung, Qin Él, Powell y Kevin Bruce Schatz **Titulo:** SEMANTIC INDEXING FOR A COMPLETE SUBJECT DISCIPLINE (Indexación semántica para una disciplina objetiva completa) **Tipo de material:** Artículo – tesis	
Autor y titulo del documento :	
Resumen y palabras claves (Enfoque)	
Disciplina: señala el área del saber desde donde se define y aborda el objeto de estudio	Indexación Semántica y semántica escalable para grandes colecciones de documentos
Paradigma conceptual: indica la postura teórica de y/o metodológica que orienta la investigación.	La investigación se basa en el desarrollo de una tecnología de semántica escalable utilizando indexación semántica para grandes colecciones de documentos. Además, evaluar la escalabilidad de las técnicas de indexación semántica en colecciones a gran escala con el objetivo de crear bancos de pruebas.
Referentes teóricos: describe los autores específicos en los cuales se apoya el autor y los fundamentos disciplinarios y teóricos que apuntalan la investigación	
Conceptos principales: señala el soporte técnico de	• Semántica escalable es una tecnología de

la tesis, explicaciones, problemas, ideas y conclusiones planteadas para la investigación	indexación que escala grandes colecciones de documentos. • El Prototipo de espacio intermedio (CANIS) es desarrollado por un grupo de investigadores en el programa de Gestión de la Información de DARPA. Es un entorno de análisis para la indexación semántica de la información multimedia en un banco de pruebas de colecciones reales, basadas en el concepto de indexación semántica y agrupamiento semántico. • Los índices semánticos registran la correlación conceptual de las frases nominales y se computan genéricamente independientes del dominio del objeto. • El algoritmo espacio conceptual ha sido usado para generar e integrar múltiples índices semánticos.
Hipótesis: registra las proposiciones que sirven de guía a la investigación (conjeturas sobre el objeto de estudio) (suposiciones a partir de los cuales se organizó la investigación)	• Extracción de la frase nominal y análisis de ocurrencia en el desarrollo del algoritmo espacio conceptual. • Evaluar la escalabilidad de las técnicas de indexación semántica utilizando la base de datos MEDLINE • Realizar el experimento de indexación semántica MEDSPACE para una disciplina médica.
Tesis: describe las proposiciones que sintetizan las generalizaciones sobre el objeto de estudio ("generalizaciones empíricas" existentes en las cuales se apoyó el trabajo)	• Evalúan la escalabilidad de las técnicas de indexación semántica para crear un único banco de pruebas a gran escala, utilizando la base de datos Médica de la universidad de Arizona. • Están desarrollando una tecnología de semántica escalable, la cual, es una técnica estadística que permite indexar grandes colecciones de documentos para búsquedas profundas. • El prototipo de espacio intermedio (CANIS) de DARPA, utiliza indexación semántica para soportar la navegación de conceptos. • En la indexación se utilizan los algoritmos: Extracción de frase nominal que opera en tres fases (tokenización, marcado de una parte de la oración e identificación de la frase nominal) y análisis de co-ocurrencia, que se calcula basándose en una función de similitud asimétrica. • Se realiza indexado MEDLINE, en el cual se particionan los subdominios (subtemas) y se utiliza el Medical Subject Headings (MeSH) para tomar su estructura alfabética y jerárquica
Tipo de investigación: señala el tipo de investigación (exploratoria: examina un tema o problema poco estudiado; Descriptiva: responde al qué del fenómeno analizado; o Explicativa: responde al porqué del fenómeno analizado)	Explicativa.
Resumen y palabras claves (metodología)	
Tipo de metodología: Describe el tipo de metodología utilizada (Cualitativa: Privilegia la comprensión de los fenómenos y la interacción investigador – investigado; Cuantitativa: Privilegia la cuantificación de los fenómenos estudiados y la no interacción investigador – investigado; o	Metodología Mixta: combina procesos de investigación, clasificación y resultados en porcentajes

Mixta: Combina procedimientos cualitativos y cuantitativos)	
Técnicas: Describe el tipo de herramientas utilizadas en la recolección, registro y sistematización de la información por el autor en la investigación	Documentos bibliográficos, el MEDLINE (registro bibliográfico de la biblioteca nacional de medicina) El MeSH, El prototipo de espacio intermedio (CANIS)
Resultados:	
Conclusiones: Señala las conclusiones que especifica explícitamente el autor en el documento	
Recomendaciones: Registra las recomendaciones que hace el autor en el documento	
Aspectos formales sobre el documento:	
Resumen del documento:	A través del auspicio del programa de gestión de información de DARPA, están desarrollando un entorno de análisis integrado, utilizando el prototipo de espacio intermedio que usa indexación semántica para soportar navegación entre conceptos. Estos índices semánticos registran la correlación conceptual de frases nominales y son computadas genéricamente. Usando esta tecnología les permite computar índices semánticos para una disciplina objeto, en este caso, la medicina. Para ello debieron crear un MEDSPACE (espacios conceptuales) usando los registros del MEDLINE y el árbol jerárquico MeSH.
Palabras claves:	Indexación Semántica, recuperación semántica, espacio conceptual, semántica escalable, espacio intermedio, MEDSPACE, MEDLINE, informática medica

Tabla 8. Ficha B 8 indexación semántica

Ficha Bibliográfica	
Aspectos formales sobre el documento	
Autor : Doina Ana Cernea, Esther Del Moral, Emilio Labra **Titulo:** SOAF: UN SISTEMA DE INDEXADO SEMÁNTICO DE OA EN LAS ANOTACIONES COLABORATIVAS (SOAF: Semantic Indexing System Based on Collaborative Tagging) 2008 **Tipo de material:** Artículo – tesis	
Resumen y palabras claves (Enfoque)	
Disciplina: señala el area del saber desde donde se define y aborda el objeto de estudio	Indexación Semántica con anotaciones especificas en sistemas colaborativos
Paradigma conceptual: indica la postura teorica de y/o metodológica que orienta la investigación.	Buscan combinar técnicas de recuperación de información basadas en extracción automática de semántica, con anotaciones especificas realizadas por usuarios que componen la comunidad de aprendizaje.
Referentes teóricos: describe los autores específicos en los cuales se apoya el autor y los fundamentos disciplinarios y teoricos que apuntalan la investigación	
Conceptos principales: señala el soporte técnico de la tesis, explicaciones, problemas, ideas y conclusiones planteadas para la investigación	• SOAF: Semántica de Objetos de Aprendizaje basada en Folksonomias. • Sitios donde se promueve el etiquetado colaborativo: Del.icio.us, Tecnocrati, Flickr. • Folksonomia: sistemas colaborativos para la categorización no jerarquica de los recursos de internet a través de etiquetas compartidas por una comunidad o red social. • Gap Semantico: ausencia de coincidencias entre la información extraida automáticamente por ordenador y la derivada de la percepción humana de las imágenes basada en conceptos de alto nivel. Las investigaciones orientadas a minimizar el gap

	semántico refieren sus resultados a las anotaciones y al marcado semántico de recursos textuales fundamentalmente • En el caso de los recursos multimedia, se percibe un gran desajuste, llamado gap semántico, entre las características de bajo nivel extraídas de forma automática (como parámetros de color, texturas, contornos,…) y la descripción de su contenido basada en conceptos o características de alto nivel.
Hipótesis: registra las proposiciones que sirven de guía a la investigación (conjeturas sobre el objeto de estudio) (suposiciones a partir de los cuales se organizó la investigación)	• Se propone un sistema capaz de indexar automáticamente los OA de un repositorio, que combine las técnicas de extracción automática de información con tecnologías de etiquetado colaborativo. • En el caso de los recursos textuales se propone un indexado de colecciones de documentos mediante técnicas de procesamiento de lenguaje natural muy, que suponen la división del texto en una lista de términos que, más tarde, son normalizados utilizando técnicas de radicación, para eliminar las ambigüedades lingüísticas y las palabras vacías de contenido, para minimizar el grado de arbitrariedad. • Para los recursos multimedia se propone la utilización de un sistema que integre las tecnologías de la folksonomias que complementen las descripciones de los OA.
Tesis: describe las proposiciones que sintetizan las generalizaciones sobre el objeto de estudio (“generalizaciones empíricas” existentes en las cuales se apoyó el trabajo)	• Se propone una arquitectura para la extracción automática de descripción semántica de los recursos educativos multimedia basada en Latent Semantic Indexing [28] a partir de la representación de los recursos (texto o multimedia) en vectores del espacio vectorial Rn, y aplicando análisis de matrices establece las asociaciones y las conexiones entre los diversos recursos de un repositorio. • El sistema SOAF utiliza tres tipos de meta-información que pueden acompañar a los OA o a los recursos que los componen: o La semántica proporcionada de forma automática mediante indexado semántico a partir de las características de bajo nivel o Los descriptores de alto nivel proporcionados por los autores (en la mayoría de los casos se refieren a título, fecha de creación, nombre de los autores, tema, objetivos de aprendizaje,…) o La semántica aportada por los usuarios mediante anotaciones colaborativas. • Finalmente, el conjunto de etiquetas asignadas por los usuarios se pueden filtrar aplicándoles un *test* de confianza basado en la similitud de perfiles para garantizar su fiabilidad, pertinencia y validez. Así, el sistema crea y almacena los perfiles de los usuarios a partir de sus intereses y sus valoraciones subjetivas.
Tipo de investigación: señala el tipo de investigación (exploratoria: examina un tema o problema poco estudiado; Descriptiva: responde al qué del fenómeno analizado; o Explicativa: responde al porqué del fenómeno analizado)	Descriptiva:
Resumen y palabras claves (metodología)	
Tipo de metodología: Describe el tipo de metodología utilizada (Cualitativa: Privilegia la comprensión de los fenómenos y la interacción investigador - investigado; Cuantitativa:	Mixta

Privilegia la cuantificación de los fenómenos estudiados y la no interacción investigador - investigado; o Mixta: Combina procedimientos cualitativos y cuantitativos)	
Técnicas: Describe el tipo de herramientas utilizadas en la recolección, registro y sistematización de la información por el autor en la investigación	Sistemas colaborativos, repositorios de información, recursos textuales y multimedia
Resultados:	
Conclusiones: Señala las conclusiones que especifica explícitamente el autor en el documento	Este artículo propone una arquitectura SOAF que permite el indexado semántico de los OA de un repositorio combinando técnicas de extracción automática de información basado en Latent Semantic Indexing con las tecnologías de etiquetado colaborativo. El conjunto de etiquetas asignadas por los usuarios de una a de aprendizaje se filtran mediante un test de confianza a partir de la semejanza de perfiles con el objetivo de mejorar su fiabilidad, pertinencia y validez. El sistema, asimismo, recomienda una etiqueta si varios usuarios se consideran semejantes Mediante la tecnología que aquí se presenta, la meta-información de los OA obtenida incorpora el significado deducido de la práctica de las comunidades de usuarios, logrando una mejor identificación de los OA que contribuye a optimizar su reusabilidad en contextos de aprendizaje diversos.
Aspectos formales sobre el documento:	
Resumen del documento:	
Palabras claves:	Indexado semántico, *Gap* semántico, Objetos de Aprendizaje, anotaciones de los usuarios, aprendizaje colaborativo, *folksonomías*

Tabla 9. Ficha B 9 indexación semántica

1.2 FICHAS BIBLIOGRAFICAS ONTOLOGIAS

Ficha Bibliográfica	
Aspectos formales sobre el documento	
Autor : Marie Aude Aufaure, Rania Soussi, Hajer Baazaoui. **Titulo:** SIRO: ON- LINE SEMÁNTICA INFORMATION RETRIEVAL USING ONTOLOGIES (Recuperación de información semántica en línea usando ontologías) **Tipo de material:** Articulo	
Autor y titulo del documento :	
Resumen y palabras claves (Enfoque)	
Disciplina:	Ontologías de dominio, repositorios como WordNet, vector modelo.
Paradigma conceptual :	Proponer una solución a la recuperación de la información a través de ontologías ya que estas pueden ayudar al usuario a encontrar documentos de un dominio específico.
Referentes teóricos:	
Conceptos principales:	• Se recalca el gran crecimiento que ha tenido la Web, y el problema que esto presenta ya que la recuperación de información es cada vez más difícil. • Las ontologías pueden mejorar la llamada y la precisión, eliminando la ambigüedad del lenguaje natural utilizando axiomas y conceptos. • Las ontologías pueden ayudar al usuario a formular sus necesidades y tener acceso a los documentos.
Hipótesis:	• Proponer un sistema basado en ontologías que

	de alguna forma ayude a recuperar la información de una manera más eficiente. • Las ontologías son uno de los mecanismos mas utilizados para resolver el problema de la recuperación de la información. • Hacen uso de un dominio de ontología combinado con un servicio de ontología • Hacen una adaptación del vector modelo, sustituyendo los términos por conceptos. • La pregunta del usuario es enriquecida usando conceptos y relaciones del dominio ontológico y WordNet. • Este proyecto es aplicado para dominio turístico. • Este proyecto puede ser aplicado en diferentes campos.
Tesis:	• El enorme número de documentos disponibles en la web hace difícil encontrar lo que es pertinente. Por lo tanto, la búsqueda de información se vuelve más y más compleja, debido al creciente volumen de datos y de su falta de estructura. La calidad de los resultados que los tradicionales motores de búsqueda de texto proporcionan todavía no es óptima para muchos tipos de consultas de los usuarios. Especialmente, las ambigüedades de los lenguajes naturales y los conceptos abstractos son manejados inadecuadamente por los motores de búsqueda de texto. Ontologías ofrecer una solución a estos problemas. Ellas pueden ayudar a un usuario a encontrar los documentos de un dominio específico. Este documento propone un sistema de recuperación basado en ontologías. Este sistema integra los resultados de los tradicionales motores de texto completo, y por lo tanto apoya una transición gradual desde los clásicos motores de búsqueda de texto basado en las ontologías. • El propósito del proyecto ON-LINE IR es usar dos tipos de ontologías las cuales son ontologías de dominio y ontología de servicio y wordnet. La ontología de dominio que es construida manualmente es la base de la recuperación ON-LINE y contiene un dominio que es caracterizado por una lista de servicios, actividades y tareas.
Tipo de investigación:	Exploratoria
Resumen y palabras claves (metodología)	
Tipo de metodología:	•
Técnicas:	Para el proyecto hacen uso del repositorio WordNet, para lematización es usado TreeTagger, para el parseo de las páginas web utilizan DOM, la herramienta para el manejo de las ontologías usan Jena Api y Google Api para buscar en la Web. Para el vector modelo ellos se basan en el modelo propuesto por Salton.
Resultados:	
Conclusiones y recomendaciones:	Este sistema se ha realizado para aplicarlo al dominio turístico y se señala que permite mejorar la precisión de la investigación y por lo tanto la pertinencia de los documentos devueltos al usuario. Este trabajo da al usuario la oportunidad de detectar los servicios de dominio exacto. SIRO como es llamado este sistema se puede aplicar a cualquier campo, este trabajo también hace que sea posible

	proporcionar una preclasificación y filtrado de documentos para mejorar la construcción de ontologías basadas en técnicas de aprendizaje. En el futuro, este sistema se integrará a un prototipo desarrollado para la construcción semiautomática de ontologías de dominio y de servicio.
Aspectos formales sobre el documento:	
Resumen del documento:	El enorme número de documentos disponibles en la Web hace que encontrar los documentos necesarios sea una tarea difícil para usuarios. Por lo tanto, la búsqueda de información se vuelve más y más compleja, debido al creciente volumen de datos y de su falta de estructura. La calidad de los resultados de los tradicionales motores de búsqueda de texto completo no es óptima para muchos tipos de consultas de los usuarios. Especialmente, las ambigüedades de los lenguajes naturales y los conceptos abstractos son manejados inadecuadamente por los motores de búsqueda de texto completo. Las Ontologías ofrecen una solución a estos problemas. Ellas pueden ayudar a un usuario a encontrar los documentos de un dominio específico. Este proyecto propone un sistema de recuperación basado en ontologías. Este sistema integra los resultados de los tradicionales motores de texto completo, y por lo tanto apoya una transición gradual desde el clásico motor de búsqueda de texto completo a la ontología.
Palabras claves:	Ontologías, vector modelo.

Tabla 10. Ficha B 1 Ontologías

Ficha Bibliográfica	
Aspectos formales sobre el documento	
Autor : Jibran Mustafa, Sharifullah Khan, Khalid Latif. **Titulo:** ONTOLOGY BASED SEMANTIC INFORMATION RETRIEVAL (Ontología basada en la recuperación de información semántica) **Tipo de material:** Articulo	
Autor y titulo del documento :	
Resumen y palabras claves (Enfoque)	
Disciplina:	Ontologías, RDF triples.
Paradigma conceptual :	
Referentes teóricos:	
Conceptos principales:	• Ontologías de dominio se utiliza como base de conocimiento para entender los significados de los conceptos. • Una ontología es una especificación formal explícita de una conceptualización compartida • Similitud temática interpreta el significado de las palabras claves y sus relaciones, ellas utilizan una temática similar emparejando en RDF triples para concentrarse en ambos aspectos(conceptos y sus relaciones)
Hipótesis:	Para este trabajo se propone una temática similar a otros trabajos, en donde en vez de utilizar palabras claves se emplea RDF triples y se concentra en el contexto de la búsqueda del término. Para tal propósito se utiliza los siguientes componentes: crawler, source model, semantic matcher, query reformulator and ranker. El sujeto, la propiedad y el objeto de RDF triples permite al armazón de la búsqueda concentrarse en la combinación de conceptos y su relación de

	similitud al mismo tiempo. Algunos de los componentes a utilizar: **Crawler:** extrae metadatos en la forma RDF triples desde documentos residentes en un repositorio de información y los carga en un modelo fuente. **Modelo fuente**: este contiene dos componentes: • Metadato de origen: contiene metadatos contenidos en documentos de la forma RDF triples. • Metadato de contenido: ayuda a la identificación de documentos relevantes. **Semantic macher:** Re formulador de pregunta: La pregunta del usuario es reformulada utilizando sinónimos y otras relaciones, entonces la pregunta se vuelve a escribir utilizando los términos mencionados y esta pregunta se pasa al matcher semántico en la forma de RDF triples.
Tesis:	Semántica basada en las técnicas de recuperación de información comprende los significados de los conceptos que los usuarios especifiquen en sus consultas. El principal inconveniente de la semántica existente basado en técnicas de recuperación de información es que ninguno de ellos considera el contexto del concepto (s). Se propone un marco de recuperación de información semántica para mejorar la precisión de los resultados de búsqueda. En este trabajo, el enfoque de la similitud temática se emplea para la recuperación de información, a fin de captar el contexto del concepto particular (s). Ellos guardan la información de metadatos de fuente (s) en forma de triples RDF. Buscan las preguntas de los usuarios en los metadatos existentes, haciendo coincidir triples RDF en lugar de palabras clave. Los resultados de los experimentos realizados en el marco mostró mejoras en la precisión y el recuerdo en comparación con la semántica existente basado en técnicas de recuperación de información. Se propone un marco de recuperación de información semántica para mejorar la precisión de los resultados de las búsquedas y se emplea el enfoque de la similitud temática para la recuperación de información, a fin de captar el contexto del concepto particular.
Tipo de investigación:	Exploratoria
Resumen y palabras claves (metodología)	
Tipo de metodología:	•
Técnicas:	Resource Description Framework (RDF) triples.
Resultados:	Los experimentos realizados en este proyecto arrojaron excelentes resultados.
Conclusiones y recomendaciones:	Los resultados de los experimentos realizados a este proyecto han demostrado tener unos excelentes resultados ha comparación de otras técnicas utilizadas para la recuperación de la información, lo que hace que los usuarios obtengan unos resultados mas precisos en sus búsquedas. Un trabajo futuro se pretende aumentar el armazón para otros datos y tratar con otros datos heterogéneos
Aspectos formales sobre el documento:	
Resumen del documento:	
Palabras claves:	Semántica, recuperación de información, temática similar, metadatos, RDF.

Tabla 11. Ficha B 2 Ontologías

Ficha Bibliográfica	
Aspectos formales sobre el documento	
Autor : Mingxia Gao, Chunnian Liu, Furong Chen **Titulo:** AN ONTOLOGY SEARCH ENGINE BASED ON SEMANTIC ANALYSIS (Un motor de búsqueda ontológico basado en el análisis semántico) **Tipo de material:** Articulo	
Autor y titulo del documento :	
Resumen y palabras claves (Enfoque)	
Disciplina:	Ontologías, algoritmo WI-OUTOSEARCH.
Paradigma conceptual :	
Referentes teóricos:	
Conceptos principales:	• Una ontología está formada por varios términos (que se denominan conceptos) que están relacionados y limitados por diversas limitaciones estructurales. • TUCUXI captura la semántica de las páginas web a través de herramientas lingüísticas como wordnet y devuelven los resultados adecuados. • Punto de referencia ontología: Dada una Ontología, que es una norma por la que la ontología de otros se puede medir o juzgar. • Evaluación de la ontología: una ontología que debe comparar con la ontología de referencia.
Hipótesis:	• Se propone un vector de pesos y un algoritmo que analiza la entrada de los mensajes y las primeras palabras claves basadas en un conjunto de conceptos. • Se decide la influencia de conceptos establecidos en la ontología y la semántica y se crea un vector de pesos. • Se obtendrá una medida de similitud entre los mensajes de entrada y los resultados preliminares. • Desarrollo de una ontología de motor de búsqueda basado en el algoritmo WI-OUTOSEARCH.
Tesis:	Con el fin de mejorar la precisión de las búsquedas mediante el análisis semántico el documento propone conceptos - peso de vectores y algoritmo de emparejamiento (CWVMA). El algoritmo en primer lugar, analiza los mensajes de entrada y las palabras clave preliminares basados en un conjunto de conceptos de resultados. Entonces decide la regla según la influencia de los conceptos establecidos en la ontología semántica y crea vector de peso, haciendo coincidir la regla. Por fin obtiene vector resultado como fundamento de la medida de similitud entre los mensajes de entrada y los resultados preliminares. Además, se diseña y se desarrolla un motor de búsqueda de ontología basada en el algoritmo anterior WI prototipo de sistema OntoSearch. El sistema puede buscar cerca de 4 millones de páginas web de Google Web Service. Muchos de los resultados de los experimentos de explicar el algoritmo puede mejorar la precisión de búsqueda de la Ontología
Tipo de investigación:	Exploratoria
Resumen y palabras claves (metodología)	
Tipo de metodología:	•
Técnicas:	algoritmo WI-OUTOSEARCH, vector de pesos, la captura de la semántica la realiza TUCUXI.
Resultados:	
Conclusiones y recomendaciones:	Este documento propone CWVMA de acuerdo a

	factores que la literatura presenta para medir la ontología semántica. Desarrollamos WI OntoSearch que es un motor de búsqueda basado en la ontología CWVMA y hace varios experimentos por él. Un gran número de experimentos con resultados muestran el algoritmo tiene un buen efecto en la mejora de la precisión de la búsqueda de la ontología. El algoritmo con diferentes normas de congruencia, puede satisfacer diferentes necesidades. Otro trabajo interesante y atractivo es la agrupación ontologías por el algoritmo con las normas de la cartografía compleja.
Aspectos formales sobre el documento:	
Resumen del documento:	La búsqueda de información útil y la localización adecuada La ontología de la WWW o la Web Semántica es una tarea importante en el ámbito de la investigación Ontología. La diferencia entre la mayoría de los Información Ontología y la información común es que La ontología tiene una estructura semántica. Con el fin de mejorar la búsqueda precisión mediante el análisis semántico el documento propone conceptos - pesos vectores algoritmo de emparejamiento (CWVMA). El en primer lugar, el algoritmo analiza los mensajes de entrada y preliminar Palabras clave de resultados basados en conjuntos de conceptos. Entonces decide corresponden a las reglas de acuerdo a la influencia de los conceptos establecidos en la Ontología y semántica crea vector de peso, haciendo coincidir regla. Por fin obtiene vector resultado como fundamento de la medida de similitud entre los mensajes de entrada y los resultados preliminares. En Además, este documento diseñado y desarrollado una ontología motor de búsqueda basado en el algoritmo anterior - --- WI Prototipo de sistema OntoSearch. El sistema puede buscar cerca de 4 millones de páginas web de Google Web Service. Una gran cantidad de resultados de experimentos explicar el algoritmo puede mejorar la precisión de la ontología de búsqueda.
Palabras claves:	Recuperación de información, ontologías, algoritmos.

Tabla 12. Ficha B 3 Ontologías

1.3 FICHAS BIBLIOGRAFICAS RECUPERACIÓN DE INFORMACIÓN

Ficha Bibliográfica	
Aspectos formales sobre el documento	
Autor : Miguel A. Alonso, Jorge Graña, Jesús Vilares **Titulo:** RECUPERACIÓN DE INFORMACIÓN EN INTERNET **Tipo de material:** Ponencia	
Autor y titulo del documento :	
Resumen y palabras claves (Enfoque)	
Disciplina: señala el área del saber desde donde se define y aborda el objeto de estudio	Recuperación de Información en internet: Principios de la recuperación
Paradigma conceptual: indica la postura teórica de y/o metodológica que orienta la investigación.	Estudio de medidas y modelos de recuperación de información
Referentes teóricos: describe los autores específicos en los cuales se apoya el autor y los fundamentos disciplinarios y teóricos que apuntalan la investigación	En varios conceptos del documento, especialmente sobre las cadenas léxicas y la estructura de cohesión léxica, se refieren al autor Morris y sus trabajos de 1991.

Conceptos principales: señala el soporte técnico de la tesis, explicaciones, problemas, ideas y conclusiones planteadas para la investigación	• La Recuperación de información (RI, Information Retrieval), es un área de la ciencia y la tecnología que trata de la representación, almacenamiento, organización y acceso a elementos de información. • Un sistema de RI trata de determinar el grado de semejanza de cada uno de los documentos disponibles con la consulta creada por el usuario, para lo cual es necesario discernir entre documentos relevantes y no relevantes de acuerdo a la consulta. • Medidas que dependen de la ordenación de los documentos devueltos. Precisión, R-precisión, Precisión media no interpolada, Precisión media interpolada en 11 puntos. • Los modelos de recuperación de información definen la manera en que se representan las consultas y los documentos.
Hipótesis: registra las proposiciones que sirven de guía a la investigación (conjeturas sobre el objeto de estudio) (suposiciones a partir de los cuales se organizó la investigación)	• Su objetivo es presentar el enfoque de la recuperación de información de acuerdo a los modelos y sus medidas
Tesis: describe las proposiciones que sintetizan las generalizaciones sobre el objeto de estudio ("generalizaciones empíricas" existentes en las cuales se apoyó el trabajo)	• Un proceso de recuperación de información produce como salida un conjunto de documentos cuyo contenido debe satisfacer la necesidad de información del usuario. • Los sistemas realizan una indexación previa de los documentos relevantes ordenada por el grado de relevancia. A pesar de la eficiencia de muchos de estos sistemas, es difícil unificar la terminología de todo tipo de personas independientemente de su hábitat. • Existen algunas medidas de rendimiento para calificar la relevancia de los documentos mostrados al usuario, como la precisión y la cobertura. • Se emplean diferentes modelos en los sistemas de recuperación de información, con el fin de mejorar la relevancia de los documentos en las consultas. Los principales son los Booleano y vectorial. Otros también usados son los probabilísticos y basados en semántica latente.
Tipo de investigación: señala el tipo de investigación (exploratoria: examina un tema o problema poco estudiado; Descriptiva: responde al qué del fenómeno analizado; o Explicativa: responde al porqué del fenómeno analizado)	Exploratoria
Resumen y palabras claves (metodología)	
Tipo de metodología: Describe el tipo de metodología utilizada (Cualitativa: Privilegia la comprensión de los fenómenos y la interacción investigador - investigado; Cuantitativa: Privilegia la cuantificación de los fenómenos estudiados y la no interacción investigador - investigado; o Mixta: Combina procedimientos cualitativos y cuantitativos)	Cualitativa
Técnicas: Describe el tipo de herramientas utilizadas en la recolección, registro y sistematización de la información por el autor en la investigación	Estudios previos de modelos de recuperación de información y de medidas de rendimiento
Resultados:	
Conclusiones: Señala las conclusiones que especifica explícitamente el autor en el documento	En esta presentación se explora de manera un poco profunda las medidas de precisión y modelos de recuperación de información. Esto es útil para los usuarios que deseen conocer las definiciones y lo

	más relevante en el área. Es de fácil entendimiento y organización para escudriñar en el tema de la recuperación de información
Recomendaciones: Registra las recomendaciones que hace el autor en el documento	
Palabras claves:	

Tabla 13. Ficha B 1 Recuperación de información

Ficha Bibliográfica	
Aspectos formales sobre el documento	
Autor : Vadim Paz Madrid Gorelov, Ángel F. Zazo, Carlos G. Figuerola, José Luis Alonso Berrocal **Titulo:** Librerías Lucene y dotLucene para Recuperación de Información. Estudio y desarrollo de casos prácticos **Tipo de material:** informe técnico	
Autor y titulo del documento :	
Resumen y palabras claves (Enfoque)	
Disciplina: señala el área del saber desde donde se define y aborda el objeto de estudio	Tutorial básico de la utilización de la librería Lucene.
Paradigma conceptual: indica la postura teórica de y/o metodológica que orienta la investigación.	Se estudia el modelo de clases de Lucene y se explotan los principales objetos para la indexación y la búsqueda de información. Además se estudia la librería Lucene y se describe la utilización de dotLucene.
Referentes teóricos: describe los autores específicos en los cuales se apoya el autor y los fundamentos disciplinarios y teóricos que apuntalan la investigación	Greenstette: dice que el proceso de separación de palabras se compone de un montón de preguntas espinosas, de las cuales solamente unas pocas tienen respuesta perfecta.
Conceptos principales: señala el soporte técnico de la tesis, explicaciones, problemas, ideas y conclusiones planteadas para la investigación	• Término índice: es una palabra que posee significado y que se utiliza para representar un concepto. • Los sistemas de información no solo persiguen encontrar aquellos términos que mejor representen a los documentos, sino además aquellos que permiten diferenciar unos respecto a otros. Los diferentes modelos de recuperación se diferencian en el método elegido para representar los documentos y las consultas y para realizar las búsquedas de información. • Análisis léxico del texto, con el objetivo de convertir la cadena de entrada en un conjunto de palabras, y determinar el tratamiento que se realizará sobre números, signos de puntuación, tratamiento de mayúsculas y/o minúsculas, nombres propios, etc. Los separadores de palabras en español suelen ser el espacio y un conjunto más o menos reducido de signos de puntuación. • Eliminación de palabras vacías, muy frecuentes y muy poco frecuentes, con el objetivo de reducir el número de términos con valores muy poco significativos para la recuperación. Es también una buena forma de reducir el tamaño de los ficheros de datos del sistema, y así aumentar la rapidez de respuesta del sistema. • Aplicación de lematización sobre los términos resultantes para eliminar variaciones morfosintácticas y obtener lemas. La lematización consiste en elegir convencionalmente una forma de una palabra para remitir a ella todas las de su misma familia por razones de economía. • Selección de términos o grupos de términos que serán considerados términos índice.

	Normalmente se realiza sobre la naturaleza morfo-sintáctica del término, pues, como ya se ha mencionado, los términos que actúan gramaticalmente como nombres suelen poseer un mayor contenido semántico que verbos, adjetivos o adverbios.
Hipótesis: registra las proposiciones que sirven de guía a la investigación (conjeturas sobre el objeto de estudio) (suposiciones a partir de los cuales se organizó la investigación)	Los sistemas de información son la pieza clave para resolver el problema de la búsqueda rápida y eficiente de la información, por lo cual hoy en día se debe tener la capacidad de encontrar información y datos de diversos tipos y formatos de manera flexible, libre de forma que permita realizar búsquedas requiriendo el mínimo esfuerzo posible.
Tesis: describe las proposiciones que sintetizan las generalizaciones sobre el objeto de estudio ("generalizaciones empíricas" existentes en las cuales se apoyó el trabajo)	Se realiza una introducción a la apasionante disciplina que es la recuperación de información y la utilización de una de las más flexibles y poderosas librerías del marcado para crear aplicaciones de recuperación de información denominada Lucene.
Tipo de investigación: señala el tipo de investigación (exploratoria: examina un tema o problema poco estudiado; Descriptiva: responde al qué del fenómeno analizado; o Explicativa: responde al porqué del fenómeno analizado)	Descriptiva
Resumen y palabras claves (metodología)	
Tipo de metodología: Describe el tipo de metodología utilizada (Cualitativa: Privilegia la comprensión de los fenómenos y la interacción investigador - investigado; Cuantitativa: Privilegia la cuantificación de los fenómenos estudiados y la no interacción investigador - investigado; o Mixta: Combina procedimientos cualitativos y cuantitativos)	Mixta
Técnicas: Describe el tipo de herramientas utilizadas en la recolección, registro y sistematización de la información por el autor en la investigación	
Resultados:	
Conclusiones: Señala las conclusiones que especifica explícitamente el autor en el documento	Como resultado de su desarrollo se pudo concluir lo siguiente: • Tanto Lucene como dotLucene son librerías fácilmente reutilizables para aplicaciones que requieran agregar la capacidad de búsqueda en aplicaciones que contengan una fuente de datos definida y su información se obtenga como texto. • Una de las mejores ventajas de Lucene es que al ser un proyecto de código abierto y de rápida implementación, ofrece la posibilidad de agregar nuevas funcionalidades desarrollando nuevos componentes. En nuestro caso hemos agregado un analizador léxico que reconoce los patrones propios de la Lengua Española, permitiendo de esta manera la búsqueda de términos en este idioma. • Las aplicaciones creadas deben considerarse puntos de partida para el estudio y ampliación de la potencia de Lucene como herramienta de recuperación de información, fortalecida fuertemente por su flexibilidad en la obtención y asignación de parámetros de configuración iniciales, lo que las hace portables y demostrables hacia diferentes ambientes y fuentes de datos que sigan las pautas de los ejemplos mostrados en nuestro trabajo.

	• Finalmente, se resalta que Lucene define un modelo de clases compacto y de fácil comprensión, por lo que una implementación inicial completa de búsqueda e indización se puede realizar con muy pocas líneas de código y pocas instancias de objetos de Lucene. A partir de este modelo se pueden ir agregando funcionalidades basándose en la exploración del modelo completo de clases de Lucene.
Recomendaciones: Registra las recomendaciones que hace el autor en el documento	
Aspectos formales sobre el documento:	
Resumen del documento:	En este informe técnico se describe la utilización de dos librerías1 para Recuperación de Información. Después de una introducción a esta disciplina, se realiza un tutorial básico de utilización de la librería Lucene, bajo el lenguaje de programación Java, explicando en qué consiste, qué se puede hacer con ella, y poniendo ejemplo prácticos de su utilización. Se estudia el modelo de clases de Lucene, y se exploran los principales objetos para la indexación y búsqueda de información. Además del estudio e implementación de la librería Lucene, se describe la utilización de dotLucene, un puerto adicional de Lucene en .Net, con el que probar la versatilidad de Lucene en otras plataformas. Para ello se han elaborado y documentado dos ejemplos de búsqueda de información. En el primero se lleva a cabo una búsqueda de información en documentos almacenados en un árbol de directorios. Se pueden realizar búsquedas de información sobre cualquier fichero convertible a texto plano. El segundo va más allá y realiza la indexación, delimitación y búsqueda de información en documentos XML, permitiendo la búsqueda por campos concretos en este tipo de documentos.
Palabras claves:	Lucene, recuperación de información

Tabla 14. Ficha B 2 Recuperación de información

1.4 FICHAS BIBLIOGRAFICAS INDEXACION SEMANTICA Y ONTOLOGIAS

Ficha Bibliográfica	
Aspectos formales sobre el documento	
Autor : E. Desmontils y C. Jacquin **Titulo:** INDEXING A WEB SITE WITH A TERMINOLOGY ORIENTED ONTOLOGY (Indexación de un sitio web con una terminología orientada a la ontología) **Tipo de material:** Articulo	
Autor y titulo del documento :	
Resumen y palabras claves (Enfoque)	
Disciplina:	Indexación semántica,
Paradigma conceptual :	Esta investigación esta daba para solucionar uno de los problemas de la web con es la recuperación de la información.
Referentes teóricos:	
Conceptos principales:	• En los procesos de recuperación de información el problema principal es determinar el contenido específico de los documentos. • En el contexto de este proyecto la definición de ontología es la siguiente: “ontología es un

	conjunto de conceptos, donde cada uno es representado por una etiqueta, un conjunto de sinónimos de ese término y un conjunto de relaciones que conectan esos conceptos por la relación genérica especifica y la relación de composición ”
Hipótesis:	En los procesos de recuperación de información, el problema principal es determinar el contenido específico de los documentos. este proyecto propone para este tipo de problemas un proceso semiautomático, que ofrece un índice basado en el contenido de un sitio web utilizando las técnicas del lenguaje natural, este proceso no se basa en las palabras claves sino en los conceptos que representan a los términos.
Tesis:	
Tipo de investigación:	Exploratoria
Resumen y palabras claves (metodología)	
Tipo de metodología:	
Técnicas:	• Para generar todos los conceptos candidatos se utilizo el tesauro de WordNet. • Se utiliza el formato XML para el almacenamiento de las ontologías y los resultados de indexación. • Las etiquetas de voz que se dan a las páginas, se dan a través del etiquetador de Brill.
Resultados:	Se observa que la indexación semántica es una técnica muy efectiva en la solución de los problemas de recuperación de información a comparación de otras técnicas de la indexación semántica.
Conclusiones:	El proceso planteado comprende una serie de ventajas sobre los métodos de indexación tradicional e incluso sobre los métodos de anotación de un sitio web. Se estudian otras relaciones aparte de la relación genérica/especifica con el fin de mejorar el proceso de extracción de conceptos. Los resultados obtenidos de este proyecto pueden ser utilizados en diversas aplicaciones. En la actualidad se está siendo incorporado en Bonom del sistema Multi-Agente para buscar información relevante en Internet.
Recomendaciones:	
Aspectos formales sobre el documento:	
Resumen del documento:	Este artículo presenta un nuevo enfoque para indexar un sitio web. Utiliza las ontologías y las técnicas de lenguaje natural para la recuperación de información en Internet. El objetivo principal es construir un índice de estructura del sitio laWeb. Esto estructura está dada por una terminología orientada a la ontología de un dominio que se haya elegido a priori, de acuerdo con el contenido de la página web. En primer lugar, el proceso de indización utiliza las mejores técnicas de lenguaje natural para extraer bien los términos formados teniendo en cuenta los marcadores de HTML. En segundo lugar, el uso de un diccionario de sinónimos nos permite asociar conceptos candidatos a cada término. Que permite a la razón en un nivel conceptual. A continuación, para cada concepto de candidato, su capacidad para representar la página es evaluada por la determinación de su nivel de representatividad de la

	página. A continuación, se construye el índice de estructura en sí. Para cada concepto de la ontología se adjuntan las páginas del sitio Web en la que se encuentran. Por último, el ocre de los indicadores permitirá evaluar el proceso de indexación de sitio la Web por la ontología sugerido.
Palabras claves:	Recuperación de Información en Internet, páginas IndexingWeb, Ontologías, indexación semántica.
Comentarios	Este proyecto presenta enfoques que son de gran importancia para nuestro proyecto, aunque hay que tener en cuenta que el procedimiento que ellos plantean es muy general a comparación al que nosotros pretendemos hacer.

Tabla 15. Ficha B 1 Indexación Semántica y Ontologías

Ficha Bibliográfica	
Aspectos formales sobre el documento	
Autor : E. Desmontils , C. Jacquin, L.Simon **Titulo:** ONTOLOGY ENRICHMENT AND INDEXING PROCESS (Enriquecimiento de una ontología y el proceso de indexación) **Tipo de material:** Articulo	
Autor y titulo del documento :	
Resumen y palabras claves (Enfoque)	
Disciplina:	Área de la informática
Paradigma conceptual :	
Referentes teóricos:	
Conceptos principales:	Dos enfoques principales, para tener en cuenta en la semántica del documento, existe. El primero se refiere a las técnicas de anotación, enfoque basado en el uso de ontologías. Consisten en anotar manualmente los documentos utilizando ontologías. Las anotaciones se utilizan para recuperar información de los documentos. Son más bien dedicado a la solicitud de sistema de respuesta (KAON3) El segundo enfoque, para tener en cuenta el contenido del documento Web, son técnicas de recuperación de información basado en el uso de ontologías de dominio
Hipótesis:	En este artículo se presenta un método de enriquecimiento de ontología que permite tener en cuenta los conceptos que no pertenecen a la ontología. Se basa en el uso de un diccionario de sinónimos y de la heurística, a fin de añadir, a la ontología, conceptos que forman parte del diccionario de sinónimos, pero no a la ontología. Sin embargo, para condiciones que no pertenecen a la ontología, ni el tesauro, hemos desarrollado una técnica basada en web de interés. El enriquecimiento de la ontología no es completamente automático. Un experto humano tomará la decisión final de añadir o no un nuevo concepto de la ontología. Con este fin, hemos desarrollado una herramienta de visualización que ayudan a los expertos humanos para controlar el proceso de indexación y control de la desviación de la ontología potencial
Tesis:	
Tipo de investigación:	Exploratoria
Resumen y palabras claves (metodología)	
Tipo de metodología:	•
Técnicas:	•
Resultados:	

Conclusiones:	Nuestros trabajos futuros relacionados con la mejora de la heurística de enriquecimiento basado en la incorporación a la ontología de los conceptos pertenecientes a un diccionario de sinónimos. De hecho, nuestra heurística son generales. Que sería más eficaz, si se dedicaban a determinado conjunto de conceptos (en función de su profundidad en la ontología, su número hijo).También estamos trabajando en la mejora del método tema de la firma. También se están haciendo experimentos para la gestión siglas no en la fase de indexación, pero en la fase de extracción a largo plazo. Para una página web determinada, estamos tratando de determinar a qué sintagma nominal un acrónimo puede ser asociado (que es el sintagma nominal que tiene sus iniciales son las que corresponden a las siglas). Y, si la asociación no puede hacerse, se utiliza una base de siglas web.
Recomendaciones:	
Aspectos formales sobre el documento:	
Resumen del documento:	En el marco de recuperación de información Web, este trabajo presenta algunos métodos para mejorar un proceso de indexación, que utiliza la terminología de ontologías específicas orientadas a un campo de conocimiento. Así, las técnicas para enriquecer las ontologías mediante procesos de especialización, se proponen con el fin de administrar las páginas que tienen que ser indexados, pero que actualmente rechazada por el proceso de indización. Este proceso de especialización de la ontología que se haga bajo la supervisión de ofrecer a los expertos del dominio de una ayuda para decisiones relativas a su ámbito de aplicación. El enriquecimiento propuesto se basa en algunas heurísticas para gestionar la especialización de la ontología y que puede ser controlado mediante una herramienta gráfica para la validación. Categorías y descriptores de asunto: H.3.1 [Análisis de contenido y de Index]
Palabras claves:	Ontología, enriquecimiento, aprendizaje supervisado, Thesaurus, proceso de indexación, recuperación de información en la Web.

Tabla 16. Ficha B 2 Indexación Semántica y Ontologías

Ficha Bibliográfica	
Aspectos formales sobre el documento	
Autor : Shahrul Azman Noah, Lailatulqadri Zakaria, Arifah Che Alhadi, Tengku Mohd Tengku Sembok, Saidah Saad **Titulo:** TOWARDS BUILDING SEMANTIC RICH MODEL FOR WEB DOCUMENTS USING DOMAIN ONTOLOGY (Hacia la construcción de un modelo rico en semántica para documentos web usando ontologías de dominio) **Tipo de material:** Articulo	
Autor y titulo del documento :	
Resumen y palabras claves (Enfoque)	
Disciplina:	Área de la informática
Paradigma conceptual :	
Referentes teóricos:	
Conceptos principales:	
Hipótesis:	Se desarrolla una herramienta para extraer la información semántica contenida en los documentos web, mediante el análisis del lenguaje natural y la ontología de dominio especifico. La idea es construir una base de datos especifica que contenga los índices semánticos de los documentos. Esto está destinado a servir e integrar documentos de

	sistemas de recuperación semántica dentro de un grupo o comunidades.
Tesis:	
Tipo de investigación:	Exploratoria
Resumen y palabras claves (metodología)	
Tipo de metodología:	
Técnicas:	
Resultados:	
Conclusiones:	La ontología de dominio desempeña un papel importante en el apoyo a las tareas de clasificación de documentos y la organización de estos mismos. En este trabajo se presento una ontología de dominio combinado con una técnica de análisis del lenguaje natural, que puede ser explotado no solo para extraer los conceptos más importantes de los documentos sino también para construir el contenido semántico de los documentos web.
Recomendaciones:	
Aspectos formales sobre el documento:	
Resumen del documento:	El acceso y la extracción de significados semánticos de documentos web es crucial para la realización de la Web Semántica. Si bien la web ofrece la flexibilidad de hacer la información fácilmente disponible, es bastante difícil encontrar un camino de describir, clasificar y presentar esta información con contenido semántico ricos. Por lo tanto, la semántica contenido de la información de los documentos web deben ser determinados con el fin de hacer la información más accesible enredado a los motores de búsqueda y otras aplicaciones. En este trabajo se propone un enfoque destinado a ayudar en la construcción de modelos de documentos semánticos con técnica de análisis de lenguaje natural y un dominio específico de la ontología. .
Palabras claves:	

Tabla 17. Ficha B 3 Indexación Semántica y Ontologías

Ficha Bibliográfica	
Aspectos formales sobre el documento	
Autor : Samaneh CHAGHERI, Catherine ROUSSEY, Sylvie CALABRETTO, Cyril DUMOULIN **Titulo:** SEMANTIC INDEXING OF TECHNICAL DOCUMENTATION (Indexación semántica de documentación técnica) **Tipo de material:** Articulo	
Autor y titulo del documento :	
Resumen y palabras claves (Enfoque)	
Disciplina:	Documentación técnica, ontologías, estructura lógica e indexación semántica.
Paradigma conceptual :	El crecimiento continuo de documentos estructurados almacenados en las empresas ha causado diferentes esfuerzos en el desarrollo de sistemas de recuperación basados en la estructura del documento.
Conceptos principales:	La estructura lógica describe el papel de cada elemento de un documento. Cada tipo de elemento corresponde a una unidad lógica en el documento, como el título o capítulo. Todos estos elementos lógicos están organizados como un árbol para representar la relación de inclusión entre elementos.
Hipótesis:	En este artículo se presenta un nuevo modelo de indexación semántica que ayude a mejorar el

	proceso de indexación y la recuperación de información.
Tesis:	Se realiza un modelo de indexación semántica que explota tanto las estructuras lógicas y el contenido semántico de los documentos.
Tipo de investigación:	Exploratoria
Resumen y palabras claves (metodología)	
Tipo de metodología:	
Técnicas:	Indexación semántica, recuperación de información.
Resultados:	
Conclusiones y recomendaciones:	Se recomienda la utilización de un recurso semántico como WordNet como modelo de la semántica del contenido del documento
Aspectos formales sobre el documento:	
Resumen del documento:	En consecuencia, es necesario organizar los documentos con el fin de recuperar rápidamente información crítica. La gestión de este creciente volumen de documentos que requiere la clasificación de documentos que se basa en técnicas de indexación. Así pues, cuánto pertinentes de la fase de indexación es, más relevante será la clasificación. La documentación técnica es, por naturaleza muy estructurada. Por ejemplo, la estructura lógica describe el papel y la naturaleza de los elementos del documento (introducción, título, sección, y así uno …) y los vínculos lógicos entre ellos (Un capítulo se compone de una parte y así). Esta estructura facilita la presentación de documentos y mejora la precisión de indexación. La recuperación de los sistemas de información clásica no utiliza la estructura lógica, ni el concepto que figura en el contenido textual de los documentos. El documento semántico es descrito por los conceptos que pertenecen a un recurso semántico. En este contexto, proponemos un nuevo modelo de indexación semántica que explota tanto las estructuras lógicas y el contenido semántico de de los documentos.
Palabras claves:	Documentación técnica, estructura lógica, estructura semántica, indexación semántica, la ontología

Tabla 18. Ficha B 4 Indexación Semántica y Ontologías

Ficha Bibliográfica	
Aspectos formales sobre el documento	
Autor : Song Jun-feng **Titulo:** ONTOLOGY-BASED INFORMATION RETRIEVAL MODEL FOR THE SEMANTIC WEB (ontología basada en modelo de recuperación de información para la web semántica) **Tipo de material:** Artículo – tesis	
Autor y titulo del documento :	
Resumen y palabras claves (Enfoque)	
Disciplina: señala el área del saber desde donde se define y aborda el objeto de estudio	Indexación Semántica y ontologías de dominio
Paradigma conceptual: indica la postura teórica de y/o metodológica que orienta la investigación.	Se enfocan en la generación de una ontología teniendo en cuenta la traducción e integración de ontologías de dominio. Además se marca el contenido semántico por medio de un razonador lógico de descripción
Referentes teóricos: describe los autores específicos en los cuales se apoya el autor y los fundamentos disciplinarios y teóricos que apuntalan la investigación	
Conceptos principales: señala el soporte técnico de la tesis, explicaciones, problemas, ideas y conclusiones planteadas para la investigación	• Construcción de web semántica usando términos definidos en ontologías. • Los términos de la ontología son usados

	como metadatos para marcar el contenido de la web. • Índices de términos semánticos. • Traducción e integración de dominios de ontologías • Conceptualización del dominio • Ontología del dominio
Hipótesis: registra las proposiciones que sirven de guía a la investigación (conjeturas sobre el objeto de estudio) (suposiciones a partir de los cuales se organizó la investigación)	• La recuperación pretende encontrar todos los documentos relevantes que satisfagan las necesidades de información de los usuarios de una colección de documentos. • El método básico para la construcción de la web semántica es usar los términos definidos en ontología como metadatos. • El lenguaje OWL Lite permite generar una ontología a partir de la traducción e integración de la ontología de dominios.
Tesis: describe las proposiciones que sintetizan las generalizaciones sobre el objeto de estudio ("generalizaciones empíricas" existentes en las cuales se apoyó el trabajo)	• Se define "dominio" como una parte del mundo acerca del cual queremos expresar algún conocimiento. La "Conceptualización del dominio" se utiliza para extraer un conjunto de términos y de conocimiento del dominio. "Ontología del dominio" es la especificación explicita del de la conceptualización del dominio. • Después de generar la ontología, los términos definidos en ella son usados como metadatos para marcar el contenido de la web. Los índices de términos semánticos son identificados a través de la web. • La ontología basada en la recuperación de información cuenta con la semántica de los términos mientras que la recuperación de información basada en el modelo vector, depende de la sintaxis de los términos.
Tipo de investigación: señala el tipo de investigación (exploratoria: examina un tema o problema poco estudiado; Descriptiva: responde al qué del fenómeno analizado; o Explicativa: responde al porqué del fenómeno analizado)	Descriptiva
Resumen y palabras claves (metodología)	
Tipo de metodología: Describe el tipo de metodología utilizada (Cualitativa: Privilegia la comprensión de los fenómenos y la interacción investigador – investigado; Cuantitativa: Privilegia la cuantificación de los fenómenos estudiados y la no interacción investigador – investigado; o Mixta: Combina procedimientos cualitativos y cuantitativos)	Mixta: combina la descripción teórica y practica mostrando el modelo de generación de la ontología.
Técnicas: Describe el tipo de herramientas utilizadas en la recolección, registro y sistematización de la información por el autor en la investigación	Documentos bibliográficos, varias ontologías de dominio, un razonador lógico de descripción. Lenguaje OWL Lite para la ontología
Resultados:	
Conclusiones: Señala las conclusiones que especifica explícitamente el autor en el documento	La recuperación de información en la web tiene varios adeptos que buscan la manera de optimizarla. En la ontología basada en el modelo de recuperación de información, cuenta con la indexación de términos semánticos, no con la sintaxis de índice de términos. Así, las vistas lógicas de documentos y las necesidades de información de los usuarios pueden representar documentos y el rendimiento es mejor comparado con el modelo vector.

	Para realizar la ontología basada en el recuperación de información para la web semántica, se necesitaron marcar el contenido de la web con términos definidos en la ontología.
Recomendaciones: Registra las recomendaciones que hace el autor en el documento	
Aspectos formales sobre el documento:	
Resumen del documento:	
Palabras claves:	Ontologías de dominio, razonador lógico, términos de índices semánticos.

Tabla 19. Ficha B 5 Indexación Semántica y Ontologías

Ficha Bibliográfica	
Aspectos formales sobre el documento	
Autor : Fatiha Boubekeur Mohand Boughanem and Lynda Tamine-Lechani **Titulo:** SEMANTIC INFORMATION RETRIEVAL BASED ON CP-NETS (Recuperación de información semántica basada en CP-Nets) **Tipo de material:** Artículo - tesis	
Autor y titulo del documento :	
Resumen y palabras claves (Enfoque)	
Disciplina: señala el área del saber desde donde se define y aborda el objeto de estudio	Indexación semántica,
Paradigma conceptual: indica la postura teórica de y/o metodológica que orienta la investigación.	Se centran en la indexación de documentos y evaluación de consultas usando CP-Nets (Conditional Preferences Networks) para la recuperación de información.
Referentes teóricos: describe los autores específicos en los cuales se apoya el autor y los fundamentos disciplinarios y teóricos que apuntalan la investigación	
Conceptos principales: señala el soporte técnico de la tesis, explicaciones, problemas, ideas y conclusiones planteadas para la investigación	• Amplio vocabulario para expresar los mismos conceptos • Enfoque conceptual de indexación de documentos • Formulación de consultas e indexación de documentos • Identificar conceptos y reglas de asociaciones para descubrir relaciones conceptuales entre conceptos. • Documentos relevantes no son recuperados siempre. • Indexación semántica basada en técnicas de desambiguación contextual de sentidos de las palabras. Asociar las palabras extraídas de documentos o consultas a las palabras de sus propios contextos. • Indexación conceptual está basada en el uso de conceptos extraídos de ontologías y taxonomías.
Hipótesis: registra las proposiciones que sirven de guía a la investigación (conjeturas sobre el objeto de estudio) (suposiciones a partir de los cuales se organizó la investigación)	• El amplio vocabulario para expresar los mismos conceptos en varios dominios, genera un problema de ambigüedad en la recuperación de información. • Los documentos relevantes no son recuperados si no comparten términos con la consulta. • Términos irrelevantes que tienen palabras comunes son recuperados incluso si no semánticamente equivalentes. • Se debe generar todas las asociaciones de significado entre los ítems de una base de datos.

Tesis: describe las proposiciones que sintetizan las generalizaciones sobre el objeto de estudio ("generalizaciones empíricas" existentes en las cuales se apoyó el trabajo)	• La indexación semántica es basada en técnicas para la eliminación de ambigüedad contextual en el sentido de una palabra. Es decir, consiste en asociar las palabras extraídas de un documento o consulta en palabras en su propio contexto. Otro enfoque elaborado usa las representaciones jerárquicas derivadas de ontologías para computarizar la distancia semántica o similitud semántica entre palabras que para ser comparadas. • La indexación conceptual es basada en el uso de conceptos extraídos de ontologías y taxonomías de acuerdo a documentos indexados en vez de usar palabras simples. El proceso de indexación esta dado por: primero identificar los términos multi-palabra en el texto del documento y luego hacer concordar los términos con conceptos en la ontología. • En el enfoque de indexación basado en CP-Nets, las relaciones entre nodos expresa las dependencias conceptuales entre los conceptos relacionados. Es basado en el uso de una ontología general WordNet como recurso para extraer conceptos representativos de documentos y la técnica de reglas de asociación para descubrir los conceptos latentes de relaciones conceptuales.
Tipo de investigación: señala el tipo de investigación (exploratoria: examina un tema o problema poco estudiado; Descriptiva: responde al qué del fenómeno analizado; o Explicativa: responde al porqué del fenómeno analizado)	
Resumen y palabras claves (metodología)	
Tipo de metodología: Describe el tipo de metodología utilizada (Cualitativa: Privilegia la comprensión de los fenómenos y la interacción investigador - investigado; Cuantitativa: Privilegia la cuantificación de los fenómenos estudiados y la no interacción investigador - investigado; o Mixta: Combina procedimientos cualitativos y cuantitativos)	
Técnicas: Describe el tipo de herramientas utilizadas en la recolección, registro y sistematización de la información por el autor en la investigación	
Resultados:	
Conclusiones: Señala las conclusiones que especifica explícitamente el autor en el documento	La recuperación de información basada en CP-Nets, proporciona un enfoque flexible que permite una consistencia en la indexación conceptual basada en CP-Nets y el uso de una ontología para identificación, ponderación y eliminación de ambigüedad en los términos del dominio. Además se crean reglas de asociación para derivar relaciones dependientes de contexto entre términos importantes para mayor expresividad en la representación de documentos. Este enfoque evalúa la relevancia de un documento respecto a una consulta dada teniendo en cuenta básicamente un grafo simple.

Tabla 20. Ficha B 6 Indexación Semántica y Ontologías

2 ANEXO B

2.1 SURVEY SOBRE CREACION DE INDICES SEMANTICOS

SURVEY SOBRE CREACION DE ÍNDICES SEMÁNTICOS

MIGUEL ÁNGEL NIÑO ZAMBRANO
INGENIERO DE SISTEMAS, MAGISTER EN INFORMÁTICA
DEPARTAMENTO DE SISTEMAS, FACULTAD DE INGENIERÍA ELECTRÓNICA Y TELECOMUNICACIONES
PROFESOR ASOCIADO, UNIVERSIDAD DEL CAUCA
manzamb@unicauca.edu.co
POPAYÁN, CAUCA, COLOMBIA

DICNORY JIMENA PEREZ URBANO
ESTUDIANTE DE INGENIERÍA DE SISTEMAS
DEPARTAMENTO DE SISTEMAS, FACULTAD DE INGENIERÍA ELECTRÓNICA Y TELECOMUNICACIONES
UNIVERSIDAD DEL CAUCA
dicnory@unicauca.edu.co
POPAYÁN, CAUCA, COLOMBIA

DIAN MARIBEL PEZO ARTEAGA
ESTUDIANTE DE INGENIERÍA DE SISTEMAS
DEPARTAMENTO DE SISTEMAS, FACULTAD DE INGENIERÍA ELECTRÓNICA Y TELECOMUNICACIONES
UNIVERSIDAD DEL CAUCA
dpezo@unicauca.edu.co
POPAYÁN, CAUCA, COLOMBIA

RESUMEN

El presente Survey expone una recopilación de las principales investigaciones existentes alrededor de la recuperación de la información (RI) y la indexación semántica (IS). Se enfoca principalmente en la creación y uso de técnicas como Ontologías e Índices semánticos para la recuperación de información. A través de este análisis se obtiene una visión general y específica de los modelos, algoritmos y herramientas utilizadas en la indexación semántica y proporciona una fuente importante en el uso de la misma por parte de los autores que deseen aplicarla. Gracias a sus resultados relevantes en la recuperación de información, la indexación semántica se utiliza como complemento de varias herramientas de búsqueda y se sustenta en la investigación de varios autores que se referenciarán en este contenido.

PALABRAS CLAVE: Indexación Semántica, Ontologías, tesauros, modelos de indexación semántica, herramientas de indexación.

ABSTRACT

This Survey presents a compilation of the main existing research around the information retrieval (IR) and semantic indexing (IS). It focuses primarily on the creation and use of techniques such as ontologies and semantic indexes for information retrieval. Through this analysis gives an overview and specific of the models, algorithms and tools used in semantic indexing and provides an important source in the use of it by authors who wish to apply. Thanks to its outstanding results in information retrieval, semantic indexing is used like complement of several search tools and is based on the research paper of several authors which will be referenced in this content.

KEYWORDS: Semantic Indexing, Ontology, thesauri, semantic indexing model, indexing tools.

1. INTRODUCCIÓN

Desde hace una década, varios proyectos [1][2][3] han propuesto diversas soluciones que mejoran la relevancia [4][5] de la información en la búsquedas Web, desarrollando o mejorando las técnicas actuales de recuperación de información. Una de las mejoras es la utilización de nuevas metodologías enmarcadas bajo el nombre de búsqueda semántica. Cada uno de estos proyectos utiliza diferentes técnicas con las cuales han obtenido resultados muy favorables, en especial las técnicas que aplican las ontologías y la construcción de índices semánticos; al parecer, éstos últimos han empezado a utilizarse en variados estudios [6][7][8][9], en dónde la semántica de los conceptos es el principal problema a resolver. Así, cada estudio plantea su propia forma de implementar índices semánticos, aplicando diferentes criterios, procesos y pasos, que dependen de los autores y sus objetivos. A pesar de la importancia que tienen los índices semánticos en los sistemas de recuperación de la información, los investigadores deben recurrir a un proceso largo de sensibilización y entendimiento en su construcción y uso, dificultando así las nuevas investigaciones en el área en particular.

El presente trabajo expone un survey en el área de la creación de índices semánticos, con el fin de crear una base teórica que permita construir un procedimiento general para generar índices semánticos en el entorno de la recuperación de la información Web. En el análisis del estado del arte realizado, no se encontró un estudio que plantee un procedimiento general para construir índices semánticos basados en ontologías de dominio, que oriente a los investigadores al momento de su construcción.

Por lo anterior, se analizaron diversos proyectos con la perspectiva de permitir la abstracción de un procedimiento para generar índices semánticos basados en ontologías de dominio, el cual sirve como soporte para el desarrollo de aplicaciones que tengan por objetivo mejorar la relevancia de los resultados obtenidos en la búsqueda Web.

Inicialmente el usuario encontrará las bases conceptuales sobre los índices semánticos, sus herramientas, algoritmos y los proyectos que los han implementado. Luego, se analizan los diferentes proyectos que se han realizado con índices semánticos y ontologías para el mejoramiento de la recuperación de información. Finalmente se realiza una comparación de los pasos y procedimiento realizados por las investigaciones estudiadas.

2. PLANTEAMIENTO DEL PROBLEMA

2.1 DEFINICIONES

El término indexar significa registrar ordenadamente datos e informaciones para elaborar su índice (según el diccionario de la Real Academia Española) y así referenciarlo de manera más rápida y eficiente. Se realizan varias técnicas para llevar a cabo la indexación de los documentos en la Web; la más simple es la indexación automática basada en el número de veces que se encuentra una palabra en un documento [10] y de esta manera, determinar la relevancia de ese documento para las búsquedas que contengan la palabra específica.

La indexación semántica va mas allá de buscar la ocurrencia de una palabra en los documentos, se enfoca también en asociar los conceptos con los términos o palabras en las páginas Web. Con ello se busca encontrar patrones en los datos no estructurados (documentos sin descriptores, como palabras clave o etiquetas especiales) [11] y usar los patrones de búsqueda en una mejor clasificación de los datos y precisión en la recuperación de información.

En otras palabras, el uso de índices semánticos en la búsqueda web, significa que los objetos son indexados no solo por los términos empleados, sino también por los conceptos que contienen para representarlos.

2.2 CARACTERÍSTICAS DE LOS ÍNDICES SEMÁNTICOS

Según el enfoque dado por Suarez B. Marco [10], un índice semántico tiene las siguientes características:

- Un índice semántico es multidimensional, ya que cualquier combinación de propiedades son moldeados en un concepto que definen como el número total de términos que aparecen en un documento, el cual, puede servir como un elemento de indexación.
- Los elementos de indexación son valores de atributos que pueden estar basados en complejas descripciones de objetos relacionados, como un concepto estructurado.
- Un índice semántico es altamente adaptable a las necesidades de cada proyecto. Los conceptos de indexación pueden ser añadidos o eliminados como se desee, lo cual los hace muy densos y precisos con respecto al interés de un grupo de personas.
- Dado que el índice es en realidad un conjunto de descripciones parciales de los objetos indexados, mucha información se puede extraer del índice solo, sin tener acceso a las descripciones individuales de todo.

Al realizar una indexación semántica, se utiliza la información del concepto que está dentro de los objetos indexados para mejorar la relevancia en la recuperación de información. Así, por ejemplo, la consulta "Paris Hilton", se asocia con una mujer, en vez de relacionarse con una ciudad y un hotel por separado, de esta manera se brinda al usuario un resultado satisfactorio de las consultas que realizas.

Para obtener la información anterior, se puede hacer uso de la notación semántica de los recursos Web, pero lamentablemente gran cantidad de los recursos en Internet no tienen éstos metadatos y los sistemas de recuperación de la información deben implementar estrategias como la creación de índices semánticos. Normalmente éstos utilizan la representación jerárquica derivada de las ontologías para calcular la distancia entre conceptos o similitud semántica entre las palabras que deben compararse [10].

2.3 HERRAMIENTAS

En la indexación semántica se puede utilizar algunas herramientas que proporcionan cierto grado de facilidad a la hora de crear los índices semánticos. A continuación se presentan las herramientas más importantes que han aportado a la creación uso de índices semánticos.

En la Tabla 21 se muestra un cuadro comparativo extraído de la investigación de C. Cobos [12] de las herramientas más utilizadas. En esta tabla comparativa se observa una gran ventaja de la herramienta Lucene respecto a las otras, debido a la indexación incremental y a la búsqueda realizada por cualquier campo que el usuario decida. Además es una herramienta multiplataforma.

Tabla 21: Comparación de herramientas en indexación semántica.

	LUCENE	TERRIER	XAPIAN	LEMUR
Multiplataforma	Si	No	Si	Si
Lenguaje de implementación	java	Java	c++	c++
Soporte para otros lenguajes	perl, phyton, **c#**, ruby y c++	No	perl, python, php, tcl, c# y ruby	java y c#
Archivos que indexa	pdf, word, html, htm, txt, xml, rtf, entre otros	html , pdf, word, xls, ppt, txt	pdf, word, html, htm, txt, xml, rtf, entre otros	pdf, word, html, ppt, txt, xml, rtf, entre otros
Stemming para varios idiomas	Si	Si	Si	Si
Búsqueda mientras actualiza índice	Si	No	No	No
Indexación incremental	Si	No	No	Si
Modelo de representación	Espacio vectorial	Probabilístico	Probabilístico	Probabilístico
Búsquedas por cualquier campo	Si	No	No	No
Tecnologías enSource	Si	Si	Si	Si
Ultima actualización	Versión: Lucene.Net 2.3.2 Fecha: 24/07/2009	Versión: Terrier 2.2.1 Fecha: 29/01/2009	Versión: Xapian 1.0.16 Fecha: 10/09/2009	Versión: Lemur 4.10.1 Fecha: 28/07/2009

2.3.1 El Prototipo de Espacio Intermedio

El Prototipo de espacio intermedio (Interspace Prototype) [13] es desarrollado por un grupo de investigadores en el programa de Gestión de la Información de DARPA. Es un entorno de análisis para la indexación semántica de la

información multimedia en un banco de pruebas de colecciones reales, basadas en el concepto de indexación semántica y agrupamiento semántico para navegar entre conceptos [14].

2.3.2 InfoReuser: Motor de indización y búsqueda semántico

La herramienta infoREUSER [15] es un módulo de indexación y recuperación semántica de contenidos ofimáticos (Semántico). Permite la recuperación de archivos de texto, ya sean documentos de texto plano, RTF, MS Word, PDF, HTML, MS Power Point, o MS Excel. Algunas de sus características son:

- **Lematización**: Esta característica permite recuperar documentos que contengan términos escritos en singular o plural, o con verbos conjugados.
- **Propagación por ontologías**: Esto permite que a los términos del usuario, se unan otros que el sistema entienda similares. Esta tarea se aplica a sustantivos, adjetivos y a verbos.
- **Indexación basada en relaciones**: Esto hace que el elemento clave de la indexación y la búsqueda no sean las palabras clave, sino cómo éstas se relacionan unas con otras.

2.4 ALGORITMOS

A continuación se describen algunos de los algoritmos más utilizados en los procesos de indexación semántica.

2.4.1 Algoritmo Espacio Conceptual

El algoritmo Espacio Conceptual [14] ha sido utilizado para generar e integrar múltiples índices semánticos y está basado en correlaciones estadísticas del contexto dentro de los documentos. En el proceso de este algoritmo se pueden llevar a cabo dos procesos.

- **Extracción de frase nominal**: se encarga de extraer las frases cuyo núcleo es un sustantivo y se realiza en tres fases: tokenización, etiquetado "Part-Of-Speech", que comprende el análisis léxico y el contextual; e identificación de la frase nominal.
- **Análisis de co-ocurrencia[16]**: permite calcular la información de frecuencia en que aparece la frase nominal, la cual es usada para calcular pesos para cada frase nominal en los documentos y con ellos crear el índice. Es calculado basándose en una función de similitud asimétrica.

2.4.2 Algoritmo Desambiguación del Sentido (concepto) de las Palabras WSD

El algoritmo de desambiguación1 o en ingles WSD (Word Sense Disambiguation), permite realizar una desambiguación semi-completa pero precisa, de las palabras que se reciben en una consulta realizada por el usuario [2]. Se usa en las fases de indexación y búsqueda. En el primer caso, este algoritmo permite la desambiguación de las palabras del cuerpo y en el segundo, resuelve las ambigüedades en las palabras de consulta [17].

La Desambiguación del Sentido (concepto) de las Palabras, puede expresarse en un conjunto más amplio de técnicas llamado Procesamiento de Lenguaje Natural (PLN), que básicamente "trata los fenómenos lingüísticos de diversa índole de forma automatizada mediante computadores" [18].

WSD es una fase necesaria para la consecución de acciones como el análisis sintáctico o la interpretación semántica en tareas del PLN, así como para el desarrollo de aplicaciones finales, tanto de recuperación de información, como de clasificación de textos, análisis de discurso, traducción automática o análisis gramatical, entre otras [18].

2.5 PROYECTOS CON INDEXACION SEMANTICA

Desde hace más 10 años se ha investigado y construido índices semánticos para la recuperación de información, a partir de técnicas, algoritmos y herramientas que

[1] Trata los fenómenos lingüísticos de diversa índole de forma automatizada. Elimina la ambigüedad en las palabras, que surge cuando una estructura gramatical puede ser interpretada de varias maneras y por tanto, puede confundir en el sentido de la oración. Extraído de: http://www.nosolousabilidad.com/articulos/desambiguacion.htm.

permiten un buen manejo de índices basados en la semántica de los documentos.

Los proyectos que han utilizado indexación semántica, se pueden clasificar así:

- **Por algoritmos**: Los que han utilizado en su proceso principal algoritmos como el de desambiguación, y de espacio conceptual.
- **Por modelos**: Se encuentran los que han utilizado el modelo vector y el de creación del espacio conceptual.
- **Por utilización de herramientas**: Corresponde a los proyectos que utilizan herramientas ya desarrolladas de indexación como Lucene y Lémur, entre otras.

2.5.1 Clasificación de Proyectos por Algoritmos

En el área de la recuperación de información, existen varios proyectos que han logrado una mejora en la relevancia de los documentos recuperados, a través de la implementación de algoritmos específicos. Los más destacados, son los algoritmos de desambiguación, principalmente el WSD (Word Sense Desambiguation) y el algoritmo de espacio conceptual. A continuación se describen algunos proyectos que los utilizan.

El proyecto de indexación semántica ***Semantic Indexing Using Wordnet Senses*** **[2]**, se basa en la implementación de un prototipo que combine la indexación basada en palabras y basada en sentidos o conceptos, utilizando WordNet. El proyecto establece tres etapas que comprenden:

1. **El módulo WSD:** en el que cada palabra es reemplazada con un nuevo formato como el siguiente: "Pos|Stem|POS|O.f.f set". La información obtenida del modulo WSD es usada para el principal proceso de indexación donde la palabra raíz y la ubicación están indexados junto al Synset (conjunto de sinónimos) de WordNet (si existe).
2. **Módulo de indexación.** Indexa los documentos y luego son procesados por el módulo WSD. El segmento Stem y, por separado, el Offset|POS son adicionados al índice. El proceso de indexación toma un grupo de archivos de documentos y produce un nuevo índice.
3. **Módulo de recuperación**. Rescata documentos basados en una consulta de entrada.

El algoritmo utilizado [19] comprende 10 pasos en los cuales se aplican los procedimientos ya definidos:

- Paso1: Pre-procesar el texto con etiquetas utilizando el etiquetado de "part-of-speech" [20].
- Paso2: Inicializar el conjunto de palabras desambiguadas (SDW), con las palabras de entrada (Ambiguas, SAW).
- Paso 3: Identificar nombres propios en el texto (procedimiento 1).
- Paso4: Identificar las palabras que solo tienen un sentido en WordNet (procedimiento 2).
- Paso 5: desambiguar palabras basadas en su ocurrencia (procedimiento 3).
- Paso 6: Identificar el conjunto de nombres (procedimiento 4).
- Paso 7: Identificar sinónimos en SAW y en SDW (procedimiento 5).
- Paso 8: sinonimia entre palabras de SAW (procedimiento 6).
- Paso 9: Identificar palabras de SAW con distancia máxima de 1 respecto a las palabras SDW (procedimiento 7).
- Paso 10: distancia máxima entre palabras de SAW (procedimiento 8).

Una vez terminados los procesos descritos anteriormente, crean un Benchmark con 52 textos para probar el método de desambiguación.

En sus resultados se observa la eficiencia del método WSD para dominios abiertos, lo cual demuestra que el algoritmo aplicado es adecuado en tareas de indexación semántica. Esta indexación ofrece una mejora en las técnicas actuales de recuperación de información.

La investigación realizada en ***Semantic Indexing For A Complete Subject Discipline*** [14], permitió el desarrollo de una técnica

estadística que pertenece a la semántica escalable, la cual indexa grandes colecciones para búsquedas profundas. Cabe agregar, que el presente proyecto también hace parte de la clasificación por Modelos en "Espacio Conceptual", puesto que construyen un prototipo de espacio conceptual. Para llevar a cabo este experimento, utilizaron los registros bibliográficos de la Biblioteca Nacional de Medicina (NLM) de Estados Unidos.

En el proceso de indexación se utiliza el **algoritmo espacio conceptual** adoptado en varios estudios [21][22][23] y usado para generar e integrar múltiples índices semánticos. Se realizan etapas intermedias en el proceso, las cuales se describen a continuación.

1. **Extracción de la frase nominal:** Se utilizaron algunas reglas de identificación de frases nominales y fraseo para el desarrollo de un extractor de frases. El fraseo opera en tres fases: tokenización, marcado de "Part of Speech" (parte de la palabra) basándose en el etiquetador de *Brill* [24] e Identificación de la frase nominal con AZ Noun Phraser [25]. Se guarda la frecuencia de la frase nominal para calcular sus pesos en los documentos.
2. **Análisis de co-ocurrencia**: Este análisis se calcula basándose en una función de similitud asimétrica. Su resultado es una matriz que representa una red de frases nominales y sus probables relaciones.

Para indexar MEDLINE realizaron dos pasos:

1. Dividir en segmentos los subdominios para navegar por ellos.
2. Utilizar la ordenación de Medical Subject Headings (MeSH) [26], que consiste en una estructura alfabética y jerárquica llamada MeSH. Esta tiene las propiedades de un tesauro y un sistema de clasificación.

Con la realización del proceso construyeron MEDSPACE: un experimento de indexación semántica para una disciplina médica. Para esto, se escogieron conjuntos de frases desde MeSH que describen bien los artículos o documentos y luego se creó el Prototipo Interspace [34]. Al finalizar el trabajo comenzaron a evaluar la utilidad de los sistemas vs. las necesidades de información.

2.5.2 Clasificación de Proyectos por Modelos

2.5.2.1 Espacio Vectorial

Otra forma de proporcionar más eficiencia en las búsquedas Web, es basándose en los modelos actuales de recuperación de información con algunas modificaciones en dichos modelos y arquitecturas. A continuación se describen algunos proyectos que lo usan.

El proyecto ***A Novel Approach to Semantic Indexing Based on Concept*** [27], describe el método de indexación basado en un "Concept Vector Space", es decir, el espacio vectorial de conceptos, a través del cual se representa el contenido semántico de un documento.

Para la extracción de conceptos utilizaron cadenas léxicas con los vectores de concepto y vectores de texto, así se calculan los índices semánticos y su grado de importancia semántica.

El sistema propuesto tiene cuatro componentes:

- Construcción de cadenas léxicas [28].
- Ponderación de cadenas y nombres.
- Reponderación del término basada en el concepto.
- Extracción del índice del término semántico.

En los dos primeros se discriminan las cadenas representativas de las cadenas léxicas. Las cadenas representativas son cadenas delegadas para representar un concepto característico de un documento. Se asume además que los conceptos son independientes entre sí, sin considerar su similitud. En el segundo bloque, las cadenas léxicas son empleadas para la extracción de conceptos. Son formadas usando WordNet y relaciones asociadas entre palabras. Las cadenas tienen cuatro relaciones.

- **Sinónimos:** son palabras que tienen un significado similar o idéntico entre sí, y pertenecen a la misma categoría gramatical. Ej. Carro tiene como sinónimos: coche, auto, automóvil.

- **Hiperónimo (hypernyms):** El término más alto en una jerarquía terminológica, que incluye a otros. Es el género superordinado (nivel más alto) respecto a sus especies. Ej. El hiperónimo de carro podría ser vehículo, pues está en un nivel superior en la jerarquía respecto a carro, coche, auto, etc.
- **Hipónimos (hyponyms):** Término específico y subordinado a otro más general. El hipónimo tiene al menos un atributo más que el hiperónimo o término superior a él, que lo especifica y le da identidad propia [29]. Ej. Hipónimos de carro serían: camión, camioneta, autobús, puesto que son carros pero tienen características más específicas como el tamaño y el uso de ellos.
- **Merónimo (meronyms):** Palabra que representa un miembro de, que forma parte de, ó que es sustancia de algo [30]. Ej. Merónimos de carro son sus partes: sillas, cajuela, motor y volante.

El índice semántico y el peso son extraídos de acuerdo al valor numérico de la cantidad de información y "Ratio" (proporción) de información, definidas en el proyecto así:

- **Cantidad de información**: Cantidad semántica de un texto, concepto o palabra en todo el documento. Esta magnitud es generada por la composición de todos los conceptos
- **Ratio de Información**: Es la proporción de la cantidad de información de una etiqueta comparativa con la cantidad de información de un texto, concepto o palabra. Esta medida denota la proporción de información de una palabra con respecto al concepto en el cual está incluido.

Siguiendo el proceso anterior, realizaron la comparación evaluativa con otros métodos de indexación como el estándar de frecuencia de términos (Stándard TF) y la extracción del peso semántico. Según su experimento, demostraron gráficamente que los resultados de la extracción por el peso semántico se acercan mucho a los de la extracción de índices de términos semánticos, realizada en este proyecto.

2.5.2.2 Espacio Conceptual

En el proyecto ***Performance And Implications Of Semantic Indexing In A Distributed Environment*** [31], se desarrolla un prototipo que contiene un amplio conjunto de clases y relaciones de datos para el módulo de indexación semántica, construido en un entorno distribuido de análisis. El desarrollo del prototipo se llevó a cabo en dos fases que se describen a continuación:

Fase 1: Se realiza el pre-procesamiento necesario de los documentos. Los pasos de esta fase son:

- Extracción de sintagmas (frases) nominales [32] de cada documento en una colección, los cuales se convierten en los conceptos.
- Estos conceptos son puestos en una lista global, así como una lista que pertenece a cada documento. Y se recogen varias estadísticas para cada concepto.

Fase 2: Se distribuyen las tareas de indexación a diferentes máquinas en el entorno y se utilizan una función de similitud [33] para la asociación de conceptos.

- Se inicializan los procesos en el depósito. Se utilizan mensajes a procesos maestros y esclavos.
- Luego se genera la indexación real, realizada por diferentes equipos de cómputo. Cada proceso esclavo se inicia mediante una asignación de carga de trabajo.
- El paso final es la recopilación de los resultados y el trabajo de limpieza del proceso.

Con lo anterior construyen su prototipo Espacio Intermedio, "Interspace" [34], el cual es la base de indexación semántica (la creación de un espacio conceptual). El espacio conceptual se basa en un cálculo estadístico que determina las relaciones entre los conceptos de una colección de documentos.

Como resultado del proyecto, se obtuvo un mejor rendimiento en la indexación semántica de documentos en un entorno distribuido (cinco máquinas). Concluyen que la indexación semántica en la actualidad, se puede llevar a cabo con pocos recursos en comparación con la indexación realizada en el pasado.

2.5.3 Clasificación De Proyectos por uso de Herramientas

La utilización de algunas herramientas para la indexación y sus respectivas modificaciones ha sido un factor importante para automatizar y mejorar las búsquedas, teniendo en cuenta la semántica de las palabras. Observamos que Lucene es la herramienta líder en el proceso, puesto que al utilizarla en la indexación semántica, se acopla mucho a las necesidades de información. A continuación se describe uno de los proyectos realizados con éxito utilizando dicha herramienta.

El proyecto ***The effect of Semantic Index in Information Retrieval development*** [3], propone un sistema basado en la recuperación de información con índices semánticos denominado: Semantic Information Retrieval System (SIRS), el cual incluye dos módulos importantes: Semantic Indexer SI (indexador semántico) y Query Searcher QS (buscador de consultas). En la Figura 1 se observa el modelo descrito.

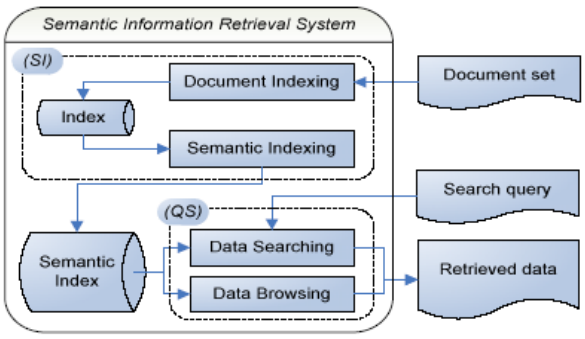


Figura 1: Modelo propuesto por SIRS [3]

Para el primer módulo se presentan dos etapas importantes:

- Indexación de documentos. Se realiza una indexación básica de documentos con la herramienta Lucene [35].
- Indexación semántica (para crear el índice semántico añadiendo más información al índice creado en el primer paso). Los usuarios pueden recuperar los datos por navegación.

El índice semántico no solo se encarga de almacenar todos los registros ordenados sino que también apoya la búsqueda de mecanismos para proporcionar los datos pertinentes.

La investigación de este proyecto se enfoca en lo siguiente.

- **Un sistema SI**: se presenta un SI de formación heurística (llamado indexación semántica) que incluye como primer paso elaborar un índice semántico con lucene y el segundo paso añadir algunas relaciones semánticas, particiones y clases semánticas.
- **Fusión de dos índices semánticos**: puede ser utilizada para combinar índices semánticos de recuperación de información o de bibliotecas digitales para crear nuevos índices. Este proceso es útil para hacer índices de varios dominios específicos y se realiza en dos pasos:
  - Fusión de dos índices semánticos con el mismo atributo y estableciendo la relación
  - Fusión de dos índices semánticos con diferentes atributos y estableciendo la relación.
- **Extracción de sub índice semántico**: Se realiza con el fin de crear un índice de dominios específicos a partir de un índice multi-dominio. Los casos para esta actividad son:
  - Extracción del sub índice semántico con el sub conjunto de atributos y sus relaciones.
  - Extracción de sub índice semántico con la restricción del documento fuente.

En la experimentación de indexación semántica se trabajó en dos pasos:

- Crear un índice normal utilizando las herramientas de Lucene.
- Crear el índice semántico, añadiendo un poco más de información en la base de

entradas en el índice creado en el primer paso.

Como conclusión de este trabajo, se pudo observar que el tamaño de los índices no es muy alto, pero el tiempo de indexación es bastante elevado, por lo tanto las tareas para su trabajo futuro es mejorar la indexación y el algoritmo de búsqueda con el fin de reducir el tiempo de procesamiento.

3. PROYECTOS CON ONTOLOGÍAS E ÍNDICES SEMÁNTICOS

Como observamos anteriormente, las ontologías y los índices semánticos son dos técnicas muy importantes en la resolución de los problemas de recuperación de información, cada una de ellas ha brindado grandes ventajas a la hora de resolver problemas de este tipo, por ello, este apartado se enfoca en la utilización conjunta de éstas dos técnicas.

En los últimos 10 años se han llevado a cabo proyectos que han hecho uso de estas técnicas obteniendo resultados muy favorables. A continuación nombraremos los más importantes.

3.1 CREACIÓN DE ÍNDICES SEMÁNTICOS CON ENRIQUECIMIENTO DE ONTOLOGÍA.

Un problema detectado, es que muchos conceptos extraídos de un documento y que pertenecen a determinado contexto, no están presentes en la ontología de dominio. Por lo cual, en el proyecto ***Ontology enrichment and indexing process*** [6], tiene como objetivo principal construir un índice de estructura de las páginas Web de acuerdo a una ontología, la cual proporciona la estructura del índice. Para llevar a cabo la construcción del índice, ellos proponen cuatro pasos generales, con los cuales se logra la construcción del mismo con estructura como se muestra en la Figura 2.

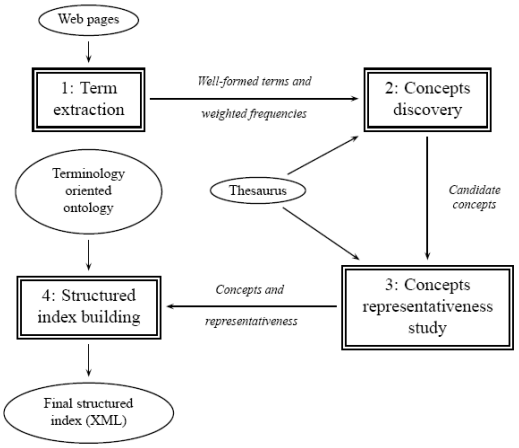


Figura 2: Proceso de Indexación según Desmontils, C.J., L. Simon [6]

- Como primer paso, a cada página se le construye un índice plano de los términos, en donde cada término es relacionado con su frecuencia ponderada.
- Utilizan el tesauro de WordNet [36] para generar los conceptos candidatos que pueden ser etiquetados por un término del índice anterior.
- Cada concepto candidato se estudia para determinar su representatividad en el contenido de la página Web. Esta evaluación se basa en la frecuencia ponderada y en las relaciones con los otros conceptos.
- Entre los conceptos candidatos, se aplica un filtro a través de la ontología y la representatividad de los conceptos. Un concepto seleccionado es un concepto que pertenece a la ontología y tiene una alta representatividad de los contenidos de la página.

Como segundo objetivo del proyecto es el enriquecimiento de la ontología, para lo cual tuvieron en cuenta dos criterios:

- El enriquecimiento de perfeccionamiento (o especialización), que trata de regresar a una ontología más especializada.
- Enriquecimiento mediante la abstracción, que trata de hacer que la ontología sea más general (por la ampliación del campo o suprimiendo conceptos muy específicos).

Se propone un método semiautomático del enriquecimiento de la ontología [37], que ofrece a los expertos un medio de comunicación de gran alcance para gestionar el dominio cubierto por la ontología. Para esto realizan cuatro pasos:

- Un índice estructurado para añadir los conceptos de la ontología que son útiles.
- Un post- tratamiento basado en una poda de la estructura del índice final.
- Una herramienta de la validación de la ontología.
- Enriquecimiento de la ontología utilizando un diccionario de sinónimos.

Para el enriquecimiento de la ontología durante la construcción del índice utilizan WordNet como diccionario de sinónimos, lo cual hace posible el proceso de agregar a la ontología algunos conceptos que están presentes en el índice plano, pero que no pertenecen a la ontología. Para determinar estos conceptos utilizan la heurística basada en rutas de hiperonimia asociadas con los conceptos en WordNet y en las relaciones "IS-A" asociada a los conceptos de la ontología.

El enriquecimiento de la ontología no es completamente automático. Un experto humano tomará la decisión final de añadir o no un nuevo concepto de la ontología, mediante una herramienta de visualización desarrollada, para ayudar en el proceso de indexación y control de la desviación de la ontología potencial.

El proceso propuesto por este proyecto trae consigo una serie de **ventajas** con relación a otros procesos de indexación tradicional. Éstas son:

1. Las páginas seleccionadas contienen, a su vez, las palabras claves y los conceptos necesarios.
2. De estos conceptos, son más representativos los temas tratados en las páginas seleccionadas.
3. Las Páginas puede comprender los conceptos necesarios y los más específicos.
4. La importancia de un concepto no sólo depende de su frecuencia del término, sino también en los marcadores de HTML y también sus relaciones con los otros conceptos de la página.

3.2 CREACIÓN DE ÍNDICE SEMÁNTICO BASADO EN LA FRECUENCIA PONDERADA, Y CÁLCULO DE REPRESENTATIVIDAD DE CONCEPTOS

El siguiente proyecto presenta una nueva idea para indexar un sitio Web, haciendo uso de ontologías y técnicas del lenguaje natural para la recuperación de información en la Internet. Esta propuesta es presentada en la investigación ***Indexing a Web Site with a Terminology Oriented Ontology*** [1], y cuyo objetivo es realizar un proceso semiautomático, que ofrece un índice basado en el contenido de un sitio Web, donde utilizan las técnicas del lenguaje natural.

El proceso de indexación que este proyecto, es similar al descrito en el anterior [6], pues ellos se basan en esta investigación. Sin embargo se presentan a continuación las especificaciones de éste, como se muestra en la Figura 3.

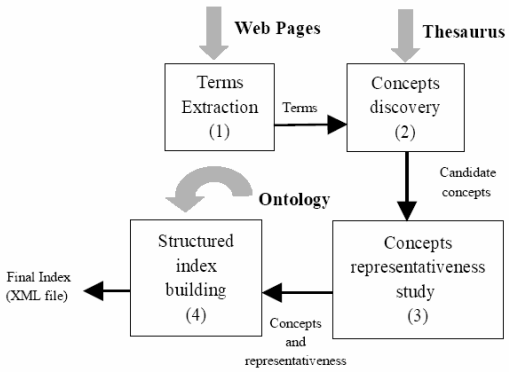


Figura 3: Proceso de Indexación [1]

En este proceso, se determinan todas las etiquetas candidatas de un concepto. Se basa en un diccionario de sinónimos y utiliza un número de heurísticas similares como los propuestos por Microkosmos [38].

Para la construcción del índice tienen en cuenta dos pasos esenciales:

- Condiciones de extracción de páginas Web y el cálculo de la frecuencia ponderada.

- Determinación de los conceptos candidatos y el cálculo de representatividad de un concepto.

Estos dos conceptos son puntos importantes a tener en cuenta en la construcción del índice, pero además cada una de ellas tiene sus respectivas partes.

Para las condiciones de extracción:

- Eliminación de los marcadores de HTML de las páginas Web.
- Dividir el texto en frases independientes.
- Lematización de las palabras incluidas en las páginas. A continuación, las páginas Web se anotan con parte de las etiquetas de voz, utilizando el etiquetador de BRILL [39].

Proceso para generar conceptos candidatos

El proceso de generación de conceptos candidatos se hace con WordNet [40], a continuación a cada concepto candidato se le calcula la representatividad de acuerdo a la frecuencia ponderada y su similitud acumulada del concepto, con relación a los otros conceptos de la página. La similitud acumulada se basa en la similitud entre dos conceptos, lo cual permite evaluar la distancia semántica entre ellos. Esta medida se define relativamente a un tesauro y la relación entre hiperónimos.

Asociación de conceptos y conjunto de sinónimos (synsets)

Los conceptos candidatos se corresponden con los conceptos de la ontología. Si un concepto está en la ontología y en la página Web, la dirección URL de esta página y su representatividad, se añade a la ontología. El proceso de evaluación permite valorar la adecuación entre las páginas y la ontología y así adoptar estrategias diferentes en función del valor de los coeficientes.

En conclusión, el proceso que proponen otorga una serie de ventajas sobre los métodos de indexación tradicional e incluso sobre los métodos de anotación Web. Además, los resultados presentados pueden ser utilizados en diversas aplicaciones. Actualmente se estudian otras relaciones genéricas y relaciones específicas, con el fin de mejorar el proceso de extracción de conceptos. Hoy día este proceso está siendo incorporado en el sistema Bomon Multiagente [41], para buscar información relevante en Internet.

En ***Towards Building Semantic Rich Model for Web Documents Using Domain Ontology*** **[9]**, se enfocan en la construcción de modelos de la Web semántica, para documentos que emplean el análisis del lenguaje natural y un conjunto de ontologías de un dominio especifico (en este caso el ámbito medico). Estos enfoques se utilizan para realizar el análisis textual, que se traduce no solo en la identificación de conceptos importantes presentados en el documento, sino también las relaciones entre estos conceptos. En este proyecto se sigue el enfoque general de la construcción del índice propuesto por Desmontils y Jacquin.

El proceso llevado a cabo es el siguiente:

Análisis de documentos: Para este análisis se realizan los siguientes pasos:

- Se toma los documentos HTML en proceso de transformación, para ser codificados y generar archivos de tipo ASCII, los cuales son documentos libres de etiquetas HTML.
- Luego los documentos son sometidos a un proceso de análisis de palabras.
- En el documento filtrado, todas las palabras vacías serán eliminadas y los conceptos seleccionados se derivan a su raíz para ser ordenados de acuerdo a la frecuencia de aparición en el documento.
- Se dividen los documentos en párrafos y luego en frases, las cuales se almacenaran en un repositorio.
- Los conceptos con alta frecuencia previamente obtenidos en el proceso de análisis de palabras, son comparados con las frases almacenadas en el repositorio, con el fin de seleccionar las frases candidatas para ser utilizadas en el análisis del lenguaje natural (NLA).

Los resultados del proceso de análisis de documentos, es una lista de los posibles conceptos candidatos, y la lista de las frases en donde los conceptos fueron encontrados.

Análisis del lenguaje natural: este análisis se define en las siguientes etapas:

- **Morfología y proceso de acceso de análisis semántico:** se analizarán las frases de entrada (oraciones que contienen conceptos candidatos) previamente almacenadas en el repositorio de frases en un árbol de análisis, utilizando el Analizador Apple Pie Parser [42].
- **Análisis semántico:** Extrae las relaciones semánticas entre los conceptos seleccionados. Se realiza ya sea por la ontología de dominio específico o por la explotación de la estructura semántica de las oraciones analizadas con la ayuda del usuario.
- **Modelo global de la semántica del documento:** Los modelos de los documentos semánticos serán almacenados y sometidos a un proceso de integración, para la creación de un modelo de documento semántico global. Este será usado para la recuperación y la navegación semántica.

Se concluye que las ontologías de dominio, juegan un papel importante en las tareas de clasificación y organización de documentos. En este proyecto se combinó una ontología de dominio con las técnicas del lenguaje natural, donde éste no sólo sirvió para extraer conceptos importantes, sino también para construir el contenido semántico de los documentos Web.

3.3 ALGORITMO DE RANKING Y MODELO VECTORIAL

Uno de los principales objetivos perseguidos en el campo de la Web semántica, radica en la mejora de las técnicas actuales de recuperación de información, mediante el uso de nuevas metodologías englobadas bajo el nombre de búsqueda semántica. ***El Proyecto de trabajo de iniciación a la investigación*** [43], se centra en este objetivo, por lo cual, realiza la implementación de un nuevo modelo de búsqueda semántica enfocada en la recuperación de información sobre grandes repositorios de documentos. Este modelo de recuperación se basa a una ontología de dominio y en bases de conocimiento.

Para este proyecto, se tienen en cuenta las propuestas de KIM [44][45] y TAP [46] que son las más completas publicadas hasta la fecha, para la construcción de bases de conocimiento y la anotación automática a gran escala. Pero este trabajo complementa a KIM y TAP con un algoritmo de ranking, específicamente diseñado para un modelo de recuperación de información basado en ontologías, utilizando un sistema de indexado semántico centrado en la ponderación de anotaciones entre los conceptos de las bases de conocimiento y los documentos almacenados en el repositorio. Además, el presente proyecto se basa en la idea de que "la búsqueda semántica sea un complemento de la búsqueda por palabra clave mientras no haya suficientes ontologías y metadatos disponibles" [47]. El proceso propuesto por este proyecto se puede ver en la Figura 4.

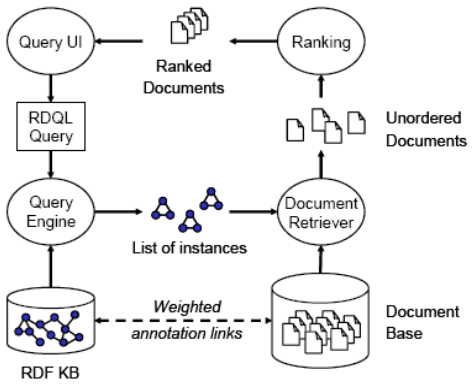


Figura 4: Vista de Modelo de Recuperación de Información Basado en Ontologías [43]

El sistema sigue estos pasos:

- Toma como entrada una pregunta formal expresada en lenguajes como RDQL (Lenguaje de consultas que permite extraer metadatos de archivos disponibles en una URL) [48] o una interfaz de formulario [49].
- La pregunta puede generarse mediante una consulta basada en palabras clave, basada en lenguaje natural, una interfaz de formulario o técnicas de usuario más sofisticadas.
- Se procede a recuperar la información que mejor se adapta a las necesidades del usuario. Este proceso se puede ver en dos fases: la primera es la consulta formal

ejecutada contra una base de conocimiento, se devuelve una lista de instancias o tuplas que cumplen los requisitos de la consulta. Como segunda fase, se utilizan las anotaciones de dichas instancias con los documentos del repositorio para recuperar el conjunto de documentos que satisfacen la consulta del usuario.

- Los documentos son ordenados y presentados al usuario siguiendo una adaptación del modelo vectorial que utiliza los pesos de las anotaciones, para aclarar el orden y así presentar al usuario los documentos que contienen la semántica que mejor responde a la necesidad del mismo.

El modelo de este trabajo puede verse como una evolución del modelo vectorial clásico [50], donde los índices basados en palabras claves, son reemplazados por bases de conocimiento fundamentadas en ontologías.

4. COMPARACION PROCEDIMIENTOS

Con la revisión anterior de los proyectos, se puede realizar una comparación de sus procedimientos y algunos de los pasos que tienen en común. Para esto se describe a continuación una tabla comparativa de cada proyecto y la descripción de los pasos que utiliza. Se marcan los pasos que realiza cada uno en la casilla correspondiente.

La tabla está dividida en dos partes: la primera nombra los pasos de la indexación automática o creación de un índice plano y la segunda parte describe los pasos de un índice semántico. Se observan 8 pasos generales con sus respectivas descripciones (breves), herramientas o técnicas y/o actividades en cada paso. Los proyectos aparecen en la columna izquierda con el número de página en que se describen en el presente survey y la respectiva referencia a su bibliografía.

Tabla 22: Comparación de procedimientos. Generación de índices tradicionales.

	PROCEDIMIENTO			
	1. Extracción términos	2. Extraer cadenas léxicas	3. Tagger o anotador (Part-Of-Speech)	4. calculo de frecuencia

PROYECTOS	Tokenizar, lematizar, análisis léxico	Frases nominales	Análsis léxico y contextual	Ponderación de frecuencia de términos
[27]				
[2]	X	X	X	
[14]	X	X	X	X
[31]		X		
[3]	X.			
[1]	X	X	X	X.
[6]	X.	X	X	X
[9]	X	X	X	X

Tabla 23: Comparación procedimientos. Generación de índices semánticos.

PROYECTOS	PROCEDIMIENTO							
	5. Extracción de conceptos			6. Análisis de conceptos		7. calculo de frecuencia		8. Asociación términos y Synsets
	WorNet	Ontología	Otro	Espacio Conceptual	Relaciones Conceptos	Representatividad	Reponderación	Asociación de términos
[27]	X			X		X	X	
[2]	X					X		X
[14]			X	X		X		
[31]			X	X		X		
[3]					X			
[1]		X			X	X		X
[6]	X	X			X	X		X
[9]		X	X		X			

La comparación anterior permite observar con mayor claridad los pasos más comunes en la creación de índices semánticos. La extracción de términos y frases nominales se realiza en varias investigaciones de manera conjunta, lo cual, según sus autores, genera buenos resultados en la recuperación de información. Así mismo el estudio de relaciones entre conceptos y/o el cálculo de representatividad, son necesarios en la construcción de dicho índice. Las herramientas más utilizadas son WordNet (tesauro) y las ontologías, los cuales proporcionan relaciones semánticas más precisas para la extracción de conceptos.

5. CONCLUSIONES

La recuperación de información es muy importante para gran parte de los usuarios de la web. Es por ello que los investigadores se han enfocado en lograr técnicas que permitan encontrar lo que realmente se busca en internet. La indexación semántica comenzó su desarrollo buscando mejoría en las búsquedas de los usuarios, por lo cual, varios investigadores han trabajado en el proceso, metodología y herramientas que permitan logros importantes a la hora de recuperar documentos.

La creación de índices semánticos sigue un proceso previo ya establecido para los motores de búsqueda, el cual se basa en una indexación automática de los documentos en la web. A partir de ahí, se crea el índice semántico, comparando generalmente, conceptos y términos para extraer relaciones semánticas existentes en el lenguaje natural. Esto proporciona mayor eficiencia en las búsquedas que se realizan, no solo por palabras clave sino también, por frases con sentido propio.

La investigación y comparación realizadas en este survey, presentaron una visión general y un tanto específica del proceso de creación de los índices semánticos, lo cual ayuda a los actuales y próximos investigadores a realizar un proceso más eficiente en la indexación semántica y mejoría en la recuperación de información.

6. REFERENCIAS

[1] C. Jacquin and E. Desmontils, "Indexing a Web Site with a Terminology Oriented Ontology", IRIN 2002, Université de Nantes 2: Standford University. pp. 181-198.

[2] M. Rada and D. Moldovan, "Semantic Indexing using WordNet Senses". Southern Methodist University: In Proceedings Of Acl Workshop On Ir & Nlp, 2000, Hongkong. p. 11.

[3] T. Nguyen and T. Phan, "The effect of Semantic Index in Information Retrieval development". International Conference on Information Integration and web-based Applications and Services iiWAS, 2008, Austria. pp. 438-441.

[4] José M. Diaz N., F.S., Mario Pérez. Recuperación de Información. . 2009 [cited 14 de julio de 2010]; Available from: http://sites.google.com/site/glosariobitrum/Home/recuperacion-de-informacion.

[5] Molina, M.P. Búsqueda y Recuperación de Información. 2009 [cited 27 de abril de 2010]; Available from: http://www.mariapinto.es/e-coms/recu_infor.htm.

[6] E. Desmontils, C.J. and L. Simon, "Ontology enrichment and indexing process". 2003, Institut de Recherche en Informatique de Nantes 2, rue de la Houssinire. p. 18.

[7] S. Chagheri, C. Roussey, S. Calabreto, C. Dumoulin, "Semantic Indexing of Technical Documentation", in Laboratoire d'InfoRmatique en Image et Systèmes d'information. 2009, Université de LYON: Toulouse, France. p. 12.

[8] Song Jun-feng, Z.W., Xiao W., Li G., Xu Z, "Ontology-Based Information Retrieval Model for the Semantic Web", in Proceedings of the 2005 IEEE International Conference on e-Technology, e-Commerce and e-Service. 2005, IEEE Computer Society: Washington, DC, USA. pp. 152 – 155.

[9] Shahrul A. N., L.Z., Arifah C. A., Tengku M. T. S., Saidah S., "Towards Building Semantic Rich Model for Web Documents Using Domain Ontology", in Proceedings of the 2004 IEEE/WIC/ACM International Conference on Web Intelligence. 2004, National University of Malaysia. p. 769 - 770.

[10] Suárez Barón M., S.V.K., "An Approach to Semantic Indexing and Information Retrieval",. Revista Facultad de Ingeniería Universidad de Antioquia, 2009. **48**: pp. 14.

[11] Dr. Clara Yu, D.J.L.C., Aaron Coburn. "The Semantic Indexing Project knowledge search" 2003 [cited 12 de Diciembre de 2009]; Available from: http://www.knowledgesearch.org/.

[12] Cobos L. Carlos A., A.R.J.K., Constaín D. William A., "Hibridación De La Mejor Búsqueda Armónica Global Y El Algoritmo K-Means Para El Clustering De Documentos Web", 2009, [Pregrado Thesis]. Universidad Del Cauca: Popayan. p. 85.

[13] Powell, K.R., The Interspace Prototype: An anaysis Environment for Semantic Interoperability. 1997. [cited 10 de mayo de 2010]; Available from: http://www.canis.uiuc.edu/INTERSPACE/.

[14] Yi-Ming Chung, Q.H., Kevin P. and B. Schatz, "Semantic Indexing for a Complete Subject Discipline", in Proceedings of the fourth ACM conference on Digital libraries 1999, University of Illinois at Urbana-Champaign, Champaign,

IL 61820: International Conference on Digital Libraries, Berkeley, California, United States pp. 39-48.

[15] Company, R. infoREUSER - Motor de búsqueda semántico. 2010 [cited 12 de abril de 2010]; Available from: http://www.reusecompany.com/producto.aspx?id=4.

[16] Gómez, E.E., "Una nota metodológica sobre los análisis cualitativos. El análisis de las relaciones entre los elementos: el análisis de las frecuencias y co-ocurrencias. Sistema de Información Científica", Redalyc, 2009. **18**: p. 57.

[17] Moreno L. P., Ferrández S., Roger S., Ferrández A., Aguilar A., "Nueva Propuesta de Desambiguación de Sentidos de Palabras para nombres en un sistema de Búsqueda de Respuestas". Procesamiento del Lenguaje Natural, 2006, vol. **36**, pp. 47-53.

[18] Leal, E.T. "La Desambiguación del Sentido de las Palabras: revisión metodológica". Revista multidisciplinar sobre diseño, personas y tecnología 2009 [cited 10 de marzo de 2010]; Available from: http://www.nosolousabilidad.com/articulos/desambiguacion.htm.

[19] Rada M., Moldovan D.I., "An Iterative Approach to Word Sense Disambiguation", in Proceedings of the Thirteenth International Florida Artificial Intelligence Research Society Conference 2000, AAAI Press: Orlando, FL. pp. 219 - 223.

[20] Brill, E., "A Simple Rule-Based Part Of Speech Tagger", in Proceedings of the workshop on Speech and Natural Language. 1992, Association for Computational Linguistics Harriman, New York. pp. 112 - 116.

[21] H. Chen, B.S., D. Ng, J. Martinez, A. Kirchhoff, C. Lin., "A Parallel Computing Approach to Creating Engineering Concept Spaces for Semantic Retrieval: The Illinois Digital Library Initiative Project". IEEE Transactions on Pattern Analysis and Machine Intelligence, 1996, vol. **18**(8): p. 39.

[22] Hsinchun Chen, A.H., R. Sewell, y B. Schatz., "Internet Browsing and Searching: User Evaluations of Category Map and Concept Space Techniques". Journal of the American Society for Information Science, 1998, vol **49(7)**: pp. 582-603.

[23] Cherukuri Aswani Kumar, S.S., "Latent semantic indexing using eigenvalue analysis For efficient information retrieval". Int. J. Appl. Math. Comput. Sci, 2006, vol **16**: pp. 551-559.

[24] Brill, E., "A Corpus-Based Approach to Language Learning", in The Institute For Research In Cognitive Science. 1993, University of Pennsylvania: Philadelphia. p. 166.

[25] Translations, L. NPtool, a detector of English noun phrases. 1993 [cited 4 de mayo de 2010]; Available from: http://www2.lingsoft.fi/doc/nptool/.

[26] Stuart Nelson: Head, M.S.H. Medical Subject Headings. 2010 [cited 14 de abril de 2010]; Available from: http://www.nlm.nih.gov/mesh/.

[27] Kang, B.-Y., "A Novel Approach to Semantic Indexing Based on Concept", in Proceedings of the 41st Annual Meeting on Association for Computational Linguistics. 2003, Association for Computational Linguistics: Sapporo, Japan. pp. 44-49.

[28] J. Morris, G.H., "Lexical Cohesion Computed by Thesaural Relations as an Indicator of the Structure of Text", in Computational Linguistics. 1991. pp. 21 - 48.

[29] Barite, M. Diccionario de Organización y representación del Conocimiento: Clasificación, Indización, terminología. 2000 [cited 10 de mayo de 2010]; Available from: http://www.eubca.edu.uy/diccionario/letra_h.htm.

[30] Costa, J. Meronimia. 2006 [cited 10 de mayo de 2010]; Available from: http://www.solodisenio.com/que-es-meronimia/.

[31] Conrad T. K. Chang, B.R.S., "Performance and Implications of Semantic Indexing in a Distributed Environment", in Proceedings of the eighth international conference on Information and knowledge management 1999, ACM, New York: Kansas City, Missouri, United States. pp. 391-398.

[32] Tuggy, D. Lecciones para un curso del náhualt moderno 2002 [cited 12 de mayo de 2010]; Available from: http://www.sil.org/~tuggyd/nahuatllecciones/l07/lecc_07_nlv.htm.

[33] Hsinchun Chen, K.J.L., "Automatic Construction of Networks of Concepts characterizing Document Databases". IEEE Transactions on Systems, man, and cybernetics, 1992. **22**: p. 18.

[34] Schatz, B.R., "Information Retrieval in Digital Libraries: Bringing Search to the Net". Science - Bioinformática, 1997, vol. **275**: pp. 327 - 334.

[35] Cutting, D. The Apache Lucene. 2010 [cited; Available from: http://lucene.apache.org.

[36] George A. Miller, R.B., C. Fellbaum, D. Gross, and K. Miller, "Introduction to WordNet: An

On-line Lexical Database". International Journal of lexicography, 1993, vol. **3**: pp. 235-244.

[37] Joerg-U. Kietz, A. Maedche, R. Volz, "A Method for Semi-Automatic Ontology Acquisition from a Corporate Intranet", in Swisslife Information Systems Research Lab, Zuerich, Switzerland. 2000, AIFB, Univ. Karlsruhe: Karlsruhe, Germany. pp. 2-6.

[38] T. O'Hara, K.M., and S. Niremburg, "Lexical Acquisition with WordNet and Microkosmos Ontology", in Proceedings of the ACL Workshop on the Use of WordNet in NLP. 1998. pp. 94-101.

[39] Brill, E., "Transformation-based error-driven learning and natural language processing: a case study in Part-of-speech", Tagging Computational Linguistics, 1995, vol. **21**(4): pp. 543-565.

[40] University, P. WordNet 3.0 Princenton University 2009 [cited 2 de febrero de 2010]; Available from: http://wordnet.princeton.edu/wordnet.

[41] S Cazalens, E.D., C Jacquin, and P Lamarre, "A Web Site Indexing Process for an Internet Information Retrieval Agent System" in Proceedings of the First International Conference on Web Information Systems Engineering (WISE'00). 2000: Hong Kong , China. pp. 254-258.

[42] Satoshi Sekine, R.G. Apple Pie Parser - Proteus Project. 2002 [cited 21 de abril de 2010]; Available from: http://nlp.cs.nyu.edu/app/.

[43] Miriam Fernández, T.P.C.A., "Proyecto de trabajo de Iniciación a la investigación", in Escuela Politécnica Superior, Universidad Autónoma de Madrid: Madrid. p. 5.

[44] Popov, B., Kiryakov, A., Ognyanoff, D., Manov, D., Kirilov, A., "KIM – A Semantic Platform for Information Extaction and Retrieval". Journal of Natural Language Engineering, 2004, vol. **10**: pp. 375-392.

[45] Kiryakov, A., Popov, B., Terziev, I., Manov, Ognyanoff, D., "Semantic Annotation, Indexing, and Retrieval". Journal of Web Sematics Journal of Web Semantics, 2004, vol. **2**: pp. 49-79.

[46] Castells, P., Fernández, M., Vallet, D., Mylonas, P., Avrithis, Y., "Self-Tuning Personalized Information Retrieval in an Ontology-Based Framework", in 1st IFIP International Workshop on Web Semantics (SWWS 2005). 2005: Agia Napa, Cyprus. pp. 977-986.

[47] J. Mayfield, T. Finin, "Information retrieval on the Semantic Web: Integrating inference and retrieval", in Proceedings of the SIGIR Workshop on the Semantic Web Workshop - ACM. 2003, The Johns Hopkins University and University of Maryland: Toronto, Canada. p. 7.

[48] Junior Sinche, J.Fierro, Servicios Web Semánticos. 2008 [cited 12 de mayo de 2010]; Available from: http://www.slideshare.net/guesta5bc77/servicios-web-semnticos-presentation.

[49] A. Maedche, S. Staab., N. Stojanovic, R. Studer, and Y. Sure, "SEmantic portAL — The SEAL approach". Spinning the Semantic Web, 2003: p. 27.

[50] G. Salton, M.J. McGill, "Introduction to Modern Information Retrieval", McGraw-Hill, Editor. 1983: New York, p. 400.

3 ANEXO C

3.1 ONTOLOGIAS

Las ontologías son una parte importante en la investigación. A continuación se presentan los componentes y tipos de ontologías, además de las herramientas para trabajar en su construcción.

3.1.1 COMPONENTES DE LA ONTOLOGÍA

Gruber [1] menciona los diferentes componentes de las ontologías que sirven para representar el conocimiento de un dominio. A continuación se mencionan estos componentes:

COMPONENTES	DEFINICIONES
Conceptos	También llamados clases, son ideas básicas que se intentan formalizar. Describen y representan conceptos relevantes de dominio. Se utilizan para estructurar el conocimiento y percepción del mundo circulante.
Propiedades	Están asociadas a los conceptos y describen sus características y atributos.
Atributos	Expresan cualidades de un concepto y el valor de dichas cualidades que pueden ser tanto escalar como literal.
Relaciones	Propiedades que conectan dos conceptos entre sí de manera biunívoca. Representan la interacción y enlace entre los conceptos de un dominio.
Funciones	Son un tipo concreto de relación donde se identifica un elemento mediante el cálculo de una función que considera varios elementos de la ontología.
Axiomas	O reglas de restricción, son teoremas que se declaran sobre relaciones que deben cumplir los elementos de la ontología. Modelan sentencias lógicas que se verifican siempre.
Instancias	Son entidades que pertenecen a una determinada clase. Se utilizan para representar objetos determinados de un concepto.

Tabla 24. Componentes de la Ontología

3.1.2 TIPOS DE ONTOLOGÍAS

En el entorno podemos encontrar diferentes tipos de ontologías. A continuación nombraremos algunos tipos de ontologías según su clasificación.

Clasificación Según el conocimiento contenido se puede clasificar en dos tipos de ontologías:	
Primera clasificación según Van Heijst [2]	Segunda clasificación según Mizoguchi ey at, 1995 [3]
Ontologías terminológicas, lingüísticas: son aquellas que especifican los términos usados para representar conocimiento en el dominio.	**Ontologías de dominio:** contiene todos los conceptos asociados a un dominio particular.
Ontologías de información: especifican la estructura de los registros de la base de datos.	**Ontologías de tarea:** establecen la forma en la cual se pueden usar para realizar tareas específicas.
Ontologías para moldear conocimiento: especifican conceptualizaciones de conocimiento. Están tienen una estructura interna mucho más rica que las ontologías anteriores, y son de gran interés para los desarrolladores de sistemas basados en conocimiento.	**Ontologías generales:** contiene descriptores generales sobre objetos, eventos, relaciones temporales, relaciones causales, modelos de compartimiento y funcionales.

Tabla 25. Clasificación por conocimiento contenido

Clasificación Según la motivación Se clasifica en dos tipos:	
Primera clasificación según Davis et al, 1993 [4].	Segunda clasificación según Poli [5]
Ontologías para la representación del conocimiento: explican las conceptualizaciones que subyacen de los formalismos de representación del conocimiento.	**Ontologías generales:** tiene que ver con las categorías fundamentales y sus conexiones de dependencia. Con respecto a las categorías fundamentales hay dificultad de manejar este nivel supremo, por ello es de máxima importancia emplear una organización de categorías principales que sea lo más trasparente posible.
Ontologías genéricas: definen conceptos genéricos en diferentes áreas. Estas también son conocidas como ontologías abstractas o súper-teóricas por que permiten definir conceptos abstractos y pueden ser usadas para definir conceptos de forma más específica en diferentes dominios.	**Ontologías categóricas:** estudian las diversas formas en las que una categoría se da cuenta de los diversos niveles ontológicos, determinando la posible presencia de una teoría general que subsume sus concretizaciones.
Ontologías de dominio: definen conceptualizaciones específicas del dominio. Las metodologías actuales de adquisición del conocimiento distinguen entre ontologías y conocimiento de dominio; el último describe situaciones actuales del dominio, mientras que las ontologías imponen descripciones sobre la estructura y el contenido del conocimiento del dominio.	**Ontologías de dominio:** se refieren a la estructuración detallada de un contexto de análisis con respecto a los subdominios que lo componen.
Ontologías de aplicación: están ligadas al desarrollo de una aplicación concreta. Estas ontologías cubren los aspectos relacionados con aplicaciones particulares. Estás también toman conceptos de ontologías de dominio genéricas así como métodos específicos para realizar la tarea, por lo que no son muy adecuadas para ser reutilizadas.	**Ontologías genéricas:** con ellas se pueden clasificar los términos en varios niveles, esto significa que cada término debería ser accesible por defecto únicamente en su sentido genérico, mientras que sus significados especializados quedan para cuando se active una ontología de dominio especifica.
	Ontología regional: analiza las categorías y sus conexiones de interdependencia para cada nivel ontológico.
	Ontología aplicada: estas ontologías son la aplicación concreta de entorno ontológico a un objeto específico.

Tabla 26. Clasificación por Motivación

Clasificación según otros aspectos También se pueden clasificar los distintos tipos de ontologías atendiendo diversos aspectos, entre ellas podemos destacar las más importantes, esta clasificación es dada por Steve (1998)[6]	
Según el ámbito del conocimiento	**Ontologías generales:** son las ontologías de nivel más alto, ya que describen conceptos generales (espacio, tiempo, materia, objeto, etc.) **Ontologías de dominio:** describen el vocabulario de un dominio concreto del conocimiento. **Ontologías específicas:** son ontologías especializadas que describen los conceptos para un campo limitado del conocimiento o una aplicación concreta.
Según el tipo de agente al que vaya destinado:	**Ontologías lingüísticas:** se vinculan a aspectos lingüísticos, esto es, aspectos gramáticos, semánticos y sintácticos destinados a su utilización por los seres humanos. **Ontologías no lingüísticas:** destinadas a ser utilizadas por robots y agentes inteligentes. **Ontologías mixtas:** combinan las características de las anteriores.
Según el grado o	**Ontologías descriptivas**: incluyen descripciones, taxonomías de conceptos, relaciones

nivel de abstracción y razonamiento lógico que permitan:	entre los conceptos y propiedades, pero no permiten inferencias lógicas. **Ontologías lógicas:** permiten inferencias lógicas mediante la utilización de una serie de componentes como la inclusión de axiomas, etc.

Tabla 27. Clasificación según otros Aspectos

3.1.3 HERRAMIENTAS PARA TRABAJAR CON ONTOLOGÍAS [7]

Para la construcción de una ontología, la principal herramienta son los editores de ontologías. Estos suelen ser desarrollados para un tipo de lenguaje específico, pero muchos de ellos incorporaron módulos para soportar otros lenguajes de especificación.

Se realizo una comparación de las diferentes herramientas para trabajar con ontologías (editores, APIs, etc.) existentes, con el fin de tener una perspectiva clara de sus principales características y que sirva como base para seleccionar la mejor dependiendo del proyecto. Para esto se darán las definiciones de cada una de las herramientas:

Protégè: es una herramienta para el desarrollo de ontologías y sistemas basados en conocimiento.

Sesame: es un API para java, es decir un entorno para el desarrollo de aplicaciones, es el lenguaje de programación java para la Web semántica.

Jena: es un framework de código abierto para desarrollar aplicaciones en java con tecnologías de la Web semántica. Jena permite gestionar todo tipo de ontologías, almacenarlas y realizar consultas en ellas.

OntoEdit: es una herramienta de edición de ontologías que apoya el desarrollo y el mantenimiento de las mismas, utilizando medios gráficos en un entorno Web, pretende además que las ontologías puedan ser almacenadas y posteriormente manipuladas por base de datos relacionales [8].

WebODE: es una herramienta que basa la construcción de ontologías en el método Methontology, permite exportar el conocimiento a diferentes lenguajes de especificación. La última versión es WebODE 2.0.9 de noviembre de 2003 [9].

Herramientas	Frames	OWL	RDF Shema	XML Shema	DAML	RDF	Libre	OIL
Protégè	X	X	X	X			X	
Sesame:			X			X		
Jena:		X	X		X	X	X	
OntoEdit:						X	X	X
WebODE:		X			X			X

Tabla 28. Herramientas para trabajar con ontologías

Como se observa en la **¡Error! No se encuentra el origen de la referencia.**, la mejor herramienta para trabajar con ontologías es protege, ya que la herramienta soporta dos formas de modelar ontologías: con Frames y OWL, también las ontologías generadas pueden ser

exportadas a varios formatos incluyendo RDF Shema, OWL y RDF Shema. Además se actualiza con bastante regularidad y se le pueden agregar módulos y plugins con nuevas funcionalidades.

3.1.4 REPOSITORIOS DE ONTOLOGÍAS

Existen repositorios de ontologías, donde algunos de ellos son de uso privado , pero también podemos encontrar repositorios libres como WordNet[10], que es empleado como recurso para muchas aplicaciones que trabajan con procesamiento de lenguaje natural (PNL) y la RI. WordNet es una gran base de datos léxica de ingles donde los sustantivos, verbos, adjetivos y adverbios son agrupados en un conjunto de sinónimos (synsets) y cada uno de ellos expresa un concepto diferente, los synsets están vinculados entre sí, por medio de la semántica conceptual y las relaciones léxicas. WordNet constituye un intento de reflejar el modelo de memoria léxica basado en redes semánticas propuesto por Collins y Quillian [11] en un modelo lexicográfico de organización léxica.

También encontramos el repositorio UAB[2] que está diseñado principalmente para ser un lugar central de almacenamiento para las ontologías, el cual podría ser de utilidad para los desarrolladores de herramientas para realizar pruebas.

Entre otros repositorios de ontologías existe uno creado y desarrollado en el marco del proyecto DARPA, en los laboratorios de Knowledge Sharing Effort (KSE)[3] de la Universidad de Stanford [12]. Inicialmente presentaba un editor central para editar ontologías utilizando el lenguaje Ontolingua. Por ello también lleva su nombre: **Ontolingua Server**[4]. Actualmente ofrece un soporte automático para las tareas de integración y diagnostico de múltiples ontologías, además permite la creación y mantenimiento de ontologías distribuidas en la Web, implementando la 'Stanford's Ontology Algebra'[13]. Este servidor ofrece las herramientas necesarias para crear ontologías, integrarlas con otras ya existentes, e incorporarlas a nuevos productos de software. También se proporcionan las herramientas necesarias para usar las ontologías en conjunto con lenguajes como KIF[5]. Por otra parte, se ha incluido una API para integrar las ontologías del servidor con agentes preparados para Internet [15].

3.2 RECUPERACIÓN DE INFORMACIÓN

En la recuperación de información es importante tener en cuenta los modelos, las herramientas, y técnicas que permiten una ejecución eficiente de los sistemas. Algunos de los elementos más importantes a la hora de hablar de recuperación de información en la Web son Los índices, tesauros y palabras claves, los cuales también se describen en los siguientes apartados.

---

2 TONES Ontology Repository, Disponible en http://owl.cs.manchester.ac.uk/repository/.

3 NECHES, Robert. *The Knowledge Sharing Effort*. Disponible en: *http://www-ksl.stanford.edu/knowledge-sharing/papers/kse-overview.html*.

4 Repositorio disponible en http://ontolingua.stanford.edu/.

5 KIF: "es un lenguaje propuesto como estándar para describir cosas. Es legible tanto para personas como para maquinas, además fue diseñado para funcionar como medidor entre la traducción de otros lenguajes. La descripción del lenguaje incluye una especificación para su sintaxis y otra para su semántica" 14. Jimenez, A. *KIF – Knowledge Interchange Format*. 2007 [cited 9 de abril de 2010]; Available from: http://alfonsojimenez.com/uncategorized/kif-knowledge-interchange-format/.

3.2.1 MODELOS DE RECUPERACIÓN DE INFORMACIÓN

Los modelos de recuperación de información permiten la comparación entre una consulta determinada y una colección de textos sobre los que se realiza la consulta. Para su ejecución, crean un índice determinado en función del contenido del documento de texto que se necesita recuperar[6]. Estos modelos definen la manera como se representan los documentos, las consultas y su emparejamiento en el sistema[16].

A continuación se describen los modelos de recuperación clásicos:

- ***Modelo Booleano.*** Está basado en la teoría de conjuntos y el algebra booleana. En su proceso se agrupan los documentos, los cuales están dispuestos por conjuntos de términos y asumen las preguntas como expresiones booleanas[17]. El algoritmo de decisión binaria que utiliza, permite determinar si un elemento está presente o no en el conjunto de resultados.
- ***Modelo Vectorial.*** Se basa en espacios vectoriales donde los documentos solo se emparejan parcialmente con la pregunta[7]. Asigna unos pesos no binarios a los índices de las preguntas y así comparar la similitud entre cada documento almacenado en el sistema.
- ***Modelo Probabilístico****.* En este modelo se calcula la probabilidad de que una pregunta realizada por el usuario esté relacionada con el documento revisado, lo cual lleva a determinar los documentos relevantes para cada consulta.

3.2.2 HERRAMIENTAS

3.2.2.1Bases de datos

Es una colección de información organizada por registros, campos y archivos, y que permite ser leída por un programa de computador[8]. Esto brinda a los usuarios una recuperación de información más rápida y ordenada en muchos tipos de información como referencias, textos, imágenes, etc.

3.2.2.1.1 Internet

Se definen algunos tipos de herramientas en Internet. A continuación se describen en la Tabla 29.

Concepto	Descripción
Revistas electrónicas	Son una forma de publicación periódica que tiene como soporte el uso de elementos electrónicos y de tecnologías de la información. Se caracteriza por su capacidad interactiva y su contenido está organizado mediante vínculos hipertextuales[9]
Buscadores	Permiten localizar y recuperar la información almacenada en Internet. Estas herramientas rastrean la Web con programas internos para encontrar documentos relevantes al usuario que realiza una consulta específica. Aunque realizan un trabajo de indexación de todos los documentos

[6] Extraído de *Recuperación y organización de la Información.* Disponible en http://modelosderecuperacioni.iespana.es/.

[7] Extraído de *Modelo de Recuperación vectorial.* Disponible en http://modelosderecuperacioni.iespana.es/.

[8] Extraído de *Definición de Base de datos*. Disponible en http://www.masadelante.com/faqs/base-de-datos.

[9] Extraído de http://caribe.udea.edu.co/~hlopera/revista_electronica.html.

	disponibles, no pueden llegar a todo el contenido cuando el documento se encuentra en una base de datos, así, esta información solo aparece dinámicamente [18]
Directorios	Son listas ordenadas que permiten acceder a la información de forma estructurada y jerárquica, con lo cual clasifican la información contenida en la Web.
Meta-buscadores	Sistemas desarrollados para evitar explorar información en un buscador aparte o única base de datos, de manera que colecciona las respuestas de los buscadores y las unifica para enviarlas al usuario que realizó la consulta[10]
Buscadores selectivos	utilizan una base de datos especializada en una materia o área
Agentes inteligentes	son herramientas o entidades software que permiten localizar la información de forma automática, basándose en su propio conocimiento [19]

Tabla 29. Tipos de herramientas en Internet

3.2.2.2 Lenguajes de indexación y control terminológico

A continuación (Tabla 30) se describen los lenguajes de indexación y control terminológico más utilizados en la Web. Sus elementos son tipos o componentes de cada uno y la descripción muestra su uso.

Lenguaje de indexación y control terminológico	Elementos	Descripción
Índices: Listado de términos organizados en categorías que representan el contenido de un recurso Web. Según María Pinto [20], algunos tipos de índices son	Índice de materias	Términos ordenados según las materias que trata la base de datos, el buscador, etc.
	Índice alfabético	Listado de términos alfabéticamente
	Índice KWIC	Tipo de índice permutado en el que el contenido temático de una obra se representa mediante palabras clave de su título o de otra fuente de información del documento
	Índice KWOC	Tipo de índice permutado que varía en su presentación respecto al índice KWIC, en que las palabras clave aparecen como un encabezamiento en línea separada. Bajo cada encabezamiento aparece la totalidad de los títulos, completos o truncados, que contienen la palabra clave de que se trata.
Palabras clave (Keywords). En el Procesamiento de Lenguaje Natural, significa un dato relevante y a su vez permite representar el contenido del documento cuando se encuentra en él.		La mayoría de los buscadores utilizan Meta-Keywords para localizar los recursos, las palabras clave de cada página Web. Por esta razón es esencial que cada página tenga una etiqueta donde se incluyan las palabras clave que la identifican, a su vez, la definición exacta de cada una de ellas, pues es a partir de estas que los buscadores localizan o no un recurso
Tesauros. Se refiere a una lista estructurada de conceptos, destinados a representar de manera unívoca el contenido de los documentos y de las consultas dentro de un sistema documental determinado. Ayuda al usuario en la	**Componentes**	**Descriptores admitidos o preferentes**: Términos normalizados (han sufrido un proceso de revisión rechazando plurales, evitando sinónimos, etc.) que el tesauro los considera aptos para asignarlos a un documento y que posteriormente facilite la recuperación **Descriptores no admitidos**: son los que aun estando normalizados, no se considera adecuado utilizarlos (suelen ser sinónimos, términos no utilizados en el campo de actuación, etc.)
	Relaciones	**Asociativas**: Indican que los términos guardan alguna

[10] Extraído de http://www.um.es/gtiweb/adrico/.

indexación de los documentos y de las consultas[11]. Es una lista de términos ordenados jerárquicamente sobre un área, la cual mantiene entre sí relaciones semánticas y genéricas.		relación. **Jerárquicas**: Indican cuando un término es más específico que otro. **Sinónimos**: Indican que dos términos son sinónimos y cuál de ellos se utiliza como admitido.

Tabla 30. **Lenguajes de indexación y control terminológico.**

3.2.3 TÉCNICAS DE RECUPERACIÓN DE INFORMACIÓN

Existe algunas técnicas de recuperación de información descritas como sigue de acuerdo a las especificaciones de María Pinto [20]:

- Sistemas de recuperación de lógica difusa

Permiten establecer consultas con frases normales, de forma que se eliminen signos de puntuación, artículos, conjunciones, plurales, tiempos verbales, palabras comunes (que suelen aparecer en todos los documentos) y dejando sólo aquellas palabras que el sistema considera relevantes. La recuperación se basa en proposiciones lógicas con valores de verdadero y falso, teniendo en cuenta la localización de la palabra en el documento.

- Técnicas de ponderación de términos

La ponderación pretende darle un valor adecuado a la búsqueda dependiendo de los intereses del usuario, puesto que algunos criterios en la búsqueda tienen más valor que otros. Los documentos recuperados se encuentran en función del valor obtenido en la ponderación. El valor depende de los términos pertinentes que contengan el documento y la frecuencia con que se repita. Así, el documento más pertinente de búsqueda sería aquel que tenga representado todos los términos de búsqueda y además el que más valor tenga, independientemente de donde se localice en el documento.

- Técnica de clustering

En esta técnica se agrupan elementos similares por medio de algoritmos matemáticos (de ranking) para clasificar documentos automáticamente, considerando las similitudes en su contenido[12] y por orden de importancia.

Algoritmos utilizados para realizar la categorización (cluster):
Algoritmo K-means: El más usado debido a su eficiencia y simplicidad, se trata de minimizar la distancia promedio entre los documentos [21].
COBWEB[13]: Algoritmo de clustering jerárquico. Utiliza aprendizaje incremental.
Algoritmo EM: Generalización del K-means que se puede aplicar a una gran variedad de representaciones de documentos [21].

- Técnicas de retroalimentación por relevancia

[11] Extraído de *la representación y organización de la información a través de los tesauros*. Disponible en: http://www.eumed.net/rev/cccss/06/smg.htm.
[12] Extraído de *Búsqueda y Recuperación de Información.* Disponible en: http://www.hipertexto.info/documentos/busq_rec.htm.
[13] Extraído de *Comparación de diferentes algoritmos de clustering en la estimación de coste en el desarrollo de software*. Disponible en: http://www.sc.ehu.es/jiwdocoj/remis/docs/GarreAdis05.pdf.

Esta técnica pretende obtener el mayor número de documentos relevantes tras establecer varias estrategias de búsqueda. La idea es que, tras determinar unos criterios de búsqueda y observar los documentos recuperados se vuelva a repetir nuevamente la consulta pero esta vez con los elementos interesantes, seleccionados de los documentos primeramente recuperados. Un ejemplo de este tipo es la utilización de los Algoritmos Genéticos.

- Técnicas de stemming

Morfológicamente las palabras están estructuradas en prefijos, sufijos y la raíz. La técnica de Stemming lo que pretende es eliminar las posibles confusiones semánticas que se puedan dar en la búsqueda de un concepto, para ello trunca la palabra y busca solo por la raíz.

Algoritmos utilizados para desechar prefijos y sufijos:
Paice/Husk [22]
S-stemmer / n-gramas
Técnicas lingüísticas

- Técnicas lingüísticas

Pretenden acotar de una manera eficaz los documentos relevantes. Por esta razón, esta técnica lo consigue mediante una correcta indización en el proceso de tratamiento de los documentos con ayuda de índices, tesauros, etc.; evitando las ambigüedades léxicas y semánticas a la hora de establecer las consultas.

3.2.4 CALIDAD DE LA RECUPERACIÓN

A continuación se presentan unos criterios básicos para que la recuperación llevada a cabo sea de calidad:

CRITERIO	DESCRIPCIÓN
Consistencia	Capacidad que tiene un sistema de búsqueda en coordinar su sistema de clasificación con el lenguaje de búsqueda, permitiendo de esta manera establecer ecuaciones de búsqueda sobre términos admitidos
Exhaustividad	Es la cualidad de un sistema de información para recuperar la totalidad de los documentos relevantes que posee una colección, conforme a los requerimientos establecidos en la estrategia de búsqueda. Esta cualidad expresa la eficiencia en recuperar la mayor cantidad de documentos relevantes para el usuario
Tasa de acierto	Coeficiente que surge de dividir el número de documentos relevantes recuperados, sobre el número total de documentos relevantes de la colección. Como ejemplo de esto, se puede mencionar la búsqueda realizada por el usuario de Paris Hilton y los documentos mostrados que se relacionen con la artista en vez de los documentos sobre la ciudad (Paris) y el hotel (Hilton). El acierto en este caso se puede medir como los documentos relevantes existentes en la web pero teniendo en cuenta que el sistema de búsqueda, sólo se recuperará algunos de esos relevantes
Relevancia	Característica de un documento recuperado que cumple con la necesidades de información, es decir, representa para el usuario los documentos que realmente le sirven, por ejemplo en la búsqueda de palabras que tienen más de un significado, se necesita recuperar los documentos relacionados con la información usuario requiere
Tasa de relevancia	coeficiente que surge de dividir el número de documentos relevantes recuperados, sobre el número total de documentos recuperados
Pertinencia	Es la cualidad que tiene el documento recuperado de adaptarse a las necesidades de información, con esto se refiere a lo adecuado que el documento es respecto a la necesidad de información o consulta del usuario
Tasa de pertinencia	coeficiente que surge de dividir el número de documentos pertinentes recuperados, sobre el número total de documentos recuperados
Precisión	Capacidad que tiene el sistema de búsqueda en coordinar la ecuación con los documentos más relevantes. De otra forma son aquellos documentos relevantes

	recuperados
Tasa de precisión	Coeficiente que surge de dividir el número de documentos relevantes recuperados, sobre el número total de documentos de la colección. En este caso no solo se tienen en cuenta los documentos relevantes de una colección, sino todos los documentos de ésta, a diferencia de la tasa de acierto.

Tabla 31. Criterios en la calidad de Recuperación de Información

3.3 HERRAMIENTAS PARA LA CONSTRUCCION DE PROCEDIMIENTOS

Como parte del diseño para la creación del procedimiento se cuenta con herramientas que facilitan la abstracción de los pasos necesarios en la creación de un índice semántico. Estas abstracciones permiten obtener la estructura, reglas, restricciones y objetos presentes en dicho procedimiento. Se describen a continuación varias herramientas utilizadas para definir procedimientos.

3.3.1 META-MODELOS Y ESTÁNDARES

Existen algunos estándares y meta-modelos útiles como el estándar BPMN, el IDEFØ y el meta-modelo SPEM para la creación de métodos, procesos y procedimientos, los cuales apoyan el modelado y diseño del objetivo. Un meta-modelo de procesos “describe un conjunto de conceptos genéricos y sus interrelaciones” [23], los cuales son básicos en la definición de modelos de procesos.

En la implementación de un proceso o metodología, SPEM diferencia dos etapas, las cuales sugieren una separación entre método y proceso. En la **¡Error! No se encuentra el origen de la referencia.** se observan las diferencias.

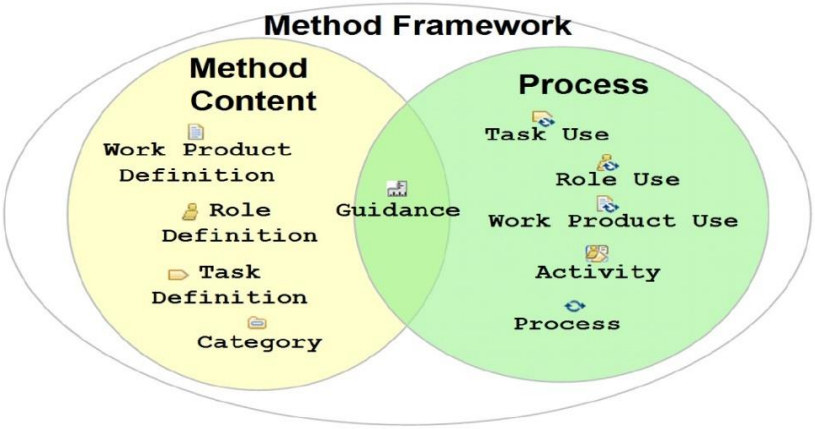


Figura 5. **Separación de proceso y método [24]**

Contenido de un proceso:

- ProcessPackage: Representa un paquete con todos los elementos del proceso
- Activity: Representa a las actividades que se ejecutan dentro del proceso, y los tareas, productos, roles asociados.
- WorkProductUse: Representa un producto de trabajo de entrada o salida, relacionado con una actividad o tarea
- RoleUse: Representa al rol (persona) que lleva a cabo una tarea o actividad dentro del proceso
- TaskUse: Representa una tarea atómica dentro de una actividad

3.3.2 DIAGRAMA DE FLUJO

Los diagramas de flujo (o flujo-gramas) son gráficas que representan la secuencia de operaciones, pasos, etapas y/o actividades que ocurren durante un proceso. Se utilizan símbolos gráficos para su representación. Puede incluir, además, la información que se considera deseable para el análisis, por ejemplo el tiempo necesario entre las etapas [25]. Favorecen la comprensión del proceso pues se define el inicio, el final, las relaciones entre las actividades y los puntos de decisión durante el proceso [26].
Los símbolos más comunes de un diagrama de flujo son:

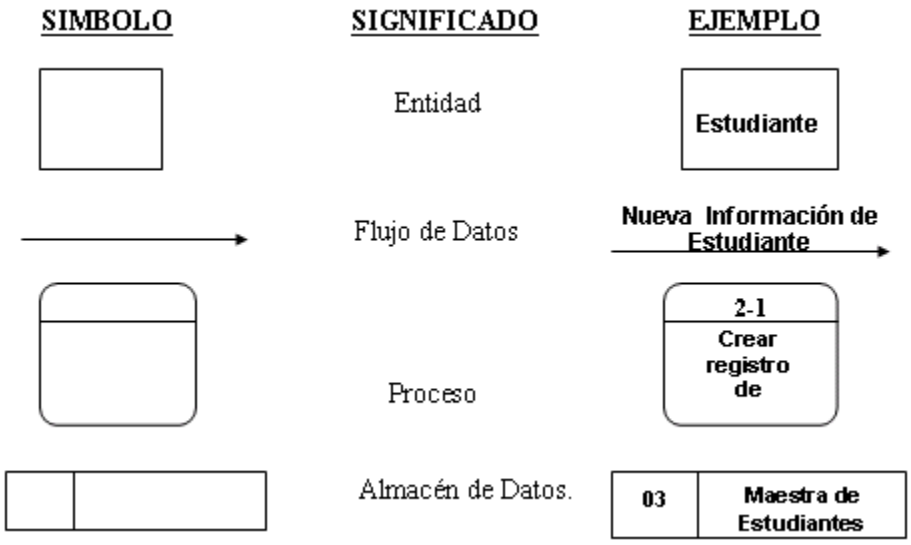


Figura 6. Símbolos más usados en el Diagrama de flujo[14]

3.3.3 DIAGRAMA DE ACTIVIDAD

El Diagrama de Actividad es un diagrama de flujo del proceso multi-propósito que hace parte del estándar UML (*Unified Modeling Language*)[15], el cual, se usa para modelar el comportamiento del sistema y también para modelar un Caso de Uso, o una clase, o un método complejo[16].

Un diagrama de actividades representa los flujos de trabajo, por pasos, del negocio y operaciones del sistema, en otras palabras, muestra el flujo de control general. Se utiliza para modelar un proceso de flujo de trabajo y/o secuencias de acciones y condiciones dentro de un proceso [27].

Los diagramas de actividades se componen principalmente de los elementos Nodos, lo cuales se pueden clasificar en [28]:

- Nodos de Acción: Reciben y pasan el control y datos a otras acciones. A su vez se diferencian en:

[14] Extraído de http://www.monografias.com/trabajos60/diagrama-flujo-datos/diagrama-flujo-datos2.shtml.

[15] Es un lenguaje gráfico para visualizar, especificar, construir y documentar un sistema. UML ofrece un estándar para describir un "plano" del sistema (modelo), incluyendo aspectos conceptuales tales como procesos de negocio y funciones del sistema. Extraído de http://www.grupoinformatica.com/biblioteca-articulos/1459-uml-lenguaje-unificado-de-modelado.html.

[16] Extraído de http://www.ibiblio.org/pub/linux/docs/LuCaS/Tutoriales/doc-modelado-sistemas-UML/multiple-html/x291.html

- Actividades: Agrupaciones de acciones que pueden poseer pre y post condición además de parámetros de entrada o de salida.
- Acción: Son consumidores/productores y reciben datos y el control de flujo para transferir estos elementos a otras acciones.

- Nodos de Control: Distribuyen el control de la ejecución y los tokens (acciones) a lo largo del diagrama.

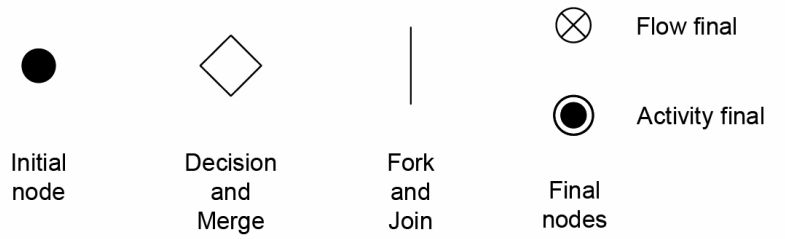


Figura 7. Nodos de control en diagramas de actividad [28].

- Nodos Objeto: Contienen datos de manera temporal a la espera de mover estos datos a lo largo del diagrama.

Otros elementos con los que se cuenta son:

- Flujos. Se diferencias dos tipos:
  - El flujo de control: para modelar el paso de una acción a otra.
  - El flujo de datos: para modelar el paso de información de una acción a otra.

Particiones.
Regiones de Expansión.
Excepciones.
Regiones de Actividad Interrumpibles.
Streaming.

3.3.4 MAPA CONCEPTUAL Y MENTAL

Los mapas conceptuales son diagramas (bidimensionales) que muestran relaciones significativas entre conceptos en determinado ámbito o área de estudio y se expresan como proposiciones. Una proposición consta de dos o más términos conceptuales unidos por palabras (palabras de enlace) para formar una unidad semántica. Las herramientas utilizadas en la realización de mapas conceptuales, permiten estructurar un conjunto de conceptos y/o ideas, además establecer relaciones entre ellos generando nuevos conceptos más complejos. [29].

El mapa mental es un diagrama que permite organizar, representar y analizar la información sobre un ámbito, con el propósito de facilitar los procesos de aprendizaje, administración y planeación organizacional así como la toma de decisiones. Mediante estas herramientas se pueden representar nuestras ideas utilizando de manera armónica las funciones cognitivas de los hemisferios cerebrales [30]. Facilitan además, la organización lógica y estructurada de los contenidos de aprendizaje, ya que permiten seleccionar, extraer y separar la información significativa o importante de la información poco útil dependiendo del área de investigación, y a su vez, se pueden insertar nuevos conceptos en la propia estructura de conocimiento [31].

Por lo general, el mapa conceptual consta de un concepto o palabra central rodeada por un grupo de ideas a las que hace referencia dicha palabra o concepto. Si se repite el proceso para cada una de estas ideas se estará construyendo un árbol de ideas derivadas de la central. Estas ideas pueden ser expresadas también a través de imágenes o dibujos que favorecen el impacto visual. La representación de un mapa conceptual se puede realizar mediante un grafo, en el cual, los nodos son los conceptos y las líneas las relaciones entre ellos. Las líneas pueden tener asociadas palabras clave que establecen el tipo de relación que une los conceptos.

Los mapas conceptuales pueden representarse en forma de estrella, con un concepto o palabra central que irradia otros conceptos relacionados (también llamado mapa mental), o en forma de árbol invertido, con el nodo raíz en la parte superior (concepto principal) y el resto de forma jerárquica descendente.

Existen varias herramientas para construir mapas mentales y conceptuales, una de las más utilizadas es FreeMind.

FreeMind [32] es una herramienta para la elaboración y manipulación de mapas conceptuales. Permite organizar y estructurar las ideas, los conceptos, su relación entre ellos y su evolución. Puede ser utilizada para cualquier área o ámbito de investigación como mecanismo o forma de plasmar lluvia de ideas para su posterior reutilización.

En los gráficos de FreeMind o mapas conceptuales se pueden incorporar todo tipo de informaciones que pueden provenir de distintas fuentes. Estas informaciones se pueden navegar por medio de una serie de nodos que se pueden expandir o contraer para mostrar dependencias de unas ideas con otras. Cada nodo permite adjuntar otros recursos como imágenes, iconos, direcciones web, etc. [33].

En la Figura 8. Mapa Conceptual en FreeMind**¡Error! No se encuentra el origen de la referencia.** se observa un ejemplo de mapa conceptual creado en la herramienta FreeMind.

Figura 8. Mapa Conceptual en FreeMind

4 ANEXO D

4.1 ELECCION DE LA ONTOLOGIA

En la actual red podemos encontrar ontologías en diferentes dominios médicos, científicos, informáticos, biomédica, educativos y demás. Cada una de estas nos brindan ventajas a la hora de trabajar en nuestros proyectos. En el caso de nuestra aplicación fue necesaria una ontología educativa que se ajuste a las necesidades del proyecto. Para realizar el proceso de selección de la ontología, se realizaron diferentes búsquedas de ontologías educativas, que manejaran el dominio de las ciencias naturales, encontrando pocas ontologías que manejaran este tema.
A continuación se describen algunas de las ontologías encontradas y sus características.

GENE ONTOLOGY (GO)
The Gene Ontology [34] es una iniciativa en bioinformática, con el fin de estandarizar la representación de los genes y productos genéticos entre especies y atributos de las bases de datos. Este proyecto proporciona un vocabulario controlado de términos para describir las características de productos de genes y también ofrece herramientas para acceder y procesar datos.
Describe tres ontologías independientes:

- Funciones moleculares.
- Procesos biológicos.
- Componentes celulares.

GO en uso
Maneja 16500 términos los cuales están asociados a una base de datos de más de 120000 genes de cerca de 20 organismos.

- Se pueden buscar las proteínas asociadas a uno o más de GO ids.
- Se pueden buscar las proteínas asociadas a un determinado termino
- Todos los términos asociados con una proteína.

GO browsers
Permite buscar términos en GO asociados a productos genéticos. Se encuentran diferentes links a varias bases de datos como secuencia, organismos específicos, etc.

PLANTONTOLOGY
El principal objetivo de PlantOntology [35] es desarrollar vocabularios controlados (ontologías) que describen las estructuras de las plantas así como su crecimiento y etapas de desarrollo, proporcionando un marco de semántico de las consultas a través de especies significativas entre las bases de datos. Esta ontología se ha desarrollado y ha mantenido el objetivo de facilitar y dar cabida a los esfuerzos de la anotación funcional en las bases de datos de la planta y por la comunidad de investigación de plantas en general.

PlantOntology no es una colección extensa de términos botánicos, sino más bien una estructura jerárquica compleja en la que los conceptos botánicos son descritos por su significado y por la relación entre sí. Su objetivo es facilitar las consultas de las bases de datos cruzadas y fomentar el uso constante de estos vocabularios en la anotación de tejido y la etapa especifica de crecimiento de los genes, las proteínas y los fenotipos.

ONTOLOGÍA PARA EL MEDIO AMBIENTE [36]
Ontología que maneja términos del medio ambiente y los diferentes hábitats. Esta ontología apoya la anotación del medio ambiente de cualquier organismo o muestra biológica.

SCIENCE ENVIRONMENT FOR ECOLOGY KNOWLEDGE (SEEK)[37]
La ciencia para el medio ambiente ecológico del conocimiento es un sistema diseñado para facilitar no solo la adquisición de datos y archivos sino también la integración, trasformación, análisis y síntesis de datos ecológicos y biodiversidad que antes eran intratables.

Los objetivos de **SEEK** son:

- Ganancia global de acceso a los datos ecológicos y la información
- Localizar y utilizar los servicios distribuidos de computo y
- Alcance de nuevos métodos para capturar, reproducir y analizar los datos mediante la aplicación de la biodiversidad ecológica y análisis y de investigación.

Después de revisar cada una de ellas se opto por escoger Plant Ontology, la cual maneja todo los conceptos relacionados a plantas (botánica).

5 ANEXO E

5.1 DIAGRAMA DE CASOS DE USO

5.1.1 CASOS DE USO FORMATO EXPANDIDO

A continuación se describen los diagramas de casos de uso en formato expandido del sistema.

Caso de uso: Obtener Ayuda

Nombre del Caso de Uso:	Obtener Ayuda
Actores:	Usuario
Propósito:	Brindar al usuario una guía para utilizar la herramienta web en caso de presentarse alguna duda.
Resumen:	Una guía para la utilización del sitio.
Tipo:	Primario.
CURSO NORMAL DE LOS EVENTOS	
Acción del Actor	**Respuesta del Sistema**
1. El usuario solicita ayuda	
	1. La página carga una guía que le servirá de ayuda al usuario a la hora de agregar la pagina web.

Tabla 32. Caso de uso Obtener Ayuda

Caso de uso: Realizar Consulta

Nombre del Caso de Uso:	Realizar Consulta
Actores:	Usuario
Propósito:	Permite al usuario realizar una consulta de un dominio en particular, teniendo en cuenta la ontología utilizada en el sitio Web.
Resumen:	Se presenta un espacio para digitar una consulta por parte del usuario sobre un tema particular, obteniendo una lista de resultados de acuerdo a sus necesidades de información.
Tipo:	Primario.
CURSO NORMAL DE LOS EVENTOS	
Acción del Actor	**Respuesta del Sistema**
1. El usuario digita una consulta	
	2. el sistema toma la consulta digitada por el usuario y la procesa.
	3. El sistema muestra una lista de resultados de acuerdo a la consulta realizada por el usuario.

Tabla 33. Caso de uso Realizar Consulta

5.1.2 CASOS DE USO REALES

A continuación se describen los casos de uso reales del sistema: Obtener Ayuda y Realizar Consulta.

Caso de uso: Obtener Ayuda

Figura 9. Caso de uso real: Obtener Ayuda

Caso de uso: Realizar Consulta

Figura 10. Caso de uso real. Realizar Consulta

5.2 DIAGRAMAS DE INTERACCIÓN

Los diagramas de interacción describen en detalle las acciones de los actores (usuario) y el sistema. A continuación se muestran los diagramas de interacción para el usuario de la aplicación.

Obtener Ayuda

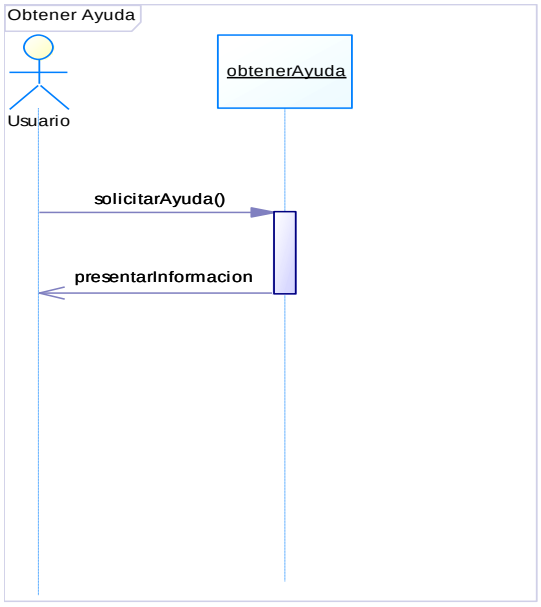


Figura 11. Diagrama de interacción Obtener Ayuda

Realizar Consulta

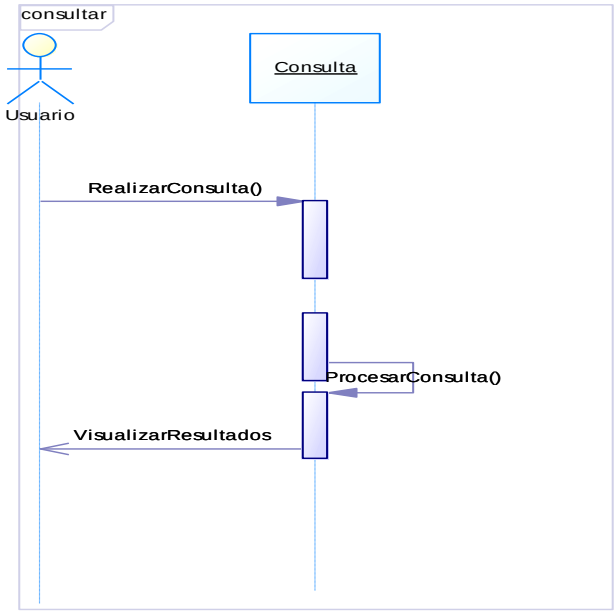


Figura 12. Diagrama de interacción Realizar Consulta

5.3 DIAGRAMA DE CLASE

Los diagramas de clase que se muestran a continuación se refieren a las clases creadas por otros autores y utilizadas como recurso para la creación de las nuevas clases del sistema.

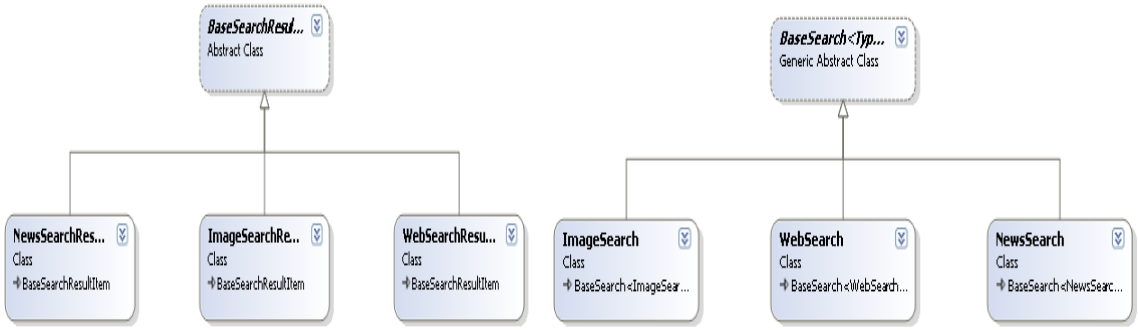


Figura 13. Diagrama de clases utilizadas

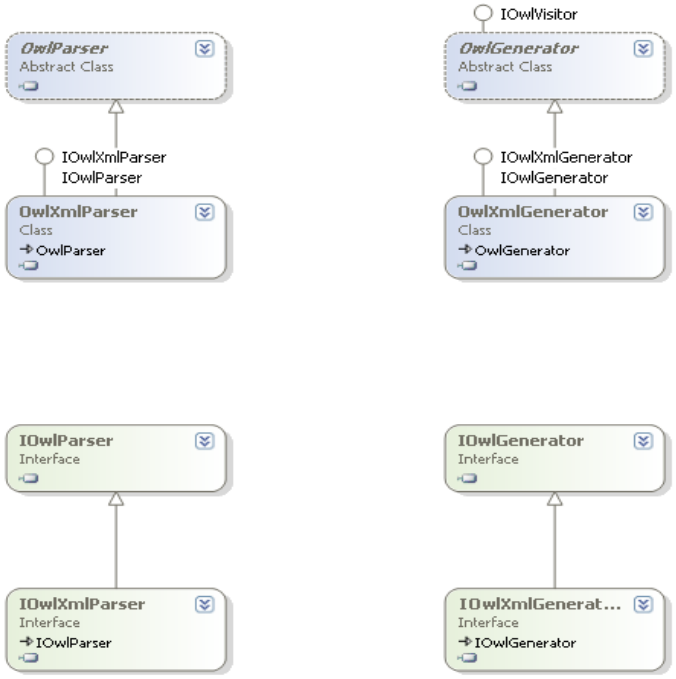


Figura 14. Clases extraídas de otros recursos

6 ANEXO F

6.1 ARQUITECTURA DE LA APLICACIÓN

A continuación se describen los patrones utilizados en la construcción del prototipo software.

Patrones utilizados
Los patrones de diseño[38] son la base para la búsqueda de soluciones a problemas comunes en el desarrollo de software con un diseño orientado a objetos. También se utilizan en ámbitos referentes al diseño de interacción o interfaces. A continuación se presenta una breve descripción de los patrones que fueron utilizados en el desarrollo del proyecto.

Patrón Singleton: es uno de los patrones más sencillos que existe. Tarata las situaciones en las que solo se permite una instancia de una clase dada. Este patrón proporciona la restricción para la única instancia de un a clase o valor de un tipo a un solo objeto, permite crear un solo objeto de la clase para evitar la redundancia de los objetos, en cualquier momento de la aplicación[17].

Patrón fachada: trata de simplificar la interface entre dos sistemas o componentes software ocultando un sistema complejo detrás de una clase que hace las veces de pantalla o fachada. Su objetivo principal es ocultar la complejidad de un sistema, el conjunto de clases o componentes que lo conforman, de modo que solo se ofrezca un punto de entrada al sistema tapado por la fachada. Una de las ventajas de usar el patrón fechada para comunicar las dos partes o componentes, es la de aislar los posibles cambios que se puedan producir en alguna de las partes[18].

Patrón Code Behind: el patrón utilizado en la plataforma .NET es el patrón Code- Behind el cual separa la interfaz grafica de la lógica, dividiendo en compilación el tiempo de ejecución y compilación previa. Este patrón permite dividir la interfaz del usuario con la definición de los objetos, además tiene una clase que desarrolla la lógica de negocio de cada control o formulario.

Proceso de construcción del prototipo para la creación de índices semánticos
Luego de haber planteado el proceso correspondiente para la creación de índices semánticas, a continuación se presentara una descripción de cómo se abordo cada uno de los pasos de que componen el procedimiento, con el fin de llevar a cabo la implementación del prototipo.

Para la realización de este se llevaron a cabo los siguientes pasos:
El dominio seleccionado para la aplicación de el procedimiento basado en ontologías de dominio, corresponde a dominio Educativo, en especial al área de las ciencias naturales para básica primaria.

[17] Extraído de http://www.programacion.com/articulo/diseno_de_software_con_patrones_114
[18] Extraído de http://www.programacion.com/articulo/diseno_de_software_con_patrones_parte_4_145#joa_patrones3_fachada

7 ANEXO G

7.1 PRUEBA DE USABILIDAD

Para la evaluación final del prototipo creado, se evaluó la usabilidad del sistema mediante una prueba realizada a estudiantes de cuarto y quinto grado (básica primaria) del Colegio Campestre Americano de la ciudad. La prueba se basa en una serie de preguntas para las cuales respondieron cada uno, según su criterio, cómo les parecía el sistema de acuerdo a su uso, visibilidad y demás factores de usabilidad, también teniendo en cuenta las observaciones y/o sugerencias que querían hacer al sistema.

Los factores que se tuvieron en cuenta, se muestran a continuación como en el documento entregado a los estudiantes de dicho colegio. Además, se consigna el número de personas que coincidieron en cada aspecto del grado de satisfacción (excelente, bueno, neutro, regular, deficiente) sobre el sistema.

Visibilidad del estado del sistema	Excelente	Bueno	Neutro	Regular	Deficiente
1.1. El sitio muestra claramente dónde se encuentra el usuario	5	3			
1.2. Los enlaces posibles de explorar están claramente señalados	5	2	1		

Relación entre sistema y mundo real	Excelente	Bueno	Neutro	Regular	Deficiente
2.1 El lenguaje es claro	7	1			
2.2. Los conceptos utilizados son entendibles	3	2	3		
2.3. Las palabras son de significado conocido	5	2	1		
2.4. Los iconos generan significado	6	2			

Consistencia y estándares	Excelente	Bueno	Neutro	Regular	Deficiente
3.1 Existe coherencia entre el nombre de un enlace y el sitio al que apunta	4	3	1		
3.2 Todos los enlaces tienen contenido	4	4			

Reconocer en lugar de recordar	Excelente	Bueno	Neutro	Regular	Deficiente
4.1 Los iconos son fácilmente reconocibles	3	2	2	1	
4.2. Los enlaces pueden identificarse claramente	4	2	2		
4.3. Es posible reconocer dónde se encuentra el usuario	3	1	3	1	

Recuperación de Información	Excelente	Bueno	Neutro	Regular	Deficiente
5.1 Encontró todos los conceptos (sobre botánica) buscados en las páginas retornadas.	4	1	3		
5.2. El buscador es fácil de usar.	7	1			

5.3. Los resultados arrojados cumplieron las expectativas de su búsqueda.	4	4			

Ayuda y documentación	Excelente	Bueno	Neutro	Regular	Deficiente
6.1 Existe algún tipo de ayuda o indicación en el sitio	5	2	1		
6.2. Cuando existe ayuda, ésta es específica	5	1	2		
6.3. La ayuda es asequible	6		2		

¿Cómo califica globalmente el sitio Web analizado?	Excelente	Bueno	Neutro	Regular	Deficiente
1. Visibilidad del estado del sistema	4	4			
2. Relación entre sistema y mundo real	4	2	2		
3. Control del usuario y libertad	4	2	1		1
4. Prevención de errores	5	1	1		1
5. Flexibilidad y eficiencia de uso	4	3	1		
6. Estética y diseño	4	3	1		
7. Ayuda y documentación	4	4			

Tabla 34. Evaluación de usabilidad de la aplicación

Al realizar la ponderación de respuestas se encontraron buenos resultados, pues la mayoría de personas estuvo de acuerdo en la calificación del sitio entre bueno y excelente, teniendo en cuenta aspectos como Visibilidad del estado y el sistema, recuperación de información y calificación global del sitio. A continuación se muestra en porcentajes los resultados obtenidos.

	Excelente	Bueno	Neutro	Regular	Deficiente
Pregunta 1.1	62,5	37,5	0	0	0
Pregunta 1.2	62,5	25	12,5	0	0
Pregunta 2.1	87,5	12,5	0	0	0
Pregunta 2.2	37,5	25	37,5	0	0
Pregunta 2.3	62,5	25	12,5	0	0
Pregunta 2.4	75	25	0	0	0
Pregunta 3.1	50	37,5	12,5	0	0
Pregunta 3.2	50	50	0	0	0
Pregunta 4.1	37,5	25	25	12,5	0
Pregunta 4.2	50	25	25	0	0
Pregunta 4.3	37,5	12,5	37,5	12,5	0
Pregunta 5.1	50	12,5	37,5	0	0
Pregunta 5.2	87,5	12,5	0	0	0
Pregunta 5.3	50	50	0	0	0
Pregunta 6.1	62,5	25	12,5	0	0
Pregunta 6.2	62,5	12,5	25	0	0
Pregunta 6.3	75	0	25	0	0
Pregunta I	50	50	0	0	0

Pregunta II	50	25	25	0	0
Pregunta III	50	25	12,5	0	12,5
Pregunta IV	62,5	12,5	12,5	0	12,5
Pregunta V	50	37,5	12,5	0	0
Pregunta VI	50	37,5	12,5	0	0
Pregunta VII	50	50	0	0	0

Tabla 35. Porcentaje de resultados para cada pregunta de usabilidad.

Como se observa, los principales porcentajes según la tabla, se encuentran en las calificaciones de bueno y excelente. Esto permite observar una buena apreciación en cuanto a la usabilidad de la aplicación y relevancia, en general, de los resultados obtenidos de algunas consultas.

8 ANEXO H

8.1 VALIDACION DEL PROTOTIPO: PRECISION-RECUERDO, INDICE MAP

A continuación se describen los resultados de la curva de precisión-recuerdo para las cinco consultas analizadas. Posteriormente se calcula el Indice MAP, promedio de la precisión encontrada.

8.2 CURVA DE PRECISIÓN-RECUERDO

La curva Precision-Recall es una medida utilizada en los sistemas de recuperación de informacion como los motores de búsqueda, que permite evaluar la eficacia [21] de resultados respecto a varias consultas realizadas por el usuario. Para ello se tiene en cuenta la cantidad de documentos recuperados y la cantidad de documentos recuperados que son relevantes de acuerdo a una consulta. La precisión y recall (recuerdo) se expresan de la siguiente manera.

8.2.1.1Average Precision (Precisión Promedio)

Promedio de los valores de precisión en los puntos en que se recupera cada documento relevante [39].

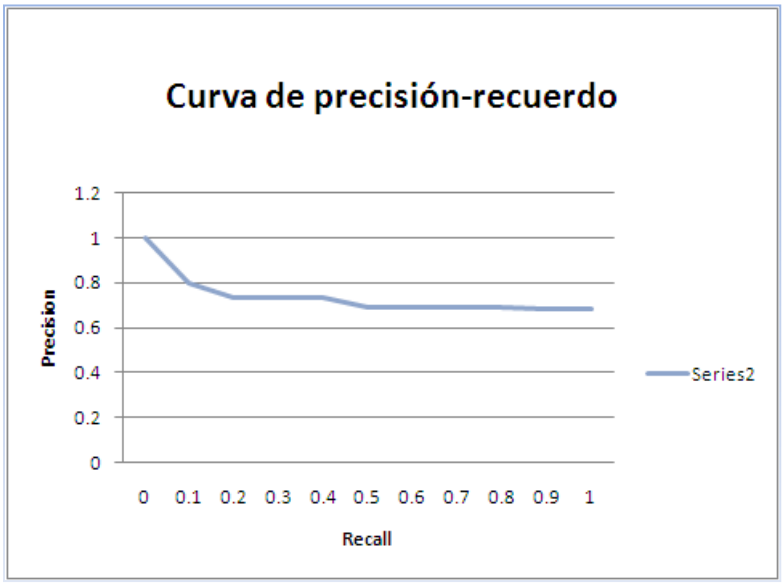
Para realizar las pruebas se tomaron 5 conceptos y se mide para cada uno, la precisión y recuerdo en las URLs retornadas. A continuación se realizan consultas y se calcula la curva de precisión-recall. Posteriormente se promedia la precisión, la cual será necesaria para el cálculo del índice MAP (Véase **¡Error! No se encuentra el origen de la referencia.**).

- **Concepto: FLOWER**

N° de documentos	Relevante	Recall	Precisión	N° de documentos	Relevante	Recall	Precisión
1	X	R=1/34=0.029	P= 1/1 =1	26	X	R=18/34=0.529	P=18/26=0.69
2	X	R =2/34 =0.058	P = 2/2= 1	27			
3	X	R =3/34 =0.088	P = 3/3=1	28	X	R=19/34=0.558	P=19/28=0.678
4				29	X	R=20/34=0.588	P=20/29=0.689
5	X	R = 4/34=0.117	P = 4/5=0.8	30			
6	X	R=5/34=0.147	P = 5/6=0.833	31	X	R=21/34=0.617	P=21/31=0.677
7				32			
8	X	R=6/34=0.176	P = 6/8=0.75	33	X	R=22/34=0.647	P=22/33=0.666
9	X	R=7/34=0.205	P = 7/9=0.777	34			
10	X	R=8/34=0.235	P = 8/10=0.8	35	X	R=23/34=0.676	P=23/35=0.657
11				36	X	R=24/34=0.705	P=24/36=0.666
12	X	R=9/34=0.264	P = 9/12=0.75	37			
13	X	R=10/34=0.294	P = 10/13=0.76	38	X	R=25/34=0.735	P=25/38=0.657
14				39	X	R=26/34=0.764	P=26/39=0.666
15	X	R=11/34=0.323	P=11/15=0.733	40	X	R=27/34=0.794	P=27/40=0.675
16				41			
17	X	R=12/34=0.352	P=12/17=0.70	42	X	R=28/34=0.823	P=28/42=0.666
18				43	X	R=29/34=0.852	P=29/43=0.674
19	X	R=13/34=0.382	P=13/19=0.684	44			
20				45	X	R=30/34=0.882	P=30/45=0.666

21	X	R=14/34=0.411	P=14/21=0.666	46	X	R=31/34=0.911	P=31/46=0.673
22	X	R=15/34=0.441	P=15/22=0.681	47	X	R=32/34=0.941	P=32/47=0.680
23				48			
24	X	R=16/34=0.470	P=16/24=0.666	49	X	R=33/34=0.970	P=33/49=0.673
25	X	R=17/34=0.5	P=17/25=0.68	50	X	R=34/34=1	P=34/50=0.68

Tabla 36. Precisión y Recall para "flower"



Recall	Precision
0	1
0.1	0.8
0.2	0.733
0.3	0.733
0.4	0.733
0.5	0.689
0.6	0.689
0.7	0.689
0.8	0.689
0.9	0.68
1	0.68

Figura 15. Curva precisión-Recall para "flower"

Interpretación de resultados

Al revisar los resultados que presenta la curva Precisión – Recall, se observa una disminución leve en la precisión a medida que el Recall aumenta, lo cual es una medida esperada, puesto que al incrementarse el número de documentos recuperados, la precisión puede disminuir un poco. Sin embargo, a pesar del aumento de estos documentos, la precisión se mantiene en buen porcentaje, es decir, no disminuye significativamente. Esto también se debe a la cantidad de resultados que contienen el concepto buscado, lo cual permite que la diferencia de precisión no sea tan amplia respecto al Recall. La curva semi-constante para la precisión y el Recall permite inferir que la relevancia es buena.

Precisión Promedio de consulta "flower"

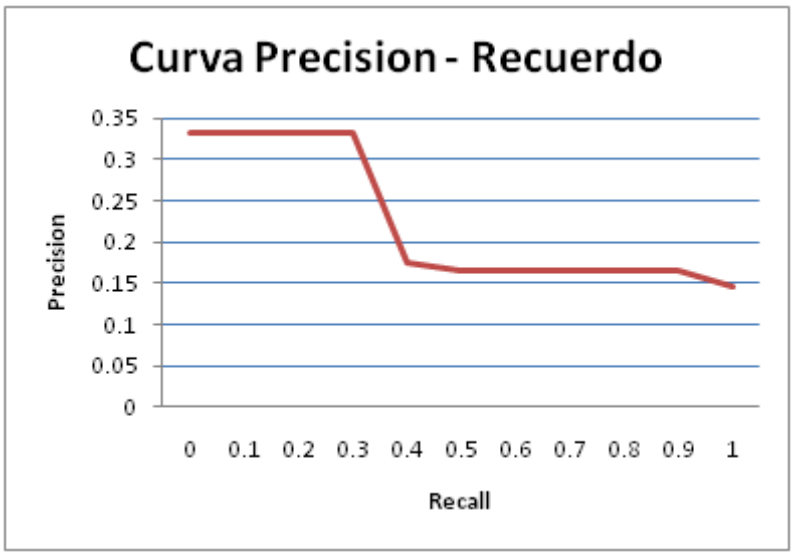
#documentos relevantes = 34
Average precision = (1+1+1+ 0.8 +0.833+0.75+0.77+0.8+0.75…) /34 = **0.686**

- **Concepto: SEEDLING**

N° de documentos	Relevante	Recall	Precision	N° de documentos	Relevante	Recall	Precisión
1				26			
2				27			

3				28			
4	X	R=1/7=0.142	P = 1/4=0.25	29			
5				30			
6	X	R=2/7=0.285	P = 2/6=0.33	31	X	R=5/7=0.714	P=5/31=0.161
7				32			
8				33			
9				34			
10				35			
11				36	X	R=6/7=0.857	P=6/36=0.166
12				37			
13				38			
14				39			
15				40			
16				41			
17	X	R=3/7=0.428	P=3/17=0.176	42			
18				43			
19				44			
20				45			
21				46			
22				47			
23				48	X	R=7/7=1	P=7/48=0.145
24				49			
25	X	R=4/7=0.571	P=4/25=0.16	50			

Tabla 37. Precisión y Recall para "seedling"



Recall	Precision
0	0.333
0.1	0.333
0.2	0.333
0.3	0.333
0.4	0.176
0.5	0.166
0.6	0.166
0.7	0.166
0.8	0.166
0.9	0.166
1	0.145

Figura 16. Curva precisión-Recall para "seedling"

Interpretación de resultados

La cantidad de documentos relevantes para la búsqueda, es relativamente poca en comparación con los documentos recuperados (50), esto ocasiona una disminución significativa en la precisión, a partir del tercer documento relevante. Sin embargo, se observa una precisión constante respecto al recall, lo cual, infiere resultados relativamente satisfactorios a pesar del incremento en la cantidad de documentos recuperados.

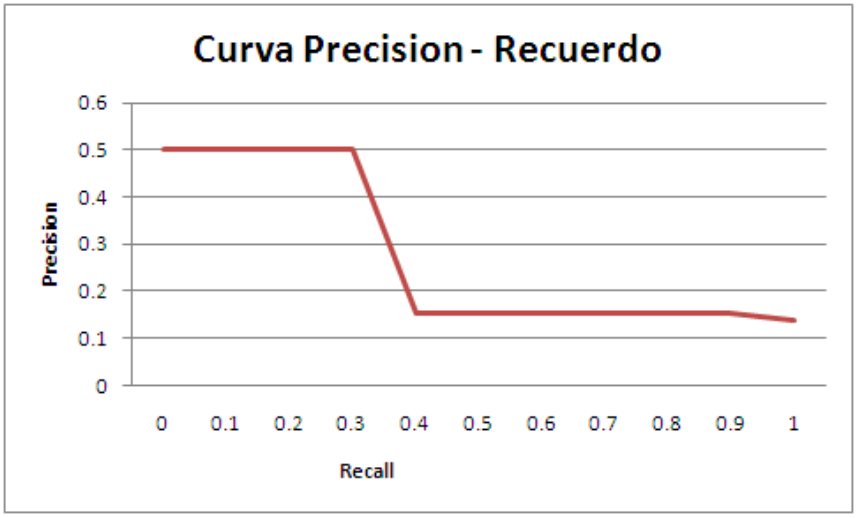
Precisión Promedio de consulta "seedling"

#documentos relevantes = 7
Average precision = (0.25 +0.333 +0.176+0.16+0.161+0.16+0.145) /7 = **0.201**

- **Concepto: PLANT STRUCTURE**

N° de documentos	Relevante	Recall	Precision	N° de documentos	Relevante	Recall	Precisión
1				26	X	R=4/5=0.8	P=4/26=0.153
2	X	R=1/5=0.2	P = 1/2=0.5	27			
3				28			
4				29			
5				30			
6				31			
7				32			
8				33			
9				34			
10				35			
11				36	X	R=5/5=1	P=5/36=0.138
12				37			
13				38			
14				39			
15				40			
16				41			
17	X	R=2/5=0.4	P=2/17=0.117	42			
18				43			
19				44			
20				45			
21				46			
22				47			
23				48			
24				49			
25	X	R=3/5=0.6	P=3/25=0.12	50			

Tabla 38. Precisión y Recall para "Plant structure"



Recall	Precision
0	0.5
0.1	0.5
0.2	0.5
0.3	0.153
0.4	0.153
0.5	0.153
0.6	0.153
0.7	0.153
0.8	0.153
0.9	0.153
1	0.138

Figura 17. Curva precisión-Recall para "Plant structure"

Interpretación de resultados

La precisión de los resultados disminuye bastante a partir del segundo documento relevante. Esto se explica en los pocos documentos relevantes para esta búsqueda. Teniendo en cuenta este resultado, se puede inferir que la precisión de los resultados disminuyó un poco por el decremento drástico en la relevancia. Sin embargo, no presenta una disminución significativa al incrementarse el número de documentos.

Precisión Promedio de consulta "plant structure"

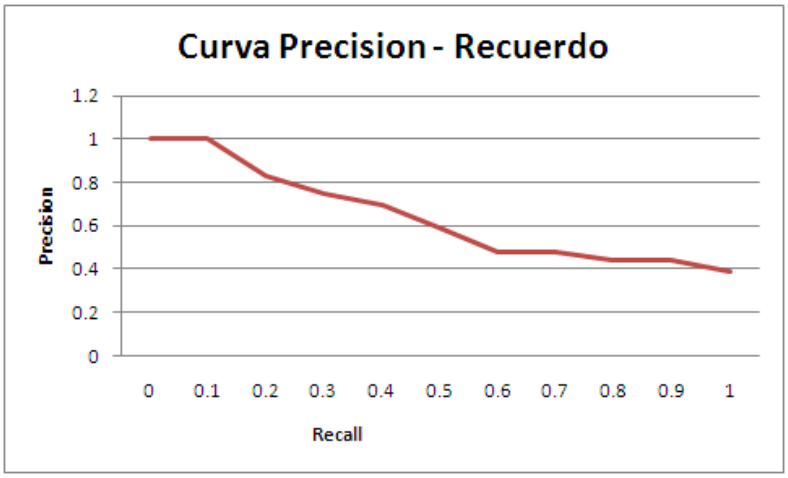
#documentos relevantes = 5
Average precision = (0.5+0.117+0.12++0.153+0.138) /5 = **0.205**

- **Concepto: SEED**

N° de documentos	Relevante	Recall	Precision	N° de documentos	Relevante	Recall	Precisión
1	X	R=1/20=0.05	P = 1/1=1	26			
2				27			
3	X	R=2/20=0.1	P = 2/3=0.666	28			
4	X	R=3/20=0.15	P = 3/4=0.75	29			
5	X	R=4/20=0.2	P = 4/5=0.8	30			
6	X	R=5/20=0.25	P = 5/6=0.83	31	X	R=13/20=0.65	P=13/31=0.419
7				32			
8	X	R=6/20=0.3	P = 6/8=0.75	33	X	R=14/20=0.7	P=14/33=0.424
9				34			
10	X	R=7/20=0.35	P = 7/10=0.7	35	X	R=15/20=0.75	P=15/35=0.428
11				36	X	R=16/20=0.8	P=16/36=0.444
12	X	R=8/20=0.4	P = 8/12=0.666	37			
13	X	R=9/20=0.45	P = 9/13=0.692	38			
14				39			
15				40			
16				41			

17	X	R=10/20=0.5	P=10/17=0.588	42			
18				43			
19				44			
20				45			
21				46	X	R=17/20=0.85	P=17/46=0.369
22	X	R=11/20=0.55	P = 11/22=0.5	47			
23				48			
24				49	X	R=18/20=0.9	P=18/49=0.367
25	X	R=12/20=0.6	P=12/25=0.48	50	X	R=19/20=0.95	P=19/50=0.38
				51	X	R=20/20=1	P=20/51=0.392

Tabla 39. Precisión y recall para "seed"



Recall	Precision
0	1
0.1	1
0.2	0.83
0.3	0.75
0.4	0.692
0.5	0.588
0.6	0.48
0.7	0.48
0.8	0.444
0.9	0.444
1	0.392

Figura 18. Curva precisión-recall para "seed"

Interpretación de resultados

La cantidad de documentos relevantes para esta búsqueda es un poco alta y por lo tanto la precisión y el Recall son relativamente altos. A pesar del incremento en el número de documentos recuperados (Recall), la precisión no disminuye de manera significativa, lo cual, es un buen indicador para el índice semántico.

Precisión Promedio de consulta "seed"

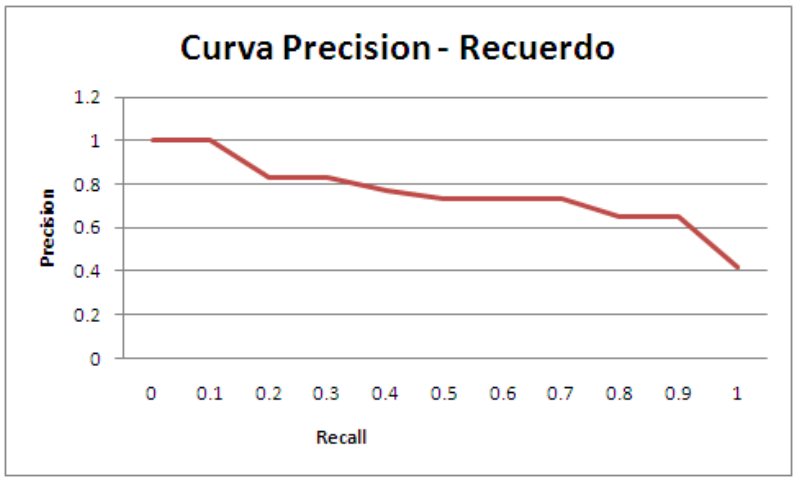
#documentos relevantes = 20
Average precision = (1+0.66+0.75+0.8+0.83+0.75+0.7+0.666+0.692+0.419…) /20
= **0.5822**

- **Concepto: LEAF**

N° de documentos	Relevante	Recall	Precision	N° de documentos	Relevante	Recall	Precisión
1	X	R=1/21=0.047	P = 1/1=1	26	X	R=17/21=0.809	P=17/26=0.653
2	X	R=2/21=0.095	P = 2/2=1	27			
3	X	R=3/21=0.142	P = 3/3=1	28			
4	X	R=4/21=0.190	P = 4/4=1	29			
5				30			

6	X	R=5/21=0.238	P = 5/6=0.833	31	X	R=18/21=0.857	P=18/31=0.580
7				32			
8	X	R=6/21=0.285	P = 6/8=0.75	33			
9				34			
10	X	R=7/21=0.333	P = 7/10=0.7	35			
11	X	R=8/21=0.380	P = 8/11=0.727	36			
12	X	R=9/21=0.428	P = 9/12=0.75	37			
13	X	R=10/21=0.476	P=10/13=0.769	38			
14				39			
15	X	R=11/21=0.523	P=11/15=0.733	40			
16				41			
17	X	R=12/21=0.571	P=12/17=0.705	42			
18				43			
19				44			
20	X	R=13/21=0.619	P=13/20=0.65	45			
21				46	X	R=19/21=0.904	P=19/46=0.413
22	X	R=14/21=0.666	P=14/22=0.636	47			
23				48			
24	X	R=15/21=0.714	P=15/25=0.6	49	X	R=20/21=0.952	P=20/49=0.408
25	X	R=16/21=0.761	P=16/25=0.64	50	X	R=21/21=1	P=21/50=0.42
				51			

Tabla 40. Precisión y recall para "leaf"



Recall	Precision
0	1
0.1	1
0.2	0.833
0.3	0.833
0.4	0.769
0.5	0.733
0.6	0.733
0.7	0.733
0.8	0.653
0.9	0.653
1	0.42

Figura 19. Curva precisión-Recall para "leaf"

Interpretación de resultados

La mayoría de documentos relevantes para esta búsqueda son los primeros recuperados del total de documentos, lo cual, ofrece una buena precisión respecto al recall y la curva muestra relevancia en los resultados a pesar del incremento en la cantidad de documentos recuperados.

Precisión Promedio de consulta "leaf"

#documentos relevantes = 21
Average precision = (1+1+1+1+0.833+0.75+0.7+0.727…) /21 = **0.712**

8.3 PRUEBA PARA ESTADISTICAS KAPPA

La validación del prototipo construido se realizo también mediante estadísticas kappa, las cuales muestran la relevancia de los resultados según los estudiantes y/o docentes que evaluaron la aplicación. Para ello, se realizaron formatos con una consulta cada uno y las respectivas urls que arroja la búsqueda. Esto con el fin de evaluar si es o no relevante para los usuarios, cada página revisada. El formato utilizado se muestra en Tabla 41, la cual se presento en el colegio Campestre Americano y en la Institución Educativa Alejandro de Humboldt, sede Yanaconas.

Las consultas realizas fueron: Embryo, Seed coat, Plant cell, Xylem, Cuticle, Flower, Seed y Fruit.

PRUEBAS DE RELEVANCIA
Nombre del Evaluador: _____
Año que cursa: _____

Consulta: **seed coat** *(esta consulta varia para cada par de jueces)*	¿Encuentra relevancia en los resultados?	
Urls retornadas para la consulta *(varían según la consulta)*	**Si**	**No**
http://www.buzzle.com/articles/plants/		
http://www.abcteach.com/directory/basics/science/plants/		
http://www.answers.com/topic/plant		
http://en.wikipedia.org/wiki/Plant		
http://www.tms.riverview.wednet.edu/lrc/plants.htm		
http://www.edhelper.com/plants.htm		
http://www.enchantedlearning.com/themes/plants.shtml		
http://www.theteachersguide.com/plantsflowers.htm		
http://www.enchantedlearning.com/subjects/plants/plant/		
http://www.newworldencyclopedia.org/entry/Plant		
http://www.neok12.com/Plants.htm		

Tabla 41. Formato de evaluación de relevancia

Con las evaluaciones realizadas se procedió al cálculo de las estadísticas kappa correspondientes.

8.3.1 Cálculos Kappa Colegio Campestre Americano

Para el cálculo de las estadísticas Kappa, primero se toman los jueces con sus respectivas consultas y por cada prueba se realiza el cálculo.

La relevancia de los documentos encontrada por el Juez 1 se consigna en filas, la del Juez 2 se muestra en las columnas. El total de urls revisadas por todos es 47.

Consultas: Embryo y Seed coat

		Relevancias de Juez 2		
		SI	NO	TOTAL
Relevancias de Juez 1	SI	7	0	7
	NO	0	11	11
	TOTAL	7	11	18

Tabla 42. Relevancias prueba 1 Campestre Americano

A continuación, se realiza el cálculo de P(A) y P(E).
Proporción observada de las veces en que los jueces estuvieron de acuerdo

$$P(A) = \frac{7+11}{18} = 1$$

$$P(no-relevante) = \frac{11+11}{18+18} = 0.611$$

$$P(relevantes) = \frac{7+7}{18+18} = 0.388$$

$\boldsymbol{Probalidad\ de\ acuerdo\ entre\ dos\ jueces}$ $(\boldsymbol{por\ azar})$

$$P(E) = P(no-relevantes)^2 + P(relevantes)^2 = 0.373 + 0.150 = 0.523$$

$\boldsymbol{Estadistica\ Kappa}$

$$K = \frac{(P(A)-P(E))}{(1-P(E))} = \frac{1-0.524}{(1-0.524)} = \frac{0.476}{0.476} = 1$$

Al obtener k = 1 indica un acuerdo total entre los jueces que revisaron los documentos. Cabe mencionar que se observo en los jueces la toma de decisiones llevados por la exactitud al encontrar las palabras claves dentro de los documentos, pero no por la relevancia en cuanto a las relaciones semánticas en el contexto del mismo. Como se explicó anteriormente, el dominio del idioma (intermedio), fue uno de los inconvenientes para una buena revisión de los documentos.

Consultas: Plant Cell y Xylem

		Relevancias de Juez 2		
		SI	NO	TOTAL
Relevancias de Juez 1	SI	5	9	14
	NO	0	5	5
	TOTAL	5	14	19

Tabla 43. Relevancias prueba 2 Campestre Americano

Proporción observada de las veces en que los jueces estuvieron de acuerdo

$$P(A) = \frac{5+5}{19} = 0.526$$

$$P(no-relevante) = \frac{14+5}{19+19} = 0.5$$

$$P(relevantes) = \frac{5+14}{19+19} = 0.5$$

$\boldsymbol{Probalidad\ de\ acuerdo\ entre\ dos\ jueces\ (por\ azar)}$

$$P(E) = P(no-relevantes)^2 + P(relevantes)^2 = 0.25 + 0.25 = 0.5$$

$\boldsymbol{Estadistica\ Kappa}$

$$K = \frac{(P(A)-P(E))}{(1-P(E))} = \frac{(0.526-0.5)}{(1-0.5)} = \frac{0.0263}{0.5} = 0.052$$

Al obtener K=0.052 se infiere muy poco acuerdo entre los jueces sobre la relevancia encontrada. Esto puede deberse también al mejor manejo del idioma en algunos estudiantes que en otros y al número de coincidencias que arrojó cada consulta.

Para Plant cell

		Relevancias de Juez 2		
		SI	NO	TOTAL
Relevancias de Juez 1	SI	7	2	9
	NO	0	1	1
	TOTAL	7	3	10

Tabla 44. Relevancias prueba 3 Campestre Americano

Proporción observada de las veces en que los jueces estuvieron de acuerdo

$$P(A) = \frac{7+1}{10} = 0.8$$

$$P(no-relevante) = \frac{3+1}{10+10} = 0.2$$

$$P(relevantes) = \frac{7+9}{10+10} = 0.8$$

***Probalidad de acuerdo entre dos jueces* (*por azar*)**

$$P(E) = P(no-relevantes)^2 + P(relevantes)^2 = 0.04 + 0.64 = 0.68$$

Estadistica Kappa

$$K = \frac{(P(A) - P(E))}{(1 - P(E))} = \frac{(0.8 - 0.68)}{(1 - 0.68)} = 0.375$$

Esta prueba, realizada por un docente y un estudiante, fue un poco más acorde entre ellos. El acuerdo de relevancia no es suficiente pero fue mejor que para los jueces anteriores.

Promedio de índice Kappa de las consultas realizadas con todos los jueces.

Promedio K = (1+0.052+0.375)/3 = 0.475

Total de documentos revisados: juez 1 y juez 2 según ¡Error! No se encuentra el origen de la referencia.

		Relevancias de Juez 2		
		SI	NO	TOTAL
Relevancias de Juez 1	SI	19	11	30
	NO	0	17	17
	TOTAL	19	28	47

Tabla 45. Relevancia total según jueces de Campestre Americano

Proporción observada de las veces en que los jueces estuvieron de acuerdo

$$P(A) = \frac{19+17}{47} = 0.765$$

$$P(no-relevante) = \frac{28+17}{47+47} = 0.478$$

$$P(relevantes) = \frac{19+30}{47+47} = 0.521$$

***Probalidad de acuerdo entre dos jueces* (*por azar*)**

$$P(E) = P(no-relevantes)^2 + P(relevantes)^2 = 0.228 + 0.271 = 0.50$$

Estadistica Kappa

$$K = \frac{(P(A) - P(E))}{(1 - P(E))} = \frac{(0.765 - 0.50)}{(1 - 0.50)} = 0.53$$

Se observó un acuerdo en un 53% teniendo en cuenta el total de los jueces y consultas. Esto debido a los inconvenientes mencionados anteriormente.

8.3.2 Cálculos Kappa Institución Educativa Alejandro de Humboldt, sede Yanaconas

Para el cálculo de las estadísticas en la institución, se toman primero las dos consultas de cada prueba y las decisiones de sus jueces. El total de URLs revisadas es 77.

Consultas: Cuticle y Embryo

		Relevancias de Juez 2		
		SI	NO	TOTAL
Relevancias de Juez 1	SI	8	0	8
	NO	0	10	10
	TOTAL	8	10	18

Tabla 46. Relevancias prueba 1 Alejandro de Humboldt

A continuación, se realiza el cálculo de P(A) y P(E).

Proporción observada de las veces en que los jueces estuvieron de acuerdo

$$P(A) = \frac{8+10}{18} = 1$$

$$P(no-relevante) = \frac{10+10}{18+18} = 0.555$$

$$P(relevantes) = \frac{8+8}{18+18} = 0.444$$

***Probalidad de acuerdo entre dos jueces* (*por azar*)**

$$P(E) = P(no-relevantes)^2 + P(relevantes)^2 = 0.308 + 0.197 = 0.506$$

Estadistica Kappa

$$K = \frac{(P(A)-P(E))}{(1-P(E))} = \frac{(1-0.506)}{(1-0.506)} = 1$$

Se observa un acuerdo total entre los jueces. Esto debido a la mejora realizada en la indexación del prototipo y la clase previa en el idioma específico. Esta vez, su acuerdo no solo se basó en el número de coincidencias sino también, en la relevancia para cada juez.

Consultas: Xylem y Seed coat

		Relevancias de Juez 2		
		SI	NO	TOTAL
Relevancias de Juez 1	SI	9	0	9
	NO	4	4	8
	TOTAL	13	4	17

Tabla 47. Relevancias prueba 2 Alejandro de Humboldt

Proporción observada de las veces en que los jueces estuvieron de acuerdo

$$P(A) = \frac{9+4}{17} = 0.764$$

$$P(no-relevante) = \frac{4+8}{17+17} = 0.352$$

$$P(relevantes) = \frac{9+13}{17+17} = 0.647$$

Probalidad de acuerdo entre dos jueces (por azar)

$$P(E) = P(no-relevantes)^2 + P(relevantes)^2 = 0.124 + 0.418 = 0.542$$

Estadistica Kappa

$$K = \frac{(P(A)-P(E))}{(1-P(E))} = \frac{(0.764-0.542)}{(1-0.542)} = 0.484$$

Consultas: Plant cell y Fruit

		Relevancias de Juez 2		
		SI	NO	TOTAL
Relevancias de Juez 1	SI	17	0	17
	NO	1	2	3
	TOTAL	18	2	20

Tabla 48. Relevancias prueba 3 Alejandro de Humboldt

Proporción observada de las veces en que los jueces estuvieron de acuerdo

$$P(A) = \frac{17+2}{20} = 0.95$$

$$P(no-relevante) = \frac{3+2}{20+20} = 0.125$$

$$P(relevantes) = \frac{18+17}{20+20} = 0.875$$

***Probalidad de acuerdo entre dos jueces* (*por azar*)**

$$P(E) = P(no - relevantes)^2 + P(relevantes)^2 = 0.015 + 0.765 = 0.781$$

Estadistica Kappa

$$K = \frac{(P(A) - P(E))}{(1 - P(E))} = \frac{(0.95 - 0.781)}{(1 - 0.781)} = 0.770$$

El índice K, muestra un acuerdo cercano entre las decisiones de los jueces. Es un buen indicador de lo que realmente busca el usuario.

Consultas: Flower y Seed

		Relevancias de Juez 2		
		SI	NO	TOTAL
Relevancias de Juez 1	SI	17	3	20
	NO	0	0	0
	TOTAL	17	3	20

Tabla 49. Relevancias prueba 4 Alejandro de Humboldt

Proporción observada de las veces en que los jueces estuvieron de acuerdo

$$P(A) = \frac{17 + 0}{20} = 0.85$$

$$P(no - relevante) = \frac{0 + 3}{20 + 20} = 0.075$$

$$P(relevantes) = \frac{12 + 13}{20 + 20} = 0.625$$

***Probalidad de acuerdo entre dos jueces* (*por azar*)**

$$P(E) = P(no - relevantes)^2 + P(relevantes)^2 = 0.0056 + 0.390 = 0.396$$

Estadistica Kappa

$$K = \frac{(P(A) - P(E))}{(1 - P(E))} = \frac{(0.85 - 0.396)}{(1 - 0.396)} = 0.751$$

Al igual que los jueces anteriores, para esta consulta, tuvieron un acuerdo cercano. Esto debido a la revisión de los documentos y la relevancia para ellos.

Promedio de índice Kappa de las consultas realizadas con todos los jueces.
Promedio K = (1+0.484+0.770+0.751)/4 = 0.751

Total documentos revisados: juez 1 y juez 2 según ¡Error! No se encuentra el origen de la referencia.

		Relevancias de Juez 2		
		SI	NO	TOTAL
Relevancias de Juez 1	SI	54	0	54
	NO	2	21	23
	TOTAL	56	21	77

Tabla 50. Relevancia total según jueces de Alejandro de Humboldt

Proporción observada de las veces en que los jueces estuvieron de acuerdo

$$P(A) = \frac{54 + 21}{77} = 0.974$$

$$P(no - relevante) = \frac{21 + 23}{77 + 77} = 0.285$$

$$P(relevantes) = \frac{56 + 54}{77 + 77} = 0.714$$

***Probalidad de acuerdo entre dos jueces* (*por azar*)**

$$P(E) = P(no - relevantes)^2 + P(relevantes)^2 = 0.081 + 0.510 = 0.795$$

Estadistica Kappa

$$K = \frac{(P(A) - P(E))}{(1 - P(E))} = \frac{(0.974 - 0.795)}{(1 - 0.795)} = 0.872$$

El acuerdo en este caso fue mayor que en el total calculado en el primer colegio evaluado y mayor que en varios casos específicos, lo cual es un buen indicador del funcionamiento de la indexación semántica de acuerdo a las necesidades de información de los usuarios.

9 ANEXO I

9.1 MANUAL DE USUARIO

El manual de usuario para el presente proyecto consta de la especificación del uso del prototipo construido, el cual se basa en realizar búsquedas en la web bajo el dominio de botánica dirigido a estudiantes de básica primaria.

9.1.1 REALIZAR CONSULTA

El usuario debe ingresar a la página con url: http://www.prometeo.unicauca.edu.co/BuscadorSemantico/SemanticIndexSearch.aspx donde aparece el buscador semántico con la interfaz propia de acuerdo al dominio establecido (botánica).

Figura 20. Realizar consulta

Después de realizar la consulta se da click en el botón “Buscar” para acceder a los resultados.

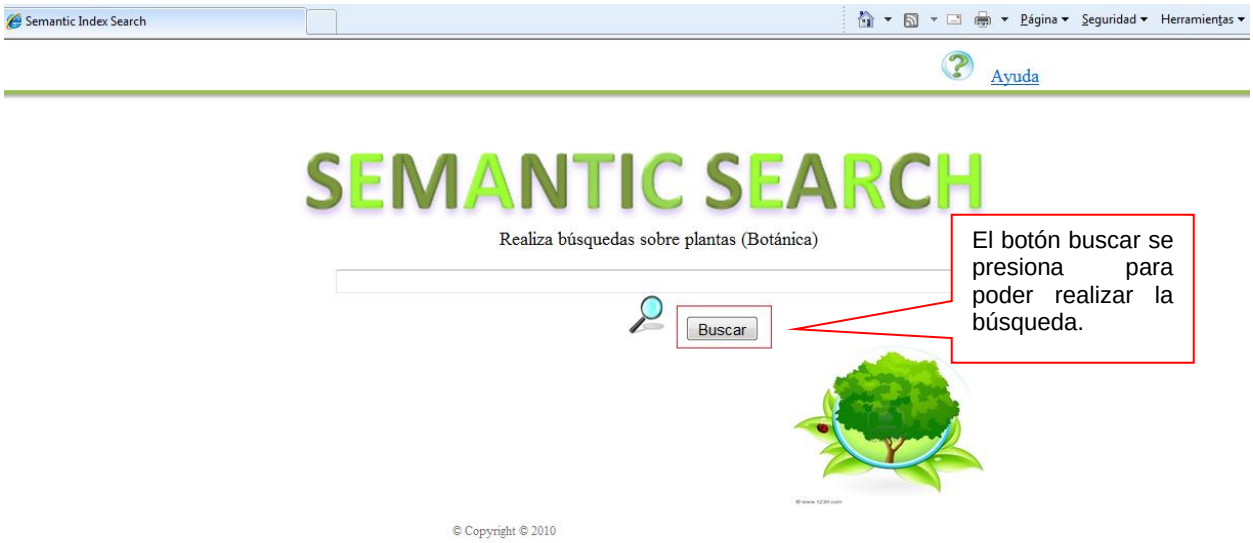


Figura 21. Botón "Buscar"

Al realizar la búsqueda aparecen los resultados en la página del explorador.

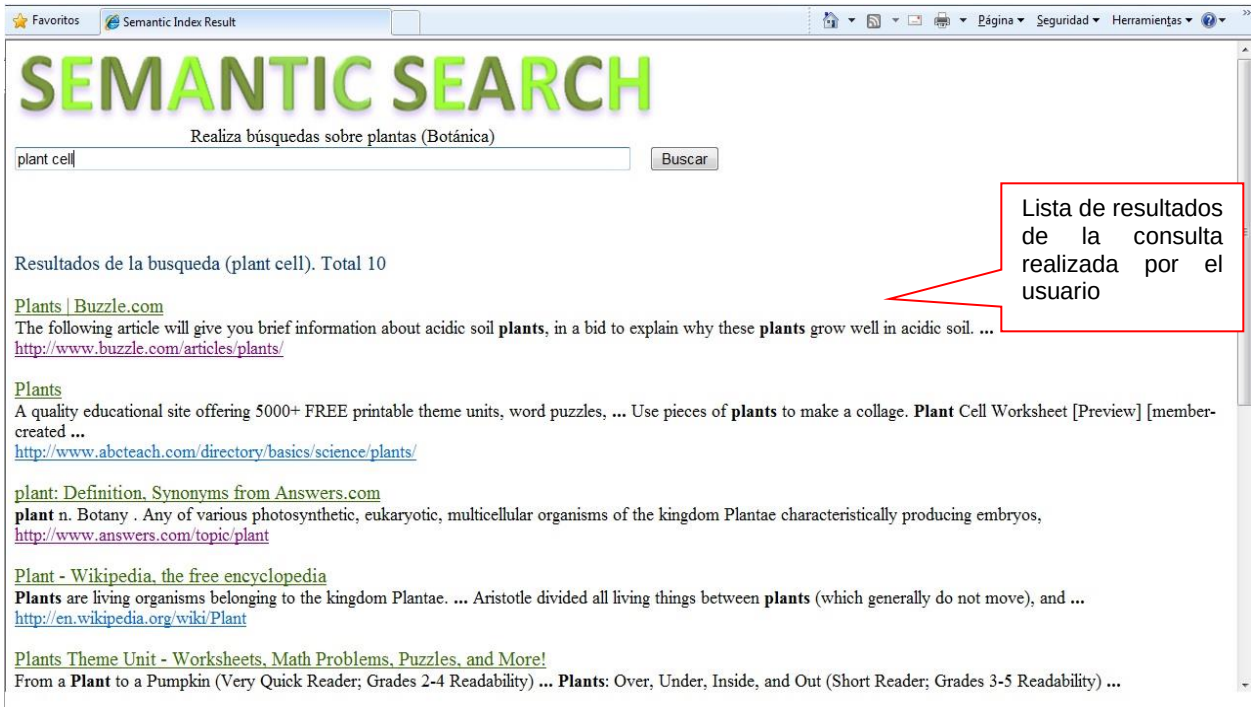


Figura 22. Lista de resultados

9.1.2 OBTENER AYUDA

Para obtener ayuda sobre la aplicación y autores de la misma, puede ingresar al botón “Ayuda” en la parte superior derecha de la página.

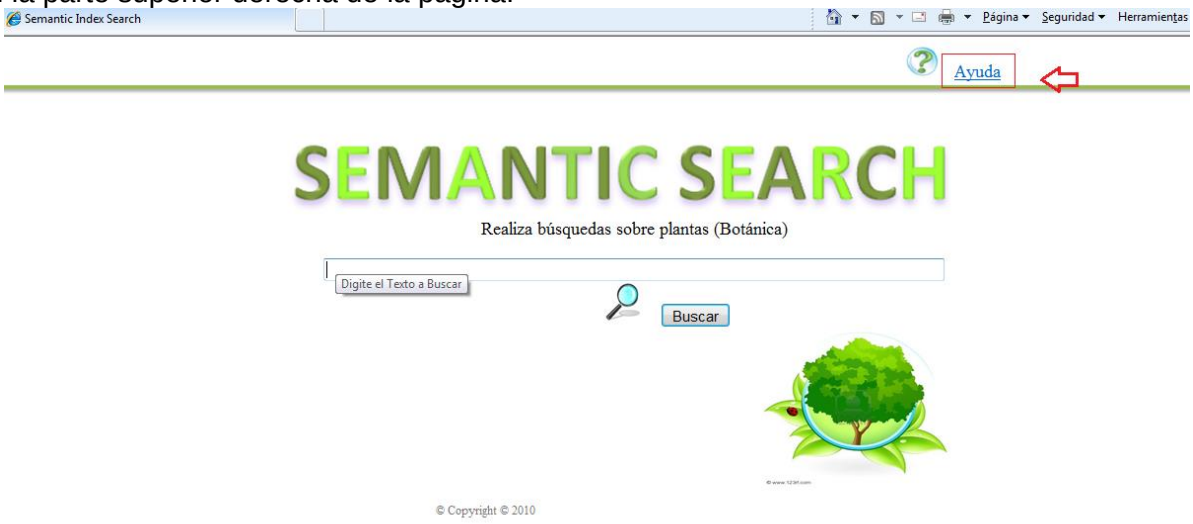


Figura 23. Obtener Ayuda

10 REFERENCIAS

1. Prada Juan José, C.S., Larraud Marina, *Recuperación de Información Bilingüe en la Web Semántica, Extraido el 20 de febrero de 2010*. 2007, Instituto de Computacion. p. 90.
2. Van Heijst, G.S., A. T., Wielinga, B. J., *Using explicit ontologies in KBS development. Extraído el 2 de febrero de 2010.* International Journal of Human-computer Studies, 1997: p. 183-292.
3. Mizoguchi, R., Vanwelkenhuysen, J., Ikeda, M., *Task Ontology for Reuse of Problem Solving Knowledge. Extraído el 2 de febrero de 2010.* Towards Very Large Knowledge Bases: KnowledgeBuilding and Knowledge Sharing, 1995: p. 46-59.
4. Davis, R., Sorbe, H., Szolovits, P. , *What is a Knowledge Representation?. Extraído el 10 de diciembre de 2009.* AI Magazine. Spring, 1993: p. 17-33.
5. Poli, R., *Levels of Reality. Extraído el 10 de diciembre de 2009.* BISCA 2000: Bolzano International Schools in Cognitive Analysis, 2000.
6. Anisleiby Fernandez H., S.C., Yudeisy Pérez G., Tatiana Villalón A., *las ontologías. Nuevos retos.* IX Congreso ISKO-España. Nuevas perspectivas para la fusión y organización del conocimiento, 2009: p. 355 - 379.
7. Leandro. *Herramientas para trabajar con ontologias, Extraido el 10 de abril de 2010*. Lea en Binario 2010 [cited.
8. Arenas Sandra, A.M.T., Guzman Jaime Alberto, *SISDEON: An information system on displacement in Colombia using ontology, Extraido el 20 de febrero de 2010.* 2009.
9. Samper Zapater José Javier, D.J.J.M.D., D. Eduardo Carrillo Zambrabo, *ONTOLOGÍAS PARA SERVICIOS WEB SEMÁNTICOS DE INFORMACIÓN DE TRÁFICO: DESCRIPCIÓN Y HERRAMIENTAS DE EXPLOTACIÓN. Extraído el 10 de abril de 2010*. 2005, UNIVERSITAT DE VALENCIA.
10. University, P. *WordNet 3.0* Princenton University 2009 [cited 2 de febrero de 2010]; Available from: http://wordnet.princeton.edu/wordnet.
11. Collins, A.M.a.Q., R. M., *Retrieval time from semantic memory.* Journal of Verbal Learning and Verbal Behavior, 1969. **8**: p. 240-247.
12. Laboratory, S.U.K.S. *Stanford KSL Network Services*. [cited 10 marzo de 2010]; Available from: http://www-ksl-svc.stanford.edu:5915/.
13. Palacios E. Juan P., C.R.C., *Modelo de Unificación Semántica de Ontologías, Aplicado al Dominio de los Archivos Digitales, Extraído el 10 de abril de 2010*, in *Departamento de Ingeniría de Sistemas Telemáticos*. 2005, UPM. p. 192.
14. Jimenez, A. *KIF – Knowledge Interchange Format*. 2007 [cited 9 de abril de 2010]; Available from: http://alfonsojimenez.com/uncategorized/kif-knowledge-interchange-format/.
15. Tramullas, J., *Agentes y Ontologías para el Tratamiento de la Información: clasificación y recuperación en Internet. Extraído el 5 de marzo de 2010*, in *Dep. de CC. de la Documentación* Universidad de Zaragoza: Zaragoza.
16. Miguel A. Alonso, J.G., Jesús Vilares, *Recuperación de Información en Internet. Extraído el 10 de mayo de 2010*, Universidad de Coruña: Coruña. p. 18.
17. Broncano, R.G. *Modelos de Recuperación, Recuperación y Organización de la Información. Extraído el 10 de mayo de 2010*. 2006 [cited; Available from: http://modelosrecuperacion.tripod.com/.

18. Felipe, J.A.R. *Recuperar información de la Internet profunda. extraído el 9 de marzo de 2010*. 2001 [cited; Available from: http://www.sociedadelainformacion.com/20011103/invisible/internetprofundo.htm.
19. Quesada, B.V. *Agentes inteligentes: definición y tipología. Los agentes de información. Extraído el 10 de febrero de 2010*. 2000 [cited; Available from: http://www.monografias.com/trabajos917/agentes-inteligentes-informacion/agentes-inteligentes-informacion.shtml.
20. Molina, M.P. *Búsqueda y Recuperación de Información*. 2009 [cited 27 de abril de 2010]; Available from: http://www.mariapinto.es/e-coms/recu_infor.htm.
21. Christopher D. Manning, P.R., Hinrich Schütze, *An Introduction to Information Retrieval, Extraído el 5 de mayo de 2010*. 2009, Cambrige University Press: Cambrige. p. 581.
22. Paice, C.D., *Another Stemmer, Extraído el 27 de abril de 2010*, in *Department of Computing* Lancaster University: Bailrigg, Lancaster.
23. Francisco Ruiz, J.V., *Guía de Uso de SPEM 2 con EPF Composer*. 2008, Universidad de Castilla-La Mancha.
24. Ruiz, F., *Introducción a la Ingeniería de Procesos Software*. 2008, Universidad de Cantabria.
25. *Diagrama de flujo de proceso*. 2009 [cited 10 de agosto de 2010]; Available from: http://148.202.148.5/cursos/id209/mzaragoza/unidad2/unidad2tres.htm.
26. Vazquez, A.M. *Herramientas organizacionales. Diagrama de flujo*. 2010 [cited 6 de septiembre de 2010]; Available from: http://www.estrucplan.com.ar/Producciones/entrega.asp?IDEntrega=526.
27. Huanca, J.C. *Lenguaje Unificado de Modelado. Diagramas*. 2010 [cited 18 de agosto de 2010]; Available from: http://www.grupoinformatica.com/biblioteca-articulos/1459-uml-lenguaje-unificado-de-modelado.html.
28. Pérez, J.D., *Notaciones y lenguajes de procesos. Una visión global.*, in *Departmento de Languajes y Sistemas Informáticos*. 2007, University of Sevilla: Sevilla. p. 109.
29. Educación, M.d. *FreeMind: mapas conceptuales*. 2009 [cited 11 de agosto de 2010]; Available from: http://recursostic.educacion.es/observatorio/web/es/software/software-general/716-freemind-mapas-conceptuales.
30. Gómez, L. *Software para elaborar mapas mentales y conceptuales*. 2008 [cited 11 de agosto de 2010]; Available from: http://manantialdevida.obolog.com/software-elaborar-mapas-mentales-conceptuales-59333.
31. Ecourban. *Mapas conceptuales*. [cited 10 de agosto de 2010]; Available from: http://www.ecourban.org/profesores/didactica/mapasconceptuales/index.html.
32. Jörg Müller, D.P. *FreeMind - free mind mapping software*. 2010 [cited 3 de septiembre de 2010]; Available from: http://freemind.sourceforge.net/wiki/index.php/Main_Page.
33. Alvarez, M.A. *FreeMind*. 2008 [cited 6 de septiembre de 2010]; Available from: http://www.desarrolloweb.com/articulos/freemind.html.
34. *the Gene Ontology* 2010 [cited 11 de agosto de 2010]; Available from: http://www.geneontology.org/.
35. *Plant Ontology*. 2010 [cited 10 de agosto de 2010]; Available from: http://www.plantontology.org/].
36. *The Environment Ontology*. 2008 [cited; Available from: http://www.environmentontology.org/.
37. *The Science Environment for Ecological Knowledge* 2008 [cited 10 de agosto de 2010]; Available from: http://seek.ecoinformatics.org/.
38. García, J. *Patrones de diseño*. 2005 [cited 27 de julio de 2010]; Available from: http://www.ingenierosoftware.com/analisisydiseno/patrones-diseno.php.

39. Joydeep Ghosh, D.L., *Performance Evaluation of Information Retrieval Systems*, Univ. of Science and Tech.