Modelación léxico semántica de descripciones de servicios web

Agosto 1, 2011

1 Resumen

Los servicios Web (WS) son somponentes de software que son almacenados e invocados a lo largo de la Web. Actualmente, los WS han ganado popularidad por las caracterśticas que ofrecen para interoperar y crear sistemas de mayor complejidad. La descripción de los WS se realiza mediante el lenguaje Web Servide Description Language (WSDL), que se centra en describir sus operaciones mediante la descripción de sus parámetros, invocación y la salida que proporciona. Para obtener mayor provecho de estos servicios y apoyar su uso y composición automatizados, es necesario mejorar los modelos de localización y acceso a los WS.
En contraste con la localización de una página Web, donde el usuario puede discriminar de entre los resultados para leer la página que le interesa, la búsqueda de un WS mediante palabras clave resulta insuficiente debido a ue el usuario no está acostrumbrado a realizar peticiones de información amplias y precisas, ni tampoco tiene la agilidad para leer documentos escritos en WSDL, esto es, textos híbridos compuestos de esquemas de operaciones y breves textos en lenguaje natural. Por tanto, este problema debe abordarse con la incorporación de información adicional en los motores de búsqueda y recuperación de información [1].
El presente trabajo incursiona en la exploración de la representación y similitud de WS apyándose en la información léxico semántica, extraída de las descripciones de los WS, con el fin de agruparlos y clasificarlos para falicitar su compresión, localización y acceso.

2 Nombre y datos personales de los participantes

2.1 Instituciones participantes e integrantes

El grupo de trabajo está integrado por especialistas en los campos de Lingüística Computacional y Tecnologías de la Información de tres instituciones nacionales: Universidad Autónoma Metropolitana Unidad Cuajimalpa (UAM-C), B. Universidad Autónoma de Puebla (BUAP), y Centro Nacional de Investigación y Desarrollo Tecnológico (CENIDET). También participan estudiantes de licenciatura, maestría y doctorado de estas tres instituciones. En orden alfabético: Para su uso en este documento, consideramos la abreviación de los nombres de los participantes indicada entre paréntesis en cada uno de ellos, y su adscripción en negritas.

- (HJ). Dr. Héctor Jiménez Salazar. Profesor de **UAM-C**. SNI I. Experiencia en diversos temas de Procesamiento del Lenguaje Natural, Recuperación de Información, y Desambiguación del Sentido de los términos.
- (CL) Dr. Christian Lemaitre y León. Profesor de **UAM-C**. Experiencia en procesamiento de lenguaje natural, sistemas multiagentes, así como en organizaciones e instituciones electrónicas.
- (AL) M.C. Wulfrano Arturo Luna Ramírez. Profesor de **UAM-C**. Experiencia en desarrollo de sistemas de tratamiento automático de textos, y algoritmos de aprendizaje automático.

Actividad	HJ	CL	AL	CR	CS	DP	AM	MT	EL	EM	PR
1.1	•				•	•			1,2		
1.2			•		•				1,2		
1.3	•	•	•	•	•	•	•	•	1,2		
2.1	•				•		•	•	3,4	1	
2.2					•	•			3,4	1	
2.3	•			•	•				3,4	1	
2.4	•	•	•	•	•	•	•	•	3,4	1	
3.1	•			•	•	•				2	•
3.2	•	•	•	•	•	•	•	•		2	•

Table 1: Participación del grupo de trabajo en cada una de las actividades.

- (AM) Dra. Azucena Montes Rendón. Profesora del **CENIDET**. SNI I. Especialidad en Procesamiento de lenguaje natural, análisis semántico en textos, modelado de representación de conocimiento, ontologías, y Web semántica.
- (DP) David Eduardo Pinto Avendaño. Profesor de **BUAP**. SNI C. Agrupamiento, expansion y evaluación de textos cortos. Recuperación de información.
- (CR) Dr. Carlos Rodríguez Lucatero. Profesor de **UAM-C**. Perfil Promep. Experiencia en algorítmica de soluciones aproximadas y probabilistas.
- (CS) Dr. Christian Sánchez Sánchez. (Responsable del proyecto). Profesor de **UAM-C**. SNI C. Experiencia en composición dinámica de servicios web, descubrimiento de servicios web, y expansión de consultas para búsqueda de servicios web.
- (MT) M.C. Mireya Tovar Vidal. Estudiante de doctorado del **CENIDET**. Perfil Promep. Experiencia en procesamiento del lenguaje natural, y agrupamiento.
- (EM) Dos estudiantes de maestría del **CENIDET**.
- (EL) Dos estudiantes de licenciatura de Tecnologías de la Información de la **UAM-C**.
- (EL) Dos estudiantes de licenciatura de Ciencias de la Computación de la **BUAP**.
- (PR) Programador contratado por honorarios.

En la Tabla 1 se muestra la participación de los miembros del grupo de trabajo.

2.2 Justificación y planteamiento del objeto de estudio

El desarrollo de *software* se apoya, en adición a los métodos convencionales, en el reuso de *software* [15] a través de programotecas para la integración de sistemas, con lo cual se aligera el esfuerzo y los tiempos de desarrollo. Los servicios web (WS) son componentes de *software* almacenados e invocados a lo largo de la web que coadyuvan a la composición de *software* de manera dinámica. Actualmente, los WS han ganado popularidad por las características que ofrecen para interoperar y crear sistemas de mayor complejidad. Estos sistemas son compuestos a partir de la conexión, configuración y la coordinación de varios de los servicios disponibles en la web. Cabe mencionar que para sacar provecho de las características de estos servicios, y apoyar la automatización de la composición, es necesario mejorar los modelos actuales de acceso a dichos servicios. La descripción de un WS, hecha mediante el lenguaje WSDL, *Web Service Description Language*, se centra en las invocaciones a las operaciones que ofrece el servicio web. Mientras la localización de una página web donde el usuario puede discriminar algunos resultados, la búsqueda de WS con palabras clave resulta insuficiente puesto que el usuario no tiene práctica en hacer peticiones amplias y precisas de información, ni tampoco la agilidad para leer documentos WSDL; es decir, textos híbridos

compuestos de esquemas de operaciones y breves textos en lenguaje natural. Por tanto, la atención a este problema con motores de búsqueda para Recuperación de Información [1] tiene que ser sustentada con información adicional. Específicamente, el presente proyecto incursiona en la exploración de la representación y similitud de servicios web apoyándose en la información léxico semántica contenida en la descripción de los servicios.

Cabe señalar que por la naturaleza de este problema se hace necesario el trabajo conjunto de especialistas de diversos campos de la coputación como es el caso de los miembros del grupo proponente, a saber: Procesamiento de Lenguaje Natural (representación, búsqueda y recuperación de información), Sevicios Web (descubrimiento y composición), Aprendizaje Automático (agrupación y clasificación), Web Semántica (Creación y uso de Ontologías Web).

```
<?xml version="1.0" encoding="utf-8" ?>
- <wsdl:definitions xmlns:soap="http://schemas.xmlsoap.org/wsdl/soap/"
    xmlns:wsdl="http://schemas.xmlsoap.org/wsdl/">
  + <wsdl:types>
  + <wsdl:message name="BusinessDemographicsByLinkIDSoapIn">
  + <wsdl:message name="BusinessDemographicsByLinkIDSoapOut">
  + <wsdl:message name="BusinessDemographicsByCompanySoapIn">
  + <wsdl:message name="BusinessDemographicsByCompanySoapOut">
  - <wsdl:portType name="CorteraBusinessVitalsSoap">
    - <wsdl:operation name="BusinessDemographicsByLinkID">
      - <documentation xmlns="http://schemas.xmlsoap.org/wsdl/">
          Gets business demographics from a Link ID
        </documentation>
        <wsdl:input message="tns:BusinessDemographicsByLinkIDSoapIn" />
        <wsdl:output message="tns:BusinessDemographicsByLinkIDSoapOut" />
      </wsdl:operation>
    - <wsdl:operation name="BusinessDemographicsByCompany">
      - <documentation xmlns="http://schemas.xmlsoap.org/wsdl/">
          Gets business demographics by company
        </documentation>
        <wsdl:input message="tns:BusinessDemographicsByCompanySoapIn" />
        <wsdl:output message="tns:BusinessDemographicsByCompanySoapOut" />
      </wsdl:operation>
    </wsdl:portType>
  + <wsdl:binding name="CorteraBusinessVitalsSoap" type="tns:CorteraBusinessVitalsSoap">
  - <wsdl:service name="CorteraBusinessVitals">
    - <documentation xmlns="http://schemas.xmlsoap.org/wsdl/">
        The Cortera Business Vitals Service retrieves business demographics
        for companies based on a Link ID or a company name.
      </documentation>
    - <wsdl:port name="CorteraBusinessVitalsSoap" binding="tns:CorteraBusinessVitalsSoap">
        <soap:address location="http://ws.strikeiron.com/StrikeIron/CorteraBusinessVitals" />
      </wsdl:port>
    </wsdl:service>
</wsdl:definitions>
```

Figure 1: Descripción de un servicio web de *Cortera Business Vitals* en WSDL. En éste se muestran dos operaciones, `BusinessDemographicsByLinkId` y `BusinessDemographicsByCompany`; sus parámetros linkID (tipo String) y Company (tipo de dato compuesto por CompanyName, Address,City, State, PostalCode y Country, todos strings), respectivamente. También aparecen textos en lenguaje natural asociados a los nodos (etiqueta `<documentation xmlns=''http://schemas.xmlsoap.org/wsdl/''>`); por ejemplo `Gets business demographics from a Link ID`.

2.3 Consideraciones sobre la originalidad de la propuesta

Nuestro enfoque relaciona la semántica (relaciones entre términos) y la sintaxis (etiquetas sobre los componentes del documento), consecuentemente incide en la representación y similitud de descripciones de servicios web: textos cortos e híbridos. Específicamente el planteamiento que hacemos sobre el acceso a WS está basado en los siguientes elementos:

1. Uso de la autoexpansión para enriquecer un WS. La finalidad es tener una representación de las descripciones

que impacte la efectividad de los métodos de agrupamiento y clasificación. Se ha probado que la autoexpansión en textos de lenguaje natural mejora el desempeño tanto en recuperación de información como en el agrupamiento de textos. En este caso extenderemos el método de autoexpansión a textos híbridos.

2. Determinación de la similitud semántico estructural mediante consideraciones sobre la similitud entre las partes de los componentes. Se trata de hacer corresponder el texto con la descripción del servicio web y combinar adecuadamente las similitudes de las partes para obtener una función de similutud sensible a las relaciones que se establecen entre el texto y la descripción de los servicios web. Por ejemplo, el texto en el WSDL correspondiente a una operación podría establecer una relación entre los parámetros de entrada y salida; con ello puede ayudarse a reforzar o debilitar el grado de similitud entre dichas partes de los componentes.

3. Clasificación basada en las relaciones semánticas de los términos. Aún cuando se hayan enriquecido las frecuencias de los términos, es conveniente extraer algunas regularidades que ayuden a la clasificación de servicios web de acuerdo al agrupamiento obtenido. Se construirá un tesauro para enriquecer la descripción de un servicio web a ser clasificado, y así comprobar la efectividad de la función de similitud.

3 Antecedentes históricos, teóricos y conceptuales

El problema de acceso a servicios web ha sido enfocado mediante el agrupamiento de una colección de WS; de esta forma se sabría qué tipo de servicio web proporciona cada WS de la colección, además, podría clasificarse un nuevo WS, o bien, proporcionar un medio para navegar entre WS. Por ejemplo, ha sido exitoso un método de agrupamiento *ad hoc* que toma en cuenta la coocurrencia de los términos para guiar la fusión y fisión de los grupos que se van generando [4]. Dicha coocurrencia se enfoca a través de los *itemsets* frecuentes. En esta dirección se han propuesto mejoras para agrupadores centrados en documentos XML como *XK-means* [26] donde se define el concepto de *tree-tuple* con lo cual la representación se expresa mediante partes del documento XML que están relacionadas por la semántica subyacente que presenta el documento original. Se recurre, entonces, al enriquecimiento léxico semántico para habilitar tal representación y, en éste como otros trabajos, se utiliza WordNet [5]. El empleo de WordNet se ha enfocado a la obtención de los sinónimos, hiperónimos, hipónimos, y cohipónimos, de los términos encontrados en el documento WSDL, con la finalidad de integrar los conceptos relacionados y, así, entablar la similitud entre los servicios web, [28]. Sin embargo, las bases de datos léxicas de carácter general, como WordNet, no capturan la terminología de dominios especializados y, por tanto, no hay una mejora sustancial en el acceso a los WS.

Otros autores han sistematizado las regularidades encontradas en los WS aplicando Análisis de Conceptos Formales (FCA) [6]; esto permite aprehender algunos conceptos subyacentes en los contenidos de WSDL, y construir una retícula que mantenga relaciones de generalidad y particularidad entre conceptos [3]. La retícula ayuda a enriquecer algunas peticiones de información como es el caso de la recuperación de WS con base en un servicio web determinado.

Por otro lado, hay quienes se apoyan en los conceptos contenidos en el WSDL para poder alinearlos, encontrando conceptos similares, y/o clasificarlos con ontologías disponibles en la Web, con el propósito de beneficiarse de las relaciones entre los conceptos de la ontología y así poder habilitar la similitud [11, 14]. Es importante reconocer que una ontología ofrece una sistematización conceptual indiscutible, aunque dicha sistematización tiene una vida efímera puesto que los WS evolucionan constantemente, asimismo serían requeridas muchas ontologías: una por dominio según el servicio web.

Cualquiera de los enfoques conlleva a considerar no solamente los términos que aparecen en los identificadores de los componentes de software (nombre de la operación, nombre de los parámetros de entrada y salida; ver, por ejemplo fig. 1) sino, además, los textos que en ocasiones aparecen en el archivo WSDL. El problema con estos textos es que son muy cortos (una decena de términos) y las colecciones de WSDL no son muy grandes. Ello dificulta la representación basada en la frecuencia de los términos y, a la vez, la aplicación de variadas formas de ponderación de términos [7]. Se ha visto que el tratamiento de expansión de fragmentos [10] provee mejoras en la representación: da relevancia a ciertas zonas de la descripción de un WS, mas no se ha determinado cómo, a través de la expansión, se logre poner en relevancia los términos más importantes.

La similitud entre dos WS debe considerar, a diferencia de los textos en lenguaje natural donde se emplea el coseno del ángulo que forman los vectores correspondientes a los textos, la estructura de los componentes de manera composicional: la similitud entre cada una de las partes de un WS contribuirá a la similitud total [4]. Los trabajos en esta dirección utilizan comúnmente una combinación de las similitudes de las partes. Estas combinaciones de similitudes pueden implicar barridos de conjuntos de datos masivos, lo cual ha podido enfrentarse mediante distancias de edición y algoritmos que consideran la estructura [32, 33] de los documentos XML [35, 13]. Aún cuando este enfoque presenta eficiencia, hasta donde se conoce, no ha habido adaptaciones de su aplicación a documentos WSDL.

3.1 Trabajos previos del grupo proponente

Uno de los retos, con los que hay que lidiar en la identificación de la función de los servicios (el descubrimiento de servicios), es que la información tanto de la consulta como de la descripción del servicio es, a menudo, incompleta, incierta o ambigua. Lo que trae como consecuencia que en muchos de los casos el usuario tenga que estar preparado a no encontrar un servicio que satisfaga la consulta o encontrar muchos servicios que lo hagan.
Sin embargo, algunos enfoques, cuando los servicios están anotados semánticamente (por medio de ontologías), han demostrado que pueden clasificarlos y ayudar a solucionar el problema de selección de servicios. Dichos enfoques son los siguientes: 1) el modelado estructural de los servicios web [23] y 2) la definición de diferentes grados de similitud en el empatamiento de la consulta y descripción, usando el razonamiento basado en lógica difusa y lógica descriptiva, [24]. Por otro lado, un enfoque alternativo, mediante 3) la expansión de las consultas y descripciones usando lógica descriptiva de acuerdo con las relaciones entre los conceptos de las ontologías, amplia la información y las opciones de búsqueda, [22]. No obstante la efectividad de estos enfoques, el problema es que muy pocos servicios están anotados semánticamente con ontologías.
El descubrimiento de servicios se ha abordado mediante el agrupamiento de colecciones de documentos WSDL para identificar los diferentes grupos que consitituyen a la colección, y así apoyar la recuperación de servicios web a través de su clasificación con el agrupamiento obtenido. En esta dirección, se ha explorado el acceso a los servicios a partir de una consulta, con base en la extracción de información semántica [31].
También se ha comprobado que en tareas de agrupamiento de textos cortos y desambiguación del sentido de una palabra, la autoexpansión de términos resulta efectiva [17, 18]. Dichas tareas han usado *corpora* desbalanceados (en tamaño de clases), y aún no muy grandes, para mejorar el desempeño de enfoques alternativos; por ejemplo, el *corpus* de la competencia SemEval 2007 se formó por no más de una centena de oraciones con el fin de descubrir los sentidos de una palabra en una oración.
La relevancia de los términos, medida tradicionalmente con $tf_{ij}idf_j$, ha sido explorada con el enfoque propuesto por Luhn [12] en cuanto a la importancia de los términos de frecuencia intermedia; tomando como base la obtención del punto de transición [27, 2, 9]. La autoexpansión y el uso de técnicas de selección de términos se han combinado en aplicaciones con gran ahorro de recursos; reducción de la dimensionalidad de los documentos hasta debajo del 10% del tamaño original [20, 16] y, en consecuencia el tiempo de procesamiento. Parte de este enfoque se ha aplicado recientemente en la sección *Web Service Discovery* del taller organizado por INEX en 2010. Se aplicaron, en esta competencia, métodos de agrupamiento y recuperación de información apoyados en parte de la experiencia previa. En los experimentos realizados se usó un *corpus* supervisado compuesto de 25 tópicos (consultas) y 1,738 documentos en formato WSDL [25]. Sin embargo, los organizadores aún no han publicado los resultados de la evaluación [34].
El problema de similitud entre servicios web obliga a tratar información que se encuentra en forma distribuida, por medio del uso distancias de edición y algoritmos probabilistas que aproximen dichas distancias de manera eficiente. Algunos resultados en este tema son los presentados en [19].

4 Preguntas y supuestos de investigación

El planteamiento es que el agrupamiento y la clasificación de WS dependen fuertemente de la representación de sus descripciones y que ella debe aprovecharse en la medida de similitud. Más detalladamente:

- **H1.** Las relaciones léxico semánticas se mantienen en los textos híbridos.

- **H2.** Los textos híbridos muy cortos pueden ser enriquecidos por autoexpansión: pueden identificarse términos importantes con base en sus frecuencias.
- **H3.** El enriquecimiento permite aplicar técnicas de selección de términos para representar los WS y, en consecuencia, es posible construir un tesauro con una colección no muy grande de WS.
- **H4.** Hay una forma de composicionalidad en el cálculo de la similitud entre dos WS que mejora el desempeño del agrupamiento.
- **H5.** La clasificación de textos híbridos con métodos apoyados en la información léxico semántica extraída mejora el desempeño.

5 Objetivos

5.1 Objetivo General

Desarrollar un modelo de representación de servicios web basado en la semántica extraída de sus descripciones con la finalidad de mejorar el agrupamiento y la clasificación de documentos WSDL.

5.2 Objetivos Particulares

1. Extraer contenido léxico semántico de los textos, métodos, mensajes, parámetros y tipos de datos de las descripciones en WSDL con base en la identificación de los términos representativos del texto híbrido.
2. Proponer métodos que impacten el acceso de servicios web basados en las relaciones subyacentes a la información contenida en los WS.
3. Evaluar cada uno de los métodos propuestos en colecciones de WS, y analizar el desempeño de los métodos a partir de los resultados obtenidos.

6 Metodología

Nuestras hipótesis serán confirmadas o rechazadas con base en pruebas empíricas. Requerimos realizar experimentos los cuales se han distribuido entre subgrupos de participantes del presente proyecto. Además de la comunicación permanente entre los participantes del grupo de investigación, tenemos previstas tres reuniones plenarias al finalizar cada cuatrimestre de cada año. En estas reuniones se presentarán los avances, se interpretarán los resultados, y se discutirá la mejor manera de realizar los experimentos posteriores. En las etapas anuales, que a continuación se presentan, se describen las acciones que llevaremos a cabo durante la ejecución del proyecto.

Etapa 1. Proponer un método para representar textos híbridos con base en la semática subyacente.

- El objetivo es conocer cómo influye la autoexpansión en la identificación de términos importantes y realizar ajustes en el filtrado de la información que representa a cada documento.
- En esta etapa se realizarán las siguientes acciones: a) normalización de los documentos de la colección (incluye limpieza, eliminación de palabras cerradas, y extracción de términos a partir de identificadores), b) cálculo de la matriz de confusión para la colección normalizada, c) autoexpansión de documentos y sus variantes: de comentarios, de operaciones, de parámetros, y combinaciones entre éstos, d) cálculo de la matriz de confusión para la colección expandida, e) comparación de resultados y ajustes. Se usará una función de similitud base, coeficiente de Jaccard, aplicado en forma global (reuniendo los términos por cada método, obtenidos de los identificadores de la operación, parámetros, y del texto en lenguaje natural).

- Ensayar algunas formas de representación de documentos con base en la autoexpansión. Se preven dos casos: bolsas de palabras usando la similitud de Jaccard, y vectorial usando el coseno. Se determinarían los *baseline* para textos normalizados y con variantes de los textos expandidos en un ambiente de Recuperación de Información. La idea es contar con una colección supervisada (clasificada), separando algunos documentos con los cuales se puedan hacer consultas para obtener los más similares (se espera recuperar documentos de su clase).
- Los resultados serán analizados para proponer la aplicación de algunos métodos de selección de términos, iniciando por el tradicional $tf_{ij}idf_j$, en el caso del modelo de espacio vectorial; para bolsas de palabras, se espera ajustar selección de términos usando información mutua o χ^2.

'roponer un método para determinar la similitud de dos WS basado en la información estructural.

- El objetivo es determinar la mejor forma de calcular la similitud entre servicios web acudiendo a la información léxico semántica distribuida en los documentos WSDL.
- Aplicar un algoritmo de "conocimiento pobre" [8] para extraer relaciones léxico semánticas de los documentos agrupados; ello puede confrontarse con las relaciones extraídas de los documentos clasificados (en una colección supervisada). De aquí, determinar una medida de efectividad sobre la representación y ajustar parámetros sobre la representación basada en la autoexpansión.
- Aplicar diversos métodos que calculan la similitud basados en la composicionalidad para conocer su efectividad. En esta parte se espera confrontar el cálculo de las matrices de confusión en documentos representados de manera diversa, con el uso del tesauro, y en diferentes colecciones.
- Analizar los resultados del cálculo de la similitud con variantes de métodos: distancia de edición arborescente y métodos de aproximación para distancias composicionales, en diversos contextos de prueba.
- Evaluar el método de similitud estructural en un escenario de agrupamiento. Estas pruebas permitirán ajustar diversos parámetros del agrupamiento con base en el análisis de las decisiones tomadas.

Eta: Adaptar y evaluar los métodos de agrupamiento y clasificación accediendo a las relaciones léxico semánticas.

- El objetivo es conocer la efectividad de los métodos desarrollados: representación de documentos y similitud entre documentos.
- Aplicar diversos métodos de agrupamiento usando la función de similitud en documentos representados mediante la autoexpansión. Comparar el mejor método con la colección de documentos no expandidos y usando la función de similitud, así como los documentos expandidos usando otra función de similitud.
- Usar el tesauro en la clasificación de un servicio web para conocer la calidad de la clasificación en diversas colecciones, además, haciendo uso de las agrupaciones. Con esto se pretende medir la efectividad global de los métodos en un escenario completo: partiendo de una colección de documentos, efectuar la representación, extraer las relaciones léxico semánticas, agrupar los documentos, y clasificar un nuevo documento que describe un servicio web.
- Explicar los resultados obtenidos en relación a diferentes colecciones de documentos WSDL.

References

[1] R. Baeza-Yates and B. Ribeiro-Neto: Modern Information Retrieval: Addison Wesley, 1999.

[2] Booth, A. D.: A law of occurrences for words of low frequency. Information and control, 10, 386-393. 1967.

[3] M. Bruno, G. Canfora, M. Di Penta, and R. Scognamiglio: An approach to support web services classification and annotation, presented at International Conference on e-Technology, e-Commerce and e-Service (EEE-05), 2005.

[4] X. Dong, A. Halevy, J. Madhavan, E. Nemes, and J. Zhang: Similarity Search for Web Services, presented at 30th VLDB Conference, Toronto, Canada, 2004.

[5] Christiane Fellbaum, *WordNet: An Electronic Lexical Database*, Cambridge, MIT Press. 1998.

[6] Bernhard Ganter, Rudolf Wille, C. Franzke: *Formal Concept Analysis: Mathematical Foundations*, Springer Verlag, 1996.

[7] Galavotti, L.; Sebastiani, F.; Simi, M.: Experiments on the Use of Feature Selection and Negative Evidence in Automated Text Categorization, Proc. of ECDL-00, 4th European Conference on Research and Advanced Technology for Digital Libraries, 59-68. 2000.

[8] Grefenstette, G.: *Explorations in Automatic thesaurus Discovery*, Kluwer Academic Publishers, Boston Hardbound, ISBN 0-7923-9468-2 July 1994.

[9] Héctor Jiménez, David Pinto & Paolo Rosso: Uso del punto de transición en la selección de términos índice para agrupamiento de textos cortos, Procesamiento del Lenguaje Natural No. 35, pp 416-421, España, 2005.

[10] Lehtonen: Preparing Heterogeneous XML for Full-Text Search, ACM Transactions on Information Systems, vol. 24, 2006.

[11] Q. A. Liang and H. Lam: Web Service Matching by Ontology Instance Categorization, presented at IEEE International Conference on Services Computing, vol.1, 2008.

[12] Luhn, H. P. The Automatic Creation of Literature Abstracts. IBM Journal of Research Development, 2(2):159165, 1958.

[13] S. Kutty, R. Nayak, Y. Li: PCITMiner - Prefix-based Closed Induced Tree Miner for Finding closed induced frequent subtrees. Sixth Australasian Data Mining Conference. Australian Computer Society. 2007.

[14] C. Marco, Z. Alejandro and C. Marcelo: Combining Document Classification and Ontology Alignment

[15] P. Henderson and J. Yang: Reusable Web Services, Software Reuse: Methods, Techniques and Tools, Lecture Notes in Computer Science, vol. 3107/2004,pages 185-194, 2004

[16] David Pinto, Héctor Jiménez & Paolo Rosso: Clustering Abstracts of Scientific Texts using the Transition Point Technique, Lecture Notes in Computer Science, Vol. 3878 A. Gelbukh (Ed.), Springer Verlag, pp 536-546, Berlin. 2006.

[17] David Pinto, Héctor Jiménez & Paolo Rosso: UPV-SI: Word Sense Induction using Self Term Expansion, Proceedings of the 4th International .Workshop on Semantic Evaluations (SemEval-2007), pages 430433, Association for Computational Linguistics, Prague, June 2007.

[18] David Pinto, Paolo Rosso, Héctor Jiménez: A self-enriching methodology for clustering narrow domain short texts, The Computer Journal, Oxford, en prensa.

[19] C. Rodríguez Lucatero: Application of ϵ-testers algorithms under sketch and streaming calculation model in robot navigation, Journal WSEAS Transactions on Computers, pag. 1484-1493, 2009.

[20] Franco Rojas López, Héctor Jiménez-Salazar & David Pinto: A competitive term selection method for information retrieval, Lecture Notes in Computer Science, Volume 4394, A. Gelbukh (Ed.), Springer Verlag, pp 468-475, 2007.

[21] M. Sabou, C. Wroe, C. Globle, and G. Mishne: Learning Domain Ontologies for Web Service Descriptions: an Experiment in Bioinformatics, presented at 2005 International World Wide Web Conference, Chiba, Japan, 2005.

[22] C. Sánchez and L. Sheremetov: Semantic Expansion of Service Descriptions, presented at 1st. International Workshop on Quantitative Semantic methods for the Internet, Monterrey, Mexico, Lecture Notes in Computer Science, Springer Verlag, 2008.

[23] C. Sánchez and L. Sheremetov: A Model for Service Discovery with Incomplete Information, presented at Proc. of 5th International Conference on Electrical Engineering, Computing Science and Automatic Control (CCE 2008), México, IEEE Computer Society Press, 2008.

[24] C. Sánchez and L. Sheremetov: A Model for Semantic Service Matching with Leftover and Missing Information. presented at 8th Int. Conf. on Hybrid Intelligent Systems, Barcelona, Spain, IEEE Computer Society Press, 2008.

[25] María Somodevilla, Beatriz Beltrán, David Pinto, Darnes Vilariñ, José Aaron: A first approach to web service discovery, Web Service Discovery Track at INEX Workshop 2010, Shlomo Geva, Jaap Kamps, Ralf Schenkel, Andrew Trotman (Eds.) The Netherlands, http://www.inex.otago.ac.nz/, december 2010.

[26] A. Tagarelli, S. Greco: Semantic Clustering of XML Documents, ACM Transactions on Information Systems, Vol. 28, No. 1, January 2010.

[27] Urbizagástegui, A. R.: Las posibililidades de la ley de zipf en la indización automática. Informe técnico, Universidad de California, Riverside. 1999.

[28] Y. Wang and E. Stroulia: Semantic Structure Matching for Assessing Web-Service Similarity Service-Oriented Computing, presented at ICSOC 2003, Lecture Notes in Computer Science, vol. 2910/2003, 194-207, 2003.

[29] Azucena Montes R., Rocío Vargas A., Hugo Estrada E., Juan G. González S., José Ruiz Ascencio: A method for Automatic Text Categorization using Word Sense Disambiguation. Computational Science and Its Applications ICCSA 2008. LNCS 5073. Springer-Verlag, ISSN 0302-9743, ISBN: 978-3-540-69840-1, Pag 1158-1169, 2008.

[30] Azucena Montes, Alejandro Reyes, Maricela Bravo, Javier Ortiz, Extraction of Semantic Information in Web Services, Polish journal in environmental studies, Miedzyzdroje, Poland, vol 17, No 4C, ISSN 1230-1485, 2008.

[31] J. Reyes Ortiz: Extracción de información semántica para la clasificación de servicios web. Tesis de maestría, CENIDET. 2008.

[32] Kuo-Chung Tai: The tree-to-tree correction problem. Journal of ACM, Vol. 26 (3), 422-433, 1979.

[33] D. Barnard, G. Clarke & N. Duncan: Tree-to-tree Correction for Document Trees. Technical Report 95-372. Department of Computing and Information Science, Queen's University, Canada. 1-44, 1995.

[34] James A. Thom and Chen Wu. Web Service Discovery Track Overview. In Proc. of the INEX 2010 Workshop. Lecture Notes in Computer Science, In press, 2011.

[35] H. Cheng, L. Jun & M. de Rougemont: Approximate validity of XML Streaming Data. Web-age Information, 2008.

7 Metas

7.1 Generación de conocimientos

1. Proponer un método para representar textos híbridos basado en la semántica subyacente (relaciones léxicas). La idea central es aplicar la autoexpansión a los términos, aumentar sus frecuencias y poder extraer relaciones léxico semánticas.

2. Proponer un método para determinar la similitud de dos WS basado en la información estructural. El método base calcula por separado similitudes de las partes de cada WS. La mejora considera incorporar al método de similitud las regularidades que se mantengan entre las partes de los WS.

3. Adaptar y evaluar los métodos de agrupamiento y clasificación accediendo a las relaciones léxico semánticas. En suma, el tesauro enriquecerá los WS a ser agrupados o clasificados.

7.2 Publicación de artículos originales en revistas científicas con arbitraje estricto

Serán realizadas al menos 10 publicaciones con arbitraje estricto. Se estima conveniente someter los métodos desarrollados a la competencia mundial sobre descubrimiento de servicios web: *INEX (INitiative for the Evaluation of XML retrieval)*. Asimismo, se presentarán los resultados en foros especializados como: CicLing, TSD o NLDB. Se harán dos publicaciones en revistas indizadas; al menos tres trabajos se enviarán a revistas indizadas de alto impacto; por ejemplo, *Language Resources and Evaluation, o Natural Language Semantics.*

7.3 Formación de recursos humanos

Este proyecto contribuirá a graduar a dos maestros en ciencias, un doctor, y cuatro licenciados en total. Consideramos la participación de estudiantes de tres niveles: 4 de licenciatura, 2 de maestría y uno de doctorado. En el caso de maestría y doctorado son estudiantes del CENIDET, y en el caso de licenciatura tanto de la BUAP como de la UAM (por determinar).

8 Cronograma de actividades

8.1 Cronograma Anual

En el caso de los estudiantes de licenciatura y maestría los temas de tesis son los siguientes:

EL1 Análisis de la representación de documentos WSDL.

EL2 Selección de términos en documentos WSDL.

EL3 Efectividad de algoritmos de similitud entre documentos WSDL.

EL4 Agrupamiento de documentos WSDL.

EM1 Similitud composicional en colecciones de documentos WSDL.

EM2 Clasificación de documentos WSDL.

El tema de tesis de la estudiante de doctorado es: Evaluación automática de ontologías de dominio restringido. El siguiente programa muestra las actividades por etapas indicando los participantes en cada una de ellas y, al final de la etapa, los productos de investigación generados.

Etapa 1. Proponer un método para representar textos híbridos con base en la semática subyacente.

- **1.1** Cuatrimestre I. Participantes: HJ, CS, DP, EL1, EL2
  Preparar el método: atender la limpieza del documento y la extracción de términos a partir de identificadores (*maximum matching algorithm*), y efectuar un análisis de pruebas que conduzcan a la forma de aplicar la autoexpansión: expansión local o global, relación entre descripción de esquema de mensajes y textos, etc. observando la matriz de confusión.
- **1.2** Cuatrimestre II. Participantes: CS, AL, EL1, EL2
  Evaluar el método en diferentes circunstancias. Por ejemplo, en una prueba de agrupamiento de los textos híbridos deberá cuidarse que la expansión respete la localidad: no expandir los términos que aparezcan en zonas diferentes dentro del documento WSDL.
- **1.3** Cuatrimestre III. Participantes: Todos
  Analizar los resultados para determinar la mejor forma de enriquecer las frecuencias y comparar con otros métodos de selección de términos usando una colección de WS común. En este proceso habrá ajuste de los métodos y elección.

Productos al final de la etapa 1:

- 3 trabajos presentados en eventos especializados con memoria en extenso y arbitraje estricto.
- 2 titulados de nivel licenciatura.
- 1 artículo enviado a una revista indizada de alto impacto.

Etapa 2. Proponer un método para determinar la similitud de dos WS basado en la información estructural.

- **2.1** Cuatrimestre IV. Paticipantes: MT, CS, AM, HJ, EL3, EL4, EM1
  Construir un tesauro a partir de la representación de los términos autoexpandidos de los WS. Para cada término se determinará su contexto, se seleccionarán los términos importantes, y ello permitirá obtener términos relacionados.
- **2.2** Cuatrimestre IV. Participantes: CS, DP, EL3, EL4, EM1
  Evaluar el tesauro con una colección supervisada a través de matrices de confusión
- **2.3** Cuatrimestre V. Participantes; CR, CS, EL3, EL4, EM1
  Desarrollar un método de similitud estructural. Se partirá de un algoritmo apoyado en el cálculo de la similitud de cada parte de una operación del servicio web, y se harán ajustes contemplando las relaciones entre las partes de acuerdo con las relaciones entre términos del tesauro.
- **2.4** Cuatrimestre VI. Participantes: Todos
  Discutir y explicar los resultados obtenidos previamente para concluir la mejor manera de proceder con la representación y el cálculo de la similitud propuesta.

Productos al final de la etapa 2:

- 3 trabajos presentados en eventos especializados con memoria en extenso y arbitraje estricto.
- 2 titulados de nivel licenciatura.
- 1 graduado de nivel maestría.
- 1 artículo enviado a una revista indizada de alto impacto.

Etapa 3. Adaptar y evaluar los métodos de agrupamiento y clasificación accediendo a las relaciones léxico semánticas.

- **3.1** Cuatrimestre VII y VIII. Participantes: CS, HJ, DP, EM2
  Aplicar varios métodos de agrupamiento para conocer el desempeño apoyándose en la función de similitud desarrollada y en varias colecciones de WS; una idea es, por ejemplo, considerar WS con y sin textos en lenguaje natural. Los resultados de la anterior actividad pueden ser un referente (*baseline*).
- **3.2** Cuatrimestre IX. Participantes: Todos
  Evaluar la clasificación en variadas colecciones de WS considerando la construcción de un tesauro. Efectuar un análisis de los desarrollos y su integración en la metodología completa.

Productos al final de la etapa 3:

- 2 trabajos presentados en eventos especializados con memoria en extenso y arbitraje estricto.
- 1 graduado de nivel maestría.
- 1 graduado de nivel doctorado.
- 1 artículo enviado a una revista indizada de alto impacto.
- 1 sitio web orientado a la experimentación sobre el acceso a documentos WSDL.

9 Requerimientos y justificación de los recursos solicitados

9.1 Recursos Humanos

El programador por honorarios tiene la función de integrar los algoritmos desarrollados a un sitio web que permita llevar a cabo experimentos en línea sobre la representación, similitud, agrupamiento y clasificación de documentos WSDL. Este sitio web de experimentación cuenta con las siguientes características:

- elige o carga una colección de documentos WSDL
- selecciona parámetros de prueba: algoritmos (de representación, de selección de términos, función de similitud, de agrupamiento, expansión de consulta, y de clasificación) ajuste para la creación del tesauro.
- crea un ambiente persistente de experimentación: crea sesiones,
- accede a una sesión previa, limpia sesión, genera reporte variados.
- además el sistema podrá clasificar en línea un documento wsdl dado por un usuario.

9.2 Infraestructura, equipamiento y recursos materiales desglosado en cuatrimestres

10 Infraestructura disponible

Se cuenta con la siguiente infraestructura:

- Equipo de cómputo. Todos los participantes cuentan con un equipo de cómputo, además del servicio que ofrece cada institución de acceso general y con equipo más sofisticado. En el caso de los estudiantes, tendrán acceso a equipo en salas destinadas a tesistas.
- Acervo bibliográfico. Tenemos una biblioteca con gran número de textos en el área de investigación. Asimismo hay acceso a texto completo de algunas revistas o colecciones de revistas; por ejemplo *Springer Verlag, Elsevier,* etc.
- *Software.* Se cuenta con *software* general (sistemas operativos, compiladores, procesadores de texto, etc.), y específico al tipo de problemas que deseamos atender (clasificadores, agrupadores, lematizadores, etc.). Dichos recursos se adecuarán a los requerimientos del proyecto o será la base para desarrollar los métodos que se desean probar.
- Colecciones de textos. En el proyecto son indispensables las colecciones de documentos WSDL. Se cuenta, ahora, con dos colecciones WSDL; una proveniente de la competencia INEX (*Web service discovery*) y otra que se ha compilado y clasificado manualmente. Ambas serán utilizadas en las pruebas, pero además es posible considerar dos colecciones más para reforzar los resultados.

En la siguiente figura se presenta el presupuesto para cada una de las acciones en las tres etapas del proyecto. En resumen, tenemos el siguiente concentrado de gastos:

Etapa	Cuatrimestre	Monto
1	1	35,380.27
1	2	115,380.27
1	3	75,380.27
2	1	64,093.87
2	2	144,093.87
2	3	104,093.87
3	1	135,380.27
3	2	135,380.27
3	3	95,380.27

10.1 Opciones adicionales de financiamiento

Este proyecto se ha semetido ante el Consejo Nacional de Ciencia y Tecnología y ha sido aprobado en el marco de la convocatoria CB-2010-01 del sistema de fondos para la investigación de ese organismo en la modalidad de jóven investigador, cuyo responsable es el Profesor Christian Sánchez Sánchez. Número de Solicitud 153315, Modalidad j3, Fondo I0017.

11 Vinculación con los planes y programas de estudio de la DCCD y la Unidad Cuajimalpa

El presente proyecto se vincula de manera directa con varias de las UEAs que se imparten en la Licenciatura en Tecnologías y Sistemas de Información (LTSI) y en otras licenciaturas de la División y de la Unidad Cuajimalpa principalmente la Licenciatura en Diseño (LD) e Ingeniería en Computación (IC). Enseguida se listan las UEAs mencionadas y se indica la manera en que está relacionada con el proyecto. En cado de que la UEA pertenezca a otra licenciatura, se señala entre paréntesis su abreviatura.

- Vinculacion directa
  - Programación de Web Estático
  - Programación de Web Dinámico
  - Laboratorios Tematicos I-III (LTSI,IC)
  - Inteligencia Artificial II. Aprendizaje
  - Seminario de Sistemas Inteligentes I-II
  - Seminario de Tecnologías de la Informacion I-II
  - Proyecto Terminal (LTSI,IC,LD,LCC)

  Los contenidos de estas materias guardan estrecha relación con el proyecto propuesto. En primer lugar, el contenido de las UEAs Programación de Web Dinámico (cuyo antecedente es Programación de Web Estático) versa principalmente con la construcción de sitios web dinámicos con un contenido que permite la interacción del usuario con los diversos tipos de información que el sitio ofrece. El proyecto a su vez toca de manera directa temas de Aprendizaje Automático, principalmente la representación, agrupación y clasificación, los cuales son revisados en la UEA correspondiente. Por otro lado, la tendencia hacia la automatización cada vez mayor de los sistemas web, mencionada anteriormente en este documento, exige la construcción de sistemas que incorporen procedimientos denominados inteligentes, ejemplo de ello son los sitios de búsqueda y recuperación de información que se apoyan en técnicas de procesamiento de lenguaje natural (como ya ha quedado de manifiesto anteriormente), siendo factible de analizar en las UEAs impartidas como seminarios. Además, la investigación referida en este proyecto puede generar diversos estudios con una temática variada tanto para las tres licenciaturas referidas como para la Licenciatura en Ciencias de la Comunicación (LCC), pues puede abordarse como tema en los Proyectos Terminales de todas ellas.

- Vinculacion indirecta
  - Programación Estructurada (LTSI,IC)
  - Estructuras de Datos (LTSI,IC)
  - Programación Orientada a Objetos (LTSI,IC)
  - Análisis y Diseõ de Algoritmos (LTSI,IC)
  - Teoría de Autómatas y Lenguajes Formales
  - Integración de Sistemas
  - Proyecto de Ingeniería de Software (IC)

Además de la vinculación planteada anteriormente, el presente proyecto se relaciona con aquellas UEAs que proporcionan conocimientos sobre el desarrollo de sistemas en general. Tal es el caso de las UEAs de programación y algoritmos, que dan la información necesaria para la construcción de *software* y las UEAs de Integración de Sistemas y Proyecto de Ingeniería de Software, cuyo contenido ofrece la capacidad de construir una solución a partir de sistemas heterogéneos. Por su parte, la materia de lenguajes formales es un antecedente en el tratamiento del lenguaje natural, cuyos métodos son de uso extenso en esta propuesta de investigación.

12 Vinculación institucional

Posibles colaboraciones con otras instituciones de investigacion y/o docencia nacionales o internacionales
En este proyecto se considera la colaboración de investigadores y estudiantes tanto de la Universidad Autónoma de Puebla (BUAP), como del Centro Nacional de Investigación y Desarrollo Tecnológico (CENIDET), cuya participación ya se ha detallado en las secciones previas de este documento. Además, la ejecución de este proyecto habilitará al grupo de Lenguajes y Razonamiento incursionar en temas relacionados con el procesamiento de texto tocando temas como la minería de la web social, lo cual incide en intereses variados como, por ejemplo, el análisis de notas periodísticas de opinión el cual es un tema de interés creciente en diversas áreas de aplicación (análisis de medios, imagen mediática, etc.).