Medio ambiente

## Acuerdo México-Francia relativo a la formación doctoral para la investigación científica, desarrollo tecnológico e innovación, Programa ECOS.

## SEP-CONACYT-ANUIES-ECOS NORD Francia

## CONVOCATORIA 2014

# FORMATO PARA LA PRESENTACIÓN DE PROYECTOS

Remitir, **a más tardar el 16 de mayo**, dos ejemplares **completamente en** **español** (original y una copia), a la Dirección General de Vinculación Estratégica de la ANUIES\*. Adjuntar **obligatoriamente** en los **dos** **ejemplares engargolados** los anexos y los *Curriculum Vitae* completos de los **responsables franceses y mexicanos,** los CVs breves de los demás investigadores tanto **franceses como mexicanos, incluyendo** los CVs del profesor o estudiante de doctorado, involucrados en el proyecto. Anexar como requisito indispensable un CD o USB con el proyecto **completo**, atendiendo las especificaciones señaladas[[1]](#footnote-1) al pie de esta página. De considerarlo pertinente para una mejor presentación de su proyecto, puede anexar el último informe de actividades de su unidad.

**Mantenga sin modificar la disposición de los cuadros.** Favor de aclarar las siglas utilizadas que no son comúnmente conocidas. Este formato de presentación lo puede solicitar en formato Word a la Lic. Argelia De León, Responsable de Proyectos de la Dirección de Relaciones Internacionales, al correo electrónico:[**adleon@anuies.mx**](mailto:adleon@anuies.mx)**.**

**1. Título del proyecto:**

Título (en negritas**): Cartografía de la diáspora calificada aplicando técnicas de minería semántica a resultados de búsquedas en la web**

**Subrayar el área del conocimiento a la cual pertenece**:

1. Educación 2. Artes y Humanidades, 3. Ciencias sociales, administración y derecho, 4. Ciencias naturales, exactas y de la

computación, 5. Ingeniería, manufactura y construcción, 6. Agronomía y veterinaria, 7. Salud, 8. Servicios. (Si el proyecto

abarca dos áreas del conocimiento, le solicitamos señalar a una como prioritaria y otra como secundaria)

**Campo específico: Procesamiento del Lenguaje Natural aplicado a la Sociología**

**Palabras clave (máximo 4): búsqueda de expertos, migración altamente calificada, minería semántica, recuperación de información**

**En México:**

Institución: **Universidad Autónoma Metropolitana**

Laboratorio (o equipo): ): **Departamento de Tecnologías de la** **Información**

Nombre del rector: **Dr. Eduardo Abel Peñalosa Castro**

**En Francia:**

Institución: **Universidad de Paris 13**

Laboratorio **1** (o equipo): **Laboratorio de Informática de Paris Norte (LIPN)**

Situación de la unidad **2**: **UMR 7030 CNRS**

Nombre del rector: **Jean-Loup Salzman**

**2. Instituciones:**

**1** Al que pertenece el responsable científico del proyecto.

**2** Indicar el tipo y el número de la unidad en el organismo INSERM, INRA, CNRS, u otro (ejemplo: **UMR 691 CNRS**).

\* ANUIES, Tenayuca 200, Colonia Santa Cruz Atoyac, Del. Benito Juárez. C.P. 03310, México, D.F. - Tel: 54 20 49 08

**Sitio web:** <http://www.anuies.mx>

**3. Responsables del proyecto 3:**

**En México:**

Nombre y apellido: **Héctor Jiménez Salazar** Grado: **Doctor**

Domicilio institucional y adscripción4: : **Universidad Autónoma Metropolitana, Unidad Cuajimalpa, Av. Vasco de Quiroga 4471, Col. Santa Fé, Cuajimalpa, Delegación Cuajimalpa de Morelos, México, D.F., C.P.05300**

|  |  |  |
| --- | --- | --- |
| Teléfono: **(+52-55) 91776650 ext. 6950.** | Fax: : **()** | Correo electrónico: **hgimenezs@gmail.com** |

**En Francia:**

Nombre y apellido: **Aldo Gangemi** Grado: **Doctor**

Dirección administrativa: **: Université Paris 13, Office B 214, building B , campus of Villetaneuse,**

**99 avenue JB Clément, 93430 Villetaneuse, France**

|  |  |  |
| --- | --- | --- |
| Teléfono**: +33 1 49 40 40 81** | Fax**: +33 1 48 26 07 12** | Correo electrónico**:gangemi@lipn.univ-paris13.fr** |

3. El responsable del proyecto deberá demostrar a través de su experiencia laboral y docente, su liderazgo en un grupo de investigación en él área del conocimiento en cuestión.

4. Requisito indispensable para el envío de notificaciones, proporcionar domicilio completo.

**4. Lista de investigadores que participan en el proyecto, incluyendo a los estudiantes de doctorado:**

**4.1** Señalar nombre y clave del programa de posgrado registrado en el Programa Nacional de Posgrados de Calidad (PNPC) del CONACYT al que pertenece el responsable y los académicos mexicanos participantes en el proyecto**:**

Clave asignada del Programa registrado en el PNPC (Requisito obligatorio): **003441**

Nombre del programa**: Maestría en Diseño, Información y Comunicación**

Nivel: **Maestría**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Nombre | Categoría | Máximo grado académico\* | Unidad de adscripción | Actividad a realizar |
| **En México** |  |  |  |  |
| Héctor Jiménez Salazar | Profesor Investigador de Tiempo Completo Titular C | Doctorado,  CINVESTAV, 2000 | Universidad  Autónoma Metropolitana,  Unidad Cuajimalpa | Co-Dirección de tesis doctoral  Y apoyo metodológico en los  procesos de representación  textual |
| Esaú Villatoro Tello | Profesor Investigador de Tiempo Completo Titular B | Doctorado,  INAOE, 2010 | Universidad  Autónoma Metropolitana,  Unidad Cuajimalpa | Apoyo metodológico en la tarea de visualización de información y aprendizaje automático. |
| David Eduardo Pinto  Avendaño | PI Titular B TC  Tiempo Completo | Doctorado,  Politécnica de  Valencia, 2008 | Facultad de Ciencias  de la Computación, BUAP | Apoyo metodológico en minería de textos, *clustering* de resultados de motores de búsqueda |
| Darnes Vilariño Ayala | PI Titular B  Tiempo Completo | Doctorado | Facultad de Ciencias  de la Computación, BUAP | Apoyo en el desarrollo de  modelos matemáticos para  el análisis cualitativo de la  propagación terminológica. |
| Raymundo Jacobo  Jiménez Muñoz | Estudiante  doctoral | Maestría | Facultad de Ciencias  de la Computación, BUAP | *\*Ver sección 8.5.3* |

\* Indicar Grado, institución y año de titulación.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Nombre | Categoría | Máximo grado académico\* | Unidad de adscripción | Actividad a realizar |
| **En Francia** |  |  |  |  |
| Aldo Gangemi | Profesor Titular | Doctorado | LIPN, Universidad de Paris 13 | Machine Reading, RDF/OWL, Web Semántica, FRED. |
| Davide Buscaldi | Profesor | Doctorado  Politécnica de  Valencia, 2010 | LIPN, Universidad de Paris 13 | Proximidad semántica, extracción de información. |
| Jorge J. García Flores | Ingeniero de  Investigación | Doctorado  Paris-Sorbonne  2007 | LIPN, Universidad de Paris 13 | Anotación de resultados de motores de búsqueda, gramáticas formales para el reconocimiento de nombres propios. |
| Dana Diminescu | Profesor | Doctorado | Télécom-Paristech | Validación de cartografías |
| Ehab Hassan | Doctorando | Maestría  Paris-Dauphine  2011 | LIPN, Universidad de París 13 | Extracción de eventos |

**5. Resumen del proyecto (el proyecto será descrito detalladamente en la Sección 8.**

**La localización de los emigrantes altamente calificados representa un problema crucial para los países en vías de desarrollo. Además de las técnicas tradicionalmente utilizadas en sociología de la migración, como las encuestas de campo o los censos poblacionales, una solución atractiva para este problema parte de la utilización de técnicas computacionales de minería semántica de textos con la finalidad de cartografiar la trayectoria profesional de los miembros de la diáspora altamente calificada a partir de información publicada en la web. Esta tarea representa un reto científico en diversas áreas del procesamiento del lenguaje natural: búsqueda y recuperación de información biográfica; desambiguación de nombres propios, términos científicos y topónimos; cálculo de proximidad semántica entre los fragmentos cortos de texto característicos de los motores de búsqueda, clasificación terminológica por disciplina del conocimiento y extracción de eventos. Nuestro objetivo es doble: por un lado cartografiar de manera individual la trayectoria profesional (geográfica y temática) de los miembros de la diáspora altamente calificada a partir de la escasa información disponible en bases de datos institucionales (como la base de ex becarios del CONACYT) y, por otro lado, permitir la búsqueda de expertos pertenecientes a la diáspora altamente calificada a partir de consultas por palabras clave (ejemplo: buscar un experto en frenado por corrientes de Foucault, de nacionalidad mexicana, de preferencia formado en Francia). La experiencia conjunta de los equipos de investigación asociados al proyecto ofrece un amplio abanico metodológico para cimentar la colaboración: en principio conformaremos un marco experimental con datos provenientes de fuentes institucionales (CONACYT, CONICET, COLCIENCIAS) así como de proyectos de investigación en sociología de las diásporas calificadas (CIDESAL). A continuación estudiaremos de manera crítica los métodos estadísticos y lingüísticos apropiados para llevar a cabo la minería semántica de los resultados de un motor de búsqueda. A partir de este estudio implementaremos algoritmos de clasificación de grupos de resultados (*clustering*) y de desambiguación por medio de WSI (*Word Sense Induction*) con la finalidad de extraer las relaciones semánticas entre un nombre propio, las instituciones, los lugares geográficos y las fechas que conforman una trayectoria profesional. De particular importancia será la caracterización epistemológica de las áreas de experiencia utilizando ontologías del conocimiento y cálculos de proximidad semántica. El resultado de este conjunto será un grafo RDF construido por medio de técnicas de *machine reading*, en donde dichas relaciones semánticas sean representadas. Este grafo servirá de base para la generación de la cartografía. El fin último de nuestra propuesta consiste en sentar las bases metodológicas para una nueva generación de herramientas de búsqueda de expertos sensibles al contexto biográfico y geográfico propio de cada persona.**

**6. Otros apoyos, públicos o privados, recibidos por parte de otros organismos\*:**

¿Ha participado anteriormente con un proyecto ECOS? Sí □ No ⌧

En caso de una respuesta afirmativa, indicar la clave, el año y el nombre del proyecto:

Anteriormente:

En curso:

Otros apoyos públicos o privados, recibidos por parte de otros organismos:

\* En Francia y en México

**7. Medios:**

**7.1 Medios solicitados en el marco del Programa SEP-CONACYT-ECOS-ANUIES para el primer año del proyecto 5:**

*a - Francia → México:*

Misiones solicitadas para el 1er año del proyecto para investigadores **con doctorado** confirmados, mencionando a los beneficiarios y la duración (máximo 15 días):

* Aldo Gangemi, misión de 7 días
* Davide Buscaldi, misión de 7 días
* Jorge García Flores, misión de 15 días

Estancias (de 1 a 2 meses) para investigadores en formación (doctorandos):

* Ehab Hassan (1 mes)

*b - México- Francia:*

Misiones solicitadas para el 1er año del proyecto para investigadores **con doctorado** confirmadas, mencionando a los beneficiarios y la duración (máximo 15 días):

* Héctor Jiménez Salazar, misión 7 días
* David Eduardo Pinto Avendaño, misión de 15 días
* Esaú Villatoro Tello, misión de 7 días

Estancias (de 1 a 2 meses) para investigadores en formación (doctorandos):

7.2 Formación doctoral: Considerando que es un requisito que el proyecto incluya la participación de por lo menos un doctorando mexicano, docente, o estudiante de una institución de educación superior mexicana o centro de investigación mexicano (indique los estudiantes mexicanos y franceses considerados en el marco del proyecto, mencionando sin falta a los beneficiarios, el organismo que financiaría la beca doctoral, y las características del doctorado. Se prestará especial atención a los doctorados en co-tutela.

*a – Profesores o estudiantes mexicanos que estudiarán el doctorado, y el organismo al que se le solicitará la beca6:*

**Raymundo Jacobo Jiménez Muñoz, doctorado por co-tutela**

**Becas CONACYT-Gobierno francés 2014**

Raymundo Jacobo Jiménez Muñoz *ha cursado estudios de maestría en la Facultad de Ciencias de la Computación de la BUAP, México, en temas de recuperación de información semántica de textos utilizando técnicas de minería de textos y ontologías bajo la dirección de Dr. David Pinto. El perfil que posee el estudiante cumple con los requerimientos de este proyecto de investigación.*

*b – Estudiantes franceses y el organismo al que se le solicitará la beca:*

**4  E**n el caso que corresponda, precise si la beca ya ha sido otorgada y el organismo la atorgó.

**5** Los recursos solicitados en el marco de la convocatoria se refieren a acciones, no a montos, ya que éstos han sido establecidos previamente para cada actividad por las instancias financiadoras.

MF= misión francesa o MM= misión mexicana (Máximo 15 días, para investigadores) o EF= estancia francesa o EM= estancia mexicana (Máximo 60 días, para estudiantes de doctorado).

6 En el caso que corresponda, precise si la beca ya ha sido otorgada y el organismo que la otorgó. Indicar la edad del estudiante de doctorado.

#### FIRMAS

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Nombre, apellido y firma del **responsable mexicano del proyecto**

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

Nombre, apellido y firma del **responsable de la unidad de investigación**

a la que pertenece el responsable del proyecto

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

Nombre, apellido, sello y firma del **titular de la institución**

o de su representante debidamente autorizado

1. **Descripción detallada del proyecto:**

En esta última parte no se deberá alterar el orden de las secciones. En caso de que sea utilizado un tratamiento de texto, se podrá ajustar libremente el tamaño de las secciones, intercalando entre ellas el número necesario de páginas, así como de anexos.

IMPORTANTE: la información presentada en México, debe coincidir con la presentada por su contraparte en Francia, con el fin de evitar valoraciones y apreciaciones muy diferentes por parte de los comités evaluadores de ambos países.

**8.1 Identificación del tema y estado del conocimiento:**

La localización de la emigración altamente calificada es un problema de vital importancia para los países en vías de desarrollo. Más allá de las fuentes demográficas, como los censos de población, las encuestas de campo sobre población activa, las fuentes gubernamentales (registros de permisos de trabajo para extranjeros, registros de visados) y las estadísticas de la [OCDE 2002], una solución atractiva para este problema se apoya en técnicas computacionales de recuperación semántica de información con el objetivo de cartografiar la trayectoria profesional de los miembros de la diáspora del conocimiento exclusivamente a partir de información publicada en Internet.

La primera aplicación de tecnologías del lenguaje humano al campo de la sociología de la migración fue el sistema *Unoporuno[[2]](#footnote-2)* [Garcia Flores et al. 2012], desarrollado en el marco del proyecto CIDESAL[[3]](#footnote-3). Este sistema permite clasificar automáticamente a potenciales miembros de la diáspora altamente calificada de acuerdo a su movimiento migratorio: 1) un movimiento lineal (se van del país para estudiar o trabajar y ya no regresan); 2) un movimiento circular (tras los estudios, regresan a su país); 3) ausencia de movimiento (la persona nunca sale de su país de origen). *Unoporuno* recibe de entrada una lista de nombres en contexto (por ejemplo: un conjunto de fichas bibliográficas). A partir de este conjunto de datos, *Unoporuno* hace minería del web con la finalidad de clasificar a las personas en cuestión por medio de procesos de aprendizaje automático según su movimiento. Gracias a los socios del proyecto CIDESAL, se logró reunir una base de aproximadamente 13000 miembros potenciales de la diáspora altamente calificada. Los registros de esta base provienen tanto de una aspiración de la base de publicaciones científicas del *ISI Web of Science* como de bases de datos de exbecarios que los organismos científicos nacionales lationamericanos proporcionaron: CONICET en Argentina, COLCIENCIAS en Colombia, SNI en Uruguay y CONACYT en México [CIDESAL 2014]. El principal límite de *Unoporuno* es que no hay recuperación alguna de información semántica, sino solamente un proceso de clasificación automática por medio de un análisis rasgos semánticos de los resultados de una consulta automática en un motor de búsqueda.

En cuanto a al hecho de cartografiar a la diáspora utilizando los datos disponibles en Internet, los avances más recientes fueron llevados a cabo durante el proyecto *e-Diáspora Atlas[[4]](#footnote-4)* o “atlas de la diáspora electrónica”. Este proyecto consistió en censar y analizar algo más de 8000 sitios web creados o frecuentados por los emigrantes del mundo entero. A partir de este análisis fueron producidos mapas de hipervínculos que caracterizaban a las diásporas nacionales según la topología creada sus hipervínculos [Diminescu 2008].

El sistema *Name-sor[[5]](#footnote-5)* es otro ejemplo significativo de aplicación de técnicas cartográficas al análisis de los movimientos migratorios. Este sistema se apoya en un análisis onomástico del origen geográfico de los nombres propios de los usuarios de *Twitter* [Shokhenmayer 2010] con el objetivo de producir un mapa de la densidad de las distintas diásporas nacionales representadas por los usuarios de esta red social.

Curiosamente ningún sistema de búsqueda de expertos (*expert finding*) ha sido nunca aplicado a la búsqueda de emigrantes altamente calificados. Esto se debe quizás a que estos sistemas no suelen tomar en cuenta la información biográfica de los expertos, puesto que sus esfuerzos metodológicos se concentran en recuperar relaciones epistemológicas estables entre un experto y su ámbito de experiencia (o dominio). A nosotros nos parece que el pasaje entre una búsqueda de expertos y una búsqueda de personas en Internet (emigrantes o no) forma parte de un continuo, y que las herramientas de búsqueda de expertos ganarían precisión si tomaran en cuenta el contexto biográfico de los expertos-objeto de su búsqueda (el país de origen, el país de formación, la trayectoria profesional). Y, de manera recíproca, el ámbito de especialidad o la pura profesión podría ser considerado como un valioso elemento de desambiguación de homónimos en las tareas de búsqueda de personas por Internet. Dos aplicaciones de búsqueda de expertos destacan en el panorama de esta tarea: *Saffron* [Monaghan et al. 2010] y *Sig.ma* [Tummarello et al. 2010]. *Saffron* aplica el procesamiento del lenguaje natural a la anotación de páginas web pertinentes para la búsqueda de expertos, mientras que *Sig.ma* explota los metadatos de las mismas. Ambos hacen abundante uso de las representaciones semánticas en RDF para caracterizar el horizonte epistemológico del ámbito de especialidad de los expertos.

Los mejores resultados reportados en la literatura en lo que a minería semántica de resultados de motores de búsqueda se refiere fueron obtenidos con el método de inducción del sentido de las palabras (*Word Sense Induction* o *WSI*) [Di Marco & Navigli 2013], que se apoya en un análisis de capas de coocurrencia léxica y en el uso de *DBpedia* y *Babelnet* para desambiguar y agrupar los resultados.

En cuanto a la minería de relaciones semánticas necesarias para reconstruir la trayectoria profesional de una persona, diversas herramientas han sido propuestas. Citaremos aquí dos de ellas: *Text Runner* [Yates et al. 2007] recupera información de un texto por medio de una cadena de procesamiento que incluye el análisis morfosintáctico de las frases que componen el texto, el reconocimiento de sintagmas nominales y la producción de un conjunto de binomios representativos de las relaciones semánticas en el texto. En la misma línea, *FRED* [Gangemi et al. 2014] es una herramienta de “lectura a máquina” (*machine reading*) capaz de extraer la mayor cantidad de información posible de un texto (entidades nombradas, taxonomías, relaciones, eventos, sentido) y de traducir el todo a un esquema RDF en donde los datos interpretados son vinculados a una fuente abierta de información como *DBpedia*.

**8.1.1. Originalidad**

La principal innovación de nuestra propuesta radica en el hecho de aplicar los últimos avances en procesamiento del lenguaje natural, como la minería semántica y la “lectura a máquina”, al problema real de localizar a los expertos pertenecientes a la diáspora calificada. Si los resultados esperados se realizan, nuestra aportación epistemológica sería doble: por el lado de las tecnologías semánticas daría lugar a una nueva generación de métodos de búsqueda de expertos que se apoyen un poco más en la información biográfica del experto; por el lado de la sociología el resultado produciría una nueva fuente de datos posible para los estudiosos de la migración, no datos de campo sino datos provenientes exclusivamente del *web.* Más allá del objetivo concreto y particular de cartografiar de manera individual la trayectoria geográfica, profesional y temática de los altamente calificados, sería interesante buscar una correlación estadística robusta entre las tendencias migratorias calculadas a partir de información del *web* y la tendencia real que se observa en las cifras de campo que reportan la OCDE o la ONU. Este resultado permitiría dar respuesta a una pregunta clásica en métodos numéricos: ¿Hasta qué punto la realidad del web permite responder a cuestiones sobre la realidad en el terreno? [Rogers 2010]

Se hace aquí necesario abrir un paréntesis para mencionar la originalidad de los socios que conforman nuestro proyecto. El LIPN cuenta con vasta experiencia en métodos que combinan el procesamiento del lenguaje natural y la representación del conocimiento. Esta vocación de construir representaciones semánticas a partir de anotaciones textuales se ha visto reforzada con la incorporación del sistema FRED y de los métodos de *machine reading* a las actividades cotidianas de investigación del laboratorio. Esta experiencia de traducción de frases cortas no estructuradas en grafos RDF/OWL haciendo uso de fuentes de conocimientos extralingüísticos estructuradas como *DBpedia* hace posible considerar la aplicación de un tratamiento semántico profundo para aquellos resultados arrojados por el motor de búsqueda Internet y considerados pertinentes para reconstruir la trayectoria profesional de una persona. Además, el laboratorio posee un buen conocimiento de los métodos de producción masiva de búsquedas automáticas en Internet, puesto que la última versión del sistema *Unoporuno* ha sido también desarrollada por el LIPN. La participación anual a la tarea de similitud semántica de SEMEVAL permite también considerar la integración de cálculos de proximidad semántica en los métodos de *clustering* de resultados del motor de búsqueda. En cuanto a la pertinencia de las cartografías de las diásporas calificadas, la presencia de un equipo como el Télécom ParisTech se vuelve indispensable. Del lado mexicano, tanto el equipo de investigación de Tecnologías de la Información de la UAM-Cujimalpa como el de Ingeniería del Lenguaje y el Conocimiento de la BUAP tienen una amplia experiencia en clasificación por *clustering* y desambiguación de mensajes de texto que bien pueden adaptarse a los textos cortos que caracterizan los resultados de los motores de búsqueda Internet. El equipo de la UAM-Cuajimalpa ha trabajado también en la recuperación de información geográfica, etapa indispensable en la producción de cartografías a partir de los datos biográficos disponibles en Internet.

**8.2 Objetivos:**

* Desarrollar métodos computacionales que permitan cartografiar automáticamente la trayectoria individual de profesionales expertos con la finalidad de localizar aquellos científicos que posean ciertas habilidades y/o conocimientos específicos. Asimismo, la visualización de dicha trayectoria científica, nos permitirá estudiar diversos fenómenos, por ejemplo, el de la migración científica.
* Establecer un mecanismo que permita producir datos con anotación semántica mediante una secuencia de tratamiento de información y *machine reading* que lleve los datos desde la expresión textual basada en lenguaje natural, hasta una representación estructurada basada en el conocimiento (RDF) a través de técnicas de aprendizaje automático y *clustering* usando el método *WSI* [Di Marco & Navigli 2013]. Esta representación RDF será la base para cartografiar la trayectoria de los expertos.
* Producir cartografías y estadísticas en base a los datos agregados de las cartografías individuales que permitan comparar estas tendencias y calcular el grado de correlación estadística entre las tendencias del *web* y las provenientes de los datos de campo de la OCDE y la ONU.

**8.2.1 Objetivos científicos particulares**

* Partir de los corpora y recursos metodológicos de proyectos de investigación en sociología de las diásporas calificadas(CIDESAL) para construir el cartografiado individual de los expertos científicos que la conforman.
* Extraer **información** semántica válida de nombres propios en contextos semánticamente ricos (por ejemplo: una ficha bibliuográfica); extraer por ejemplo topónimos, temas de investigación o nombres de organizaciones que, al ser combinados con el nombre de la persona y enviados al motor de búsqueda web, arrojen información biográfica pertinente.
* Integrar la herramientas de “lectura a máquina profunda” (*deep machine learning*) FRED [Etzioni et al. 2006 ; Gangemi et al. 2014] al sistema de búsqueda de la diáspora del conocimiento *Unoporuno* [Garcia Flores et al. 2012] para extender la funcionalidad de éste último al caso de uso tradicional de la búsqueda de expertos (*expert finding*) [Bordea & Buitelaar 2012].
* Reforzar la recuperación de elementos semánticos a partir de resultados del motor de búsqueda con métodos de aprendizaje automático.
* Construir las cartografías individuales de expertos científicos, evaluando los resultados obtenidos en base a los métodos propuestos por CIDESAL.
* Cambiar de escala a partir de las cartografías individuales construidas con la finalidad de producir agregados que permitan determinar, por ejemplo cuales fueron los países de preferencia y la tasa de migración de expertos que pertenecen a cierto país de origen.
* Calcular la correlación entre las tendencias de los agregados y las cifras poblacionales de organismos internacionales tales como la OCDE y la ONU.

**8.2.2 Objetivos institucionales**

* Fomentar la cooperación bilateral entre instituciones de educación superior mexicanas y francesas.
* Promover la generación de convenios activos entre instituciones que permitan mantener los vínculos de trabajo a largo plazo.
* Crear una sinergia entre ambos grupos de investigación, en este campo interdisciplinario, al unir expertos en computación, lingüística computacional y medio ambiente
* Contribuir en la formación de investigadores y doctorantes implicados en el proyecto en materia de cartografía temática de tópicos del medio ambiente usando métodos computacionales
* Incidir en los programas de posgrado en proceso de cooperación, fortaleciendo las líneas de investigación abordadas en los diversos laboratorios

**8.3 Metodología:**

**8.3.1 Metodología científica:**

Como paso inicial, definiremos las teorías, conceptos o ideas que pretendemos verificar, particularmente, la dilucidación de propuestas para la construcción de cartografías individuales para expertos científicos. Una tarea inicial consiste en el estudio exhaustivo del estado del arte y la evaluación preliminar de las líneas de acción, a fin de establecer un proyecto de culminación factible en tiempo y forma de acuerdo a los objetivos planteados.

Posteriormente, haremos una comprobación experimental de las hipótesis y confrontación de los resultados. En particular, usaremos métricas clásicas de evaluación para los sistemas de tratamiento automático de información que incluyen, entre otras, precisión, recall, F-Measure, mean average precisión, así como medidas de correlación entre los datos obtenidos y aquellos reportados en la literatura. En resumen, nos interesa llevar a cabo un análisis y evaluación no subjetivo, de manera tal que se corroboren adecuadamente las hipótesis planteadas.

Como tercer paso, criticaremos los métodos y técnicas desarrolladas, aceptando o rechazando las hipótesis planteadas bajo la explicación de los resultados experimentales y la teoría que subyace a los métodos para la identificación de la trayectoria científico-profesional de expertos, así como al tipo de entidades que utilizan (personas).

Finalmente, se presentarán las ventajas y desventajas de los métodos y técnicas desarrolladas, así como sus posibles aplicaciones en otros tipos de problemas.

En particular, hemos definido las siguientes hipótesis que deben ser evaluadas:

Consideramos que los sistemas de recuperación de información de imágenes que reciben consultas mediante una imagen se verán beneficiados en términos de evocación y/o precisión por la implementación de un mecanismo de representación basado en bolsa de palabras visuales.

Por tanto, se formulan las siguientes hipótesis:

**1. Es posible cartografiar la trayectoria científica de un experto, así como su movilidad profesional a partir del análisis de información presente en Internet (estructurada, semi-estructurada y no estructurada).**

**2. El modelo de cartografía de la diáspora calificada generado a partir del presente proyecto podrá ser usado para ubicar personas de acuerdo a sus conocimientos y habilidades.**

**3. Existe una correlación entre la movilidad detectada automáticamente por los modelos generados en el presente proyecto y las estadísticas mostradas en cifras poblacionales por organismos internacionales tales como la OCDE y la ONU.**

**8.3.2 Metodología de desarrollo:**

A continuación se enumeran los pasos que se seguirán para desarrollar el proyecto:

1) En primera instancia se constituirá un conjunto de datos de prueba (*testbed*) constituido por un número de personas suficiente como para poder realizar inferencia estadística.

2) Se reconstruirá el buscador de personas especializado en movilidad profesional “Unoporuno”, para usarlo como aparato de observación.

3) Se usará la salida del sistema actual “Unoporuno” como un criterio base (*baseline*) sobre el cual medir el rendimiento de los nuevos sistemas desarrollados.

4) Experimentar con métodos del Procesamiento del Lenguaje Natural (PLN) y minería de textos, complementado con técnicas de aprendizaje automático (Machine Reading), a fin de constituir una plataforma de tratamiento automático y masivo de información.

5) Modelizar el campo epistemológico del ámbito de experiencia utilizando los métodos usuales en búsqueda de expertos [Aleman-Meza et al. 2007 ; Li et al. 2006].

6) Integrar herramientas que traducen sentencias escritas en lenguaje natural a ontologías RDF/OWL con los sistemas de búsqueda de personas a fin de mejorar el proceso de refinamiento de la información de la información asociada a expertos científicos. Este paso presenta un reto mayor, debido a que, por ejemplo, FRED sólo trabaja en lengua inglesa, por lo que habrá que hacer una trasposición de recursos del inglés [Banea et al. 2008] a otras lenguas, o implementar versiones ligeras de FRED.

7) Evaluar la extracción de los datos mínimos para reconstituir una carrera científica/profesional: nacionalidad, primera formación, formación especializada, ámbito d especialización. Comparar con otras herramientas.

8) Producir datos en formatos semánticamente estructurados tipo RDF.

9) Utilizar las herramientas de visualización de datos (por ejemplo, Cytoscape[[6]](#footnote-6) o Cartodb[[7]](#footnote-7)) para cartografiar la movilidad científica y profesional de los expertos.

10) Generar datos estadísticos y mapas a gran escala a partir de las cartografías generadas.

11) Calcular la correlación entre las cifras obtenidas por el sistema a partir del *web*  y aquellas mostradas en cifras poblacionales por organismos internacionales tales como la OCDE y la ONU.

**8.3.3 Referencias bibliográficas**

[Aleman-Meza et al. 2007] Boanerges Aleman-Meza , Uldis Bojars , Harold Boley , John G. Breslin , Malgorzata Mochol , Lyndon Jb Nixon , Axel Polleres , Anna V. Zhdanova. Combining RDF vocabularies for expert finding. In Proceedings of the 4th European Semantic Web Conference (ESWC2007), number 4519 in Lecture Notes in Computer Science, (2007).

[Banea et al. 2008] Carmen Banea, Rada Mihalcea, Janyce Wiebe and Samer Hassan, Multilingual Subjectivity Analysis Using Machine Translation, in Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2008), Honolulu, Hawaii, October 2008.

[CIDESAL, 2014] CIDESAL : Construction d’Incubateurs des Diasporas du Savoir en Amérique Latine.

Rapport final d’activité. EuroAid/Commission Européenne, 2014 (in press)

[Bordea & Buitelaar 2012] Georgeta Bordea, Paul Buitelaar. Expertise Mining. Proceedings of the 21st National Conference on Artificial Intelligence and Cognitive Science, Galway, Ireland, 2010.

[Di Marco & Navigli 2013] Antonio Di Marco, Roberto Navigli: Clustering and Diversifying Web Search Results with Graph-Based Word Sense Induction. Computational Linguistics 39(3): 709-754 (2013)

[Diminescu 2008] Diminescu D. The connected migrant: an epistemological manifesto. Social Science Information (December 2008) vol. 47 no. 4 565-579

[Etzioni et al. 2006] Oren Etzioni, Michele Banko and Michael J Cafarella. Machine Reading. AAAI Conference on Artificial Intelligence, 2006.

[Gangemi et al. 2014] Aldo Gangemi, Valentina Presutti, Diego Reforgiato Recupero. Frame-based detection of opinion holders and topics: a model and a tool. IEEE Computational Intelligence, 9(1), 2014

[Garcia Flores et al. 2012] J. J. Garcia Flores, Pierre Zwigenbaum, Zhao Yue and W.A. Turner (2012) Tracking Researcher Mobility on the Web Using Snippet Semantic Analysis. In H. Ishara & K. Kanzaki, Advances in Natural Language Processing 8th International Conference on NLP, October 22-24, Kanazawa, Japan. (JapTAL 2012) Springer Lecture Notes in Computer Sciences, vol. 7614, ISBN 978-3-642-33983-7, pp. 180-191.

[Li et al. 2006] Jie Li, Harold Boley, Virendrakumar C. Bhavsar, and Jing Mei. Expert Finding for eCollaboration Using FOAF with RuleML Rules. Montreal Conference on eTechnologies MCTECH, (May 17, 2006)

[Monaghan et al. 2010] Fergal Monaghan, Georgeta Bordea, Krystian Samp, Paul Buitelaar (2010)

Exploring Your Research: Sprinkling some Saffron on Semantic Web Dog Food, 9th International Semantic Web Conference (ISWC), Semantic Web Challenge, Shanghai, China.

[OECD, 2002] International Mobility of the Highly Skilled. Policy Brief, OECD, 2002, Paris.

[Rogers 2010] Richard Rogers, "Internet Research: The Question of Method," Journal of Information Technology and Politics 7 (2010): 241-260.

[Shokhenmayer 2010] Evgeny Shokhenmayer Cartography in onomastic research / TRANS vol.18; Virtual INST World Conference. CCKS: Cities, Cultures, Knowledge Societies. Section ‘Cartographic Science as a Reflection of Knowledge Societies and Cultures’, 2010

[Tummarello et al. 2010] Giovanni Tummarello, Richard Cyganiak, Michele Catasta, Szymon Danielczyk, Renaud Delbru, Stefan Decker "Sig.ma: Live views on the Web of Data", Journal of Web Semantics: Science, Services and Agents on the World Wide Web - Volume 8, Issue 4, November 2010, Pages 355-364

[Yates et al. 2007] Alexander Yates, Michael Cafarella, Michele Banko, Oren Etzioni, Matthew Broadhead, and Stephen Soderland. 2007. TextRunner: open information extraction on the web. In Proceedings of Human Language Technologies: The Annual Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations (NAACL-Demonstrations '07)

**8.4 Calendarización de las actividades del proyecto, CRONOGRAMA:**

|  |  |  |
| --- | --- | --- |
| Año | Actividad | Objetivo |
| 1 | * Estado del arte * *Testbed* a partir del corpus CIDESAL (Conacyt + Conicet + Colciencias) * Actualización de Unoporuno * Implementación local de FRED * Guía ética para la manipulación de datos biográficos provenientes del *web* * Casos de uso | Desarrollar un marco teórico, metodológico y tecnológico de análisis y evaluación de la calidad de datos del proyecto. |
| 2 | * *Clustering* de los resultados de *Unoporuno* usando *WSI* (*Word Sense Induction*) * Ontología de ámbitos de experiencia/ compilation de recursos epistemológicos (léxicos, diccionarios, tesauros) * Anotación de entidades nombradas + evaluación * Extracción de eventos + evaluación * Gramática formal para el reconocimiento de nombres de personas en español, francés, inglés y alemán + evaluación | Identificar la metodología adecuada (incluyendo herramientas y técnicas) para usar métodos de minería semántica que permitan extraer información de calidad/relevante para el dominio en estudio de este proyecto. |
| 3 | * Integración *Unoporuno+FRED* * *Crosslingual FRED* * Producción de RDF + evaluación * Primeras cartografías individuales | Construir una herramienta multilingue que genere el grafo RDF y las primeras cartografías individuales |
| 4 | * Cartografías agregadas * Estadísticas agregadas * Correlación entre datos provenientes del *web* y datos de campo (OCDE) * Evaluación | Cartografías y estadísticas agregadas.  Correlación web vs. campo |

|  |  |
| --- | --- |
| Año | Acciones |
| 1 | 1. Misión de 15 días para 1 investigador mexicano en Francia. 2. Misión de 7 días para 2 investigadores mexicanos en Francia 3. Misión de 15 días para 1 investigador francés en México 4. Misión de 7 días para 2 investigadores franceses en México 5. Misión de 1 mes para 1 doctorando francés en México 6. Publicación conjunta de artículos de posicionamiento |
| 2 | 1. Misión de 7 días para 3 investigadores mexicanos en Francia 2. Misión de 7 días para 3 investigadores franceses en México 3. Misión de 7 días para 1 doctorando francés en México 4. Publicación conjunta de resultados científicos |
| 3 | 1. Misión de 7 días para 2 investigadores mexicanos en Francia 2. Misión de 7 días para 2 investigadores franceses en México 3. Participación en comités tutoriales de tesis de maestría y doctorado 4. Publicación conjunta de resultados científicos |
| 4 | 1. Organización de un coloquio Franco-Mexicano de doctorantes e investigadores participantes, en el marco de la temática del proyecto 2. Defensa de tesis del doctorando financiado por CONACYT 3. Publicación conjunta de resultados científicos |

**8.5 Resultados esperados y beneficios globales del proyecto, (publicaciones, eventos, etc.), considerando, entre otros, la contribución al desarrollo científico de la unidad académica mexicana que lo propone, la formación de recursos humanos, y la consolidación institucional de área de docencia e investigación.**

**Resultados esperados:**

* Corpus y método de evaluación empírica de las cartografías individuales de trayectoria profesional a partir de la *web*.
* Método para cartografiar a las diásporas del conocimiento a partir de los datos disponibles en la *web.*
* Sistema de producción de cartografías individuales de expertos, ya sea a partir del nombre de un experto o a partir de un ámbito de conocimiento y algunas restricciones biográficas (ejemplo: *busco a un experto en frenado por corrientes de Foucault de nacionalidad mexicana y que haya estudiado en Francia*).
* Ontología epistemológica de ámbitos de experiencia
* Publicaciones conjuntas entre el equipo de México y de Francia.
* Organización de un coloquio Franco-Mexicano de doctorantes e investigadores participantes, en el marco de la temática del proyecto.
* Formación de recursos humanos (estudiantes e investigadores) en temas relacionados con la minería de textos aplicables al dominio de las ciencias ambientales.

8.5.1. Indicar el impacto institucional y social de los resultados alcanzados en el proyecto.

El desarrollo de una cartografía que permita visualizar la trayectoria profesional de la diáspora calificada permitiría a los organismos que financian la investigación tener una idea más clara del grado de emigración y el grado de retorno de los exbecarios.

Los países en vías de desarrollo contarían con un método relativamente menos costoso que los censos o las encuestas de campo para localizar su talento.

Las pequeñas y medianas empresas mexicanas podrían contar con una herramienta informática que les permita vincularse con la diáspora altamente calificada, de forma y manera que esta podría vincularse con y movilizarse por el desarrollo nacional.

En términos institucionales, la integración de un proyecto de investigación que sería desarrollado de manera conjunta entre nuestra institución y un centro de investigación francés sería de mucho beneficio puesto que permitiría intercambiar conocimientos, herramientas y recursos que redundaría sin duda en un avance académico y de investigación importante.

Finalmente, dado que se contempla la formación doctoral de al menos un estudiante mexicano que será dirigido por co-tutela, este tendrá una incorporación a futuro en México que tendrá también un impacto a nivel social.

**8.5.2. Especificar la participación de cada investigador y el producto de su trabajo**

|  |  |  |
| --- | --- | --- |
| Año | Actividad | Investigador |
| 1 | * Estado del arte * *Testbed* a partir del corpus CIDESAL (Conacyt + Conicet + Colciencias) * Actualización de Unoporuno * Implementación local de FRED * Guía ética para la manipulación de datos biográficos provenientes del *web* * Casos de uso | Raymundo J. Jiménez M.  Ehab Hassan  Jorge Garcia Flores  Aldo Gangemi  David Pinto  Héctor Jiménez Salazar  Dana Diminescu  Davide Buscaldi  Esaú Villatoro Tello |
| 2 | * *Clustering* de los resultados de *Unoporuno* usando *WSI* (*Word Sense Induction*) * Ontología de ámbitos de experiencia/ compilation de recursos epistemológicos (léxicos, diccionarios, tesauros) * Anotación de entidades nombradas + evaluación * Extracción de eventos + evaluación * Gramática formal para el reconocimiento de nombres de personas en español, francés, inglés y alemán + evaluación | Raymundo J. Jiménez M.  Ehab Hassan  Jorge Garcia Flores  Aldo Gangemi  David Pinto  Esaú Villatoro Tello  Héctor Jiménez Salazar  Darnes Vilariño |
| 3 | * Integración *Unoporuno+FRED* * *Crosslingual FRED* * Producción de RDF + evaluación * Primeras cartografías individuales | Raymundo J. Jiménez M.  Jorge Garcia Flores  Aldo Gangemi  David Pinto  Esaú Villatoro Tello  Héctor Jiménez Salazar  Dana Diminescu |
| 4 | * Cartografías agregadas * Estadísticas agregadas * Correlación entre datos provenientes del *web* y datos de campo (OCDE) * Evaluación | Raymundo J. Jiménez M.  Jorge Garcia Flores  David Pinto  Héctor Jiménez Salazar  Aldo Gangemi |

**8.5.3. Describir las actividades del estudiante de doctorado a realizar en el proyecto y especificar claramente el plan de trabajo del mismo.**

|  |  |  |
| --- | --- | --- |
| Año | Actividades | Objetivo |
| 1 | * Estudio del estado del arte sobre las herramientas actuales en *expert finding*. * Estudio de técnicas para la extracción terminológica en dominios específicos y ontologías. * Análisis de representaciones textuales efectivos en la extracción de información. | Identificar los conceptos teóricos necesarios para el desarrollo del proyecto en cuestión. |
| 2 | * Desarrollo de una metodología para la minería semántica en el dominio específico del problema a tratar. * Integración entre FRED y Unoporuno. | Identificar la metodología adecuada (incluyendo herramientas y técnicas) para usar métodos de minería de textos que permitan extraer información de calidad/relevante para el dominio en estudio de este proyecto. |
| 3 | * Desarrollo de las herramientas de cartografía para la visualización de las trayectorias profesionales individuales. * Cytoscape. | Construir una herramienta de visualización cartográfica biográfica para el dominio de las diásporas calificadas. |
| 4 | * Análisis cualitativo de la agregación cartográfica y estadística. | Evaluar la exactitud de las trayectorias profesionales construidas. |

**8.6 Antecedentes de apoyo:**

El apoyo mencionado brevemente en la Sección 6, deberá ser ampliado en esta sección. En caso de ser pertinente, anexar las referencias de las publicaciones de las que sea coautor, tesis, equipamientos realizados, proyectos conjuntos, patentes, etc.

**8.7 Publicaciones:**

Lista de cinco publicaciones significativas de los dos equipos y/o de los responsables de éstos, en los últimos cinco años (relacionadas o no con el proyecto):

**Equipo Francés:**

* **Aldo Gangemi**

1. Ciaramita M, Gangemi A, Ratsch E, Rojas I, Saric J. Unsupervised Learning of Semantic Relations between Concepts of a Molecular Biology Ontology, in Proceedings of International Joint Conference on Artificial Intelligence (IJCAI2005), 2005.
2. Baldassarre C., Daga E., Gangemi A., Gliozzo A., Salvati A., Troiani G. Semantic Scout: Making Sense of Organizational Knowledge. P. Cimiano and S: Pinto (eds.): Proceedings of the Conference on Knowledge Engineering and Knowledge Management (EKAW2010), Springer, 2010.
3. Valentina Presutti, Francesco Draicchio and Aldo Gangemi. Knowledge extraction based on Discourse Representation Theory and Linguistic Frames. A. ten Teije and J. Völker (eds.): Proceedings of the Conference on Knowledge Engineering and Knowledge Management (EKAW2012), LNCS, Springer, 2012.
4. Aldo Gangemi. A Comparison of Knowledge Extraction Tools for the Semantic Web. Proceedings of ESWC2013, LNCS, Springer, 2013.
5. Andrea Giovanni Nuzzolese, Valentina Presutti, Aldo Gangemi, Alberto Musetti and Paolo Ciancarini. Aemoo: exploring knowledge on the Web. Proceedings of WebSci2103 (ACM Web Science Conference), ACM, 2013.

* **Dana Diminescu**

1. Diminescu D. (ed.): Special issue: Diasporas on the Web (2012) Social Science Information Journal, December 2012; 51 (4) Publisher: SAGE, on behalf of Maison des Sciences de l’Homme, Paris, ISSN 0539-0184, 245 pages.
2. Diminescu D. The connected migrant: an epistemological manifesto. Social Science Information (December 2008) vol. 47 no. 4 565-579
3. Diminescu D., 2007 « Le passage par l’écran : ou l’émergence de nouvelles frontières », in Actes du colloque Les frontières de l’Europe, Ed Universitatii, Bucarest, pp 263-274
4. Diminescu D., (2006), « L’usage du téléphone portable par les migrants en situations précaire », in Quatre ans de recherche urbaine (2001-2004), ACI-Ville Min de la recherche, E. Bajolet, MF Mattei, JM Rennes (sld ), Presse Universitaire François Rabelais, MSH Ville et territoire, Tours, 2006, pp90-98
5. Diminescu D., (2006), « Genèse d’une figure de migrant, in Cosmopolitiques n°11/2006, ed Apogée, Paris, pp. 63-73

* **David Buscaldi**

1. Davide Buscaldi and Paulo Rosso. A conceptual density-based approach for the disambiguation of toponyms. Int. J. Geogr. Inf. Sci., 22(3) :301–313, 2008.
2. Davide Buscaldi, Paolo Rosso, José Manuel Gómez, and Emilio Sanchis. Answering questions with ann-gram based passage retrieval engine. Journal of Intelligent Information Systems (JIIS), 34(2) :113–134, 2009.
3. Paolo Rosso, Santiago Correa, and Davide Buscaldi. Passage retrieval in legal texts. Journal of Logic and Algebraic Programming, 80(3-5) :139 – 153, 2011.
4. Guillaume Cleuziou, Davide Buscaldi, Vincent Levorato, and Gaël Dias. A pretopological framework for the automatic construction of lexical-semantic structures from texts. In CIKM ’11 : Proceedings of the 20th ACM international conference on Information and knowledge management, pages 2453–2456, New York, NY, USA, 2011. ACM.
5. Davide Buscaldi. Approaches to disambiguating toponyms. SIGSPATIAL Special, 3(2) :16–19, 2011.

* **Jorge J. Garcia Flores**

1. J. J. Garcia Flores, Pierre Zwigenbaum, Zhao Yue and W.A. Turner (2012) Tracking Researcher Mobility on the Web Using Snippet Semantic Analysis. In H. Ishara & K. Kanzaki, Advances in Natural Language Processing 8th International Conference on NLP, October 22-24, Kanazawa, Japan. (JapTAL 2012) Springer Lecture Notes in Computer Sciences, vol. 7614, ISBN 978-3-642-33983-7, pp. 180-191.
2. J. Garcia-Flores and W.A. Turner (2011) A Nominal Filter for Web Search Snippets: Using the Web to Identify Members of Latin America’s Highly Qualiﬁed Diaspora. In I.Z. Batyrshin & G. Sidorov, Mexican International Conference on Artificial Inteligence (MICAI 2011), IEEE press, ISBN: 978-1-4577-2173-1, pp. 45-50.
3. B. Djioua, J.J. Garcia Flores, A. Blais, J-P. Desclés, G. Guibert, A Jackiewicz., F. Le Priol, L. Nait-Baha, B. Sauzay (2006) EXCOM: an automatic annotation engine for semantic information. In G. Sutcliff & R. Goebel (eds.): Florida Artificial Intelligence Research Society Conference (FLAIRS 2006), Association for the Advancement of Artificial Intelligence Press, ISBN 978-1-57735-261-7, pp. 285-290 .
4. J.J. Garcia-Flores. (2004) Semantic Filtering of Textual Requirements Descriptions. In F. Méziane & E. Métais (editores): Natural Language Processing and Informations Systems (NLDB 2004), Springer, Lecture Notes in Computer Science, ISBN 978-3-540-22564-5, pp. 427-434 .

**Equipo mexicano**

* **Hector Jiménez Salazar**

1. [Esaú Villatoro-Tello](http://www.informatik.uni-trier.de/~ley/pers/hd/v/Villatoro=Tello:Esa=uacute=.html), [Christian Sánchez-Sánchez](http://www.informatik.uni-trier.de/~ley/pers/hd/s/S=aacute=nchez=S=aacute=nchez:Christian.html), Héctor Jiménez-Salazar, [Wulfrano Arturo Luna-Ramírez](http://www.informatik.uni-trier.de/~ley/pers/hd/l/Luna=Ram=iacute=rez:Wulfrano_Arturo.html), [Carlos Rodríguez-Lucatero](http://www.informatik.uni-trier.de/~ley/pers/hd/r/Rodr=iacute=guez=Lucatero:Carlos.html): UAM at INEX 2012 Relevance Feedback Track: Using a Probabilistic Method for Ranking Refinement. [CLEF (Online Working Notes/Labs/Workshop) 2012](http://www.informatik.uni-trier.de/~ley/db/conf/clef/clef2012w.html#Villatoro-TelloSJLR12)
2. [David Pinto](http://www.informatik.uni-trier.de/~ley/pers/hd/p/Pinto:David.html), [Darnes Vilariño Ayala](http://www.informatik.uni-trier.de/~ley/pers/hd/a/Ayala:Darnes_Vilari=ntilde=o.html), [Yuridiana Alemán](http://www.informatik.uni-trier.de/~ley/pers/hd/a/Alem=aacute=n:Yuridiana.html), [Helena Gómez-Adorno](http://www.informatik.uni-trier.de/~ley/pers/hd/g/G=oacute=mez=Adorno:Helena.html), [Nahun Loya](http://www.informatik.uni-trier.de/~ley/pers/hd/l/Loya:Nahun.html), Héctor Jiménez-Salazar: The Soundex Phonetic Algorithm Revisited for SMS Text Representation. [TSD 2012](http://www.informatik.uni-trier.de/~ley/db/conf/tsd/tsd2012.html#PintoAAGLJ12): 47-55
3. David Pinto, Paolo Rosso, Héctor Jiménez-Salazar: A Self-enriching Methodology for Clustering Narrow Domain Short Texts. Computer Journal. 54(7): 1148-1165 (2011)
4. [David Pinto](http://www.informatik.uni-trier.de/~ley/pers/hd/p/Pinto:David.html), [Paolo Rosso](http://www.informatik.uni-trier.de/~ley/pers/hd/r/Rosso:Paolo.html), Héctor Jiménez-Salazar: On the Assessment of Text Corpora. [NLDB 2009](http://www.informatik.uni-trier.de/~ley/db/conf/nldb/nldb2009.html#PintoRJ09): 281-290
5. [Alberto Bañuelos-Moro](http://www.informatik.uni-trier.de/~ley/pers/hd/b/Ba=ntilde=uelos=Moro:Alberto.html), Héctor Jiménez-Salazar, [José de Jesús Lavalle-Martínez](http://www.informatik.uni-trier.de/~ley/pers/hd/l/Lavalle=Mart=iacute=nez:Jos=eacute=_de_Jes=uacute=s.html): Test of complementarity on sentence extraction methods.[Procesamiento del Lenguaje Natural 40](http://www.informatik.uni-trier.de/~ley/db/journals/pdln/pdln40.html#Banuelos-MoroJL08) (2008)

* **David Eduardo Pinto Avendaño**

1. David Pinto, [Helena Gómez-Adorno](http://www.informatik.uni-trier.de/~ley/pers/hd/g/G=oacute=mez=Adorno:Helena.html), [Darnes Vilariño Ayala](http://www.informatik.uni-trier.de/~ley/pers/hd/a/Ayala:Darnes_Vilari=ntilde=o.html), [Vivek Kumar Singh](http://www.informatik.uni-trier.de/~ley/pers/hd/s/Singh:Vivek_Kumar.html): A graph-based multi-level linguistic representation for document understanding. [Pattern Recognition Letters 41](http://www.informatik.uni-trier.de/~ley/db/journals/prl/prl41.html" \l "PintoGAS14): 93-102 (2014)
2. Paolo Rosso, Marcelo Errecalde, David Pinto: Analysis of short texts on the Web: introduction to special issue. Language Resources and Evaluation 47(1): 123-126 (2013)
3. David Pinto, Paolo Rosso, Héctor Jiménez-Salazar: A Self-enriching Methodology for Clustering Narrow Domain Short Texts. Computer Journal. 54(7): 1148-1165 (2011)
4. Claudia Denicia-Carral, Manuel Montes-y-Gómez, Luis Villaseñor-Pineda and David Pinto-Avendaño, Bilingual Document Clustering: Evaluating Cognates as Features, Journal of Information and Library Science, 35(3): 265-286, 2011
5. David Pinto, Jorge Civera, Alberto Barrón-Cedeño, Alfons Juan, Paolo Rosso: A statistical approach to crosslingual natural language tasks. Journal of Algorithms 64(1): 51-60 (2009)

* **Esaú Villatoro Tello**

1. [Fernando Sánchez-Vega](http://www.informatik.uni-trier.de/~ley/pers/hd/s/S=aacute=nchez=Vega:Fernando.html), Esaú Villatoro-Tello, [Manuel Montes-y-Gómez](http://www.informatik.uni-trier.de/~ley/pers/hd/m/Montes=y=G=oacute=mez:Manuel.html), [Luis Villaseñor Pineda](http://www.informatik.uni-trier.de/~ley/pers/hd/p/Pineda:Luis_Villase=ntilde=or.html), [Paolo Rosso](http://www.informatik.uni-trier.de/~ley/pers/hd/r/Rosso:Paolo.html): Determining and characterizing the reused text for plagiarism detection. [Expert Syst. Appl. 40](http://www.informatik.uni-trier.de/~ley/db/journals/eswa/eswa40.html" \l "Sanchez-VegaVMPR13)(5): 1804-1813 (2013)
2. Esaú Villatoro-Tello, [Antonio Juárez](http://www.informatik.uni-trier.de/~ley/pers/hd/j/Ju=aacute=rez:Antonio.html), [Manuel Montes-y-Gómez](http://www.informatik.uni-trier.de/~ley/pers/hd/m/Montes=y=G=oacute=mez:Manuel.html), [Luis Villaseñor Pineda](http://www.informatik.uni-trier.de/~ley/pers/hd/p/Pineda:Luis_Villase=ntilde=or.html), [Luis Enrique Sucar](http://www.informatik.uni-trier.de/~ley/pers/hd/s/Sucar:Luis_Enrique.html): Document ranking refinement using a Markov random field model. [Natural Language Engineering 18](http://www.informatik.uni-trier.de/~ley/db/journals/nle/nle18.html#Villatoro-TelloJMPS12)(2): 155-185 (2012)
3. Esaú Villatoro-Tello, [Christian Sánchez-Sánchez](http://www.informatik.uni-trier.de/~ley/pers/hd/s/S=aacute=nchez=S=aacute=nchez:Christian.html), [Héctor Jiménez-Salazar](http://www.informatik.uni-trier.de/~ley/pers/hd/j/Jim=eacute=nez=Salazar:H=eacute=ctor.html), [Wulfrano Arturo Luna-Ramírez](http://www.informatik.uni-trier.de/~ley/pers/hd/l/Luna=Ram=iacute=rez:Wulfrano_Arturo.html), [Carlos Rodríguez-Lucatero](http://www.informatik.uni-trier.de/~ley/pers/hd/r/Rodr=iacute=guez=Lucatero:Carlos.html): UAM at INEX 2012 Relevance Feedback Track: Using a Probabilistic Method for Ranking Refinement. [CLEF (Online Working Notes/Labs/Workshop) 2012](http://www.informatik.uni-trier.de/~ley/db/conf/clef/clef2012w.html#Villatoro-TelloSJLR12)
4. Esaú Villatoro-Tello, [Ricardo Omar Chávez García](http://www.informatik.uni-trier.de/~ley/pers/hd/g/Garc=iacute=a:Ricardo_Omar_Ch=aacute=vez.html), [Manuel Montes-y-Gómez](http://www.informatik.uni-trier.de/~ley/pers/hd/m/Montes=y=G=oacute=mez:Manuel.html), [Luis Villaseñor Pineda](http://www.informatik.uni-trier.de/~ley/pers/hd/p/Pineda:Luis_Villase=ntilde=or.html), [Luis Enrique Sucar](http://www.informatik.uni-trier.de/~ley/pers/hd/s/Sucar:Luis_Enrique.html): A Probabilistic Method for Ranking Refinement in Geographic Information Retrieval. [Procesamiento del Lenguaje Natural 44](http://www.informatik.uni-trier.de/~ley/db/journals/pdln/pdln44.html#Villatoro-TelloGMPS10) (2010)
5. [Maya Carrillo](http://www.informatik.uni-trier.de/~ley/pers/hd/c/Carrillo:Maya.html), Esaú Villatoro-Tello, [Aurelio López-López](http://www.informatik.uni-trier.de/~ley/pers/hd/l/L=oacute=pez=L=oacute=pez:Aurelio.html), [Chris Eliasmith](http://www.informatik.uni-trier.de/~ley/pers/hd/e/Eliasmith:Chris.html), [Luis Villaseñor Pineda](http://www.informatik.uni-trier.de/~ley/pers/hd/p/Pineda:Luis_Villase=ntilde=or.html), [Manuel Montes-y-Gómez](http://www.informatik.uni-trier.de/~ley/pers/hd/m/Montes=y=G=oacute=mez:Manuel.html): Concept Based Representations for Ranking in Geographic Information Retrieval. [IceTAL 2010](http://www.informatik.uni-trier.de/~ley/db/conf/tal/icetal2010.html" \l "CarrilloVLEPM10): 85-96

* **Darnes Vilariño Ayala**

1. Darnes Vilariño Ayala, David Pinto, Beatriz Beltrán, Saul León, Esteban Castillo, Mireya Tovar: A Machine-Translation Method for Normalization of SMS. MCPR 2012: 293-302
2. David Pinto, [Helena Gómez-Adorno](http://www.informatik.uni-trier.de/~ley/pers/hd/g/G=oacute=mez=Adorno:Helena.html), [Darnes Vilariño Ayala](http://www.informatik.uni-trier.de/~ley/pers/hd/a/Ayala:Darnes_Vilari=ntilde=o.html), [Vivek Kumar Singh](http://www.informatik.uni-trier.de/~ley/pers/hd/s/Singh:Vivek_Kumar.html): A graph-based multi-level linguistic representation for document understanding. [Pattern Recognition Letters 41](http://www.informatik.uni-trier.de/~ley/db/journals/prl/prl41.html" \l "PintoGAS14): 93-102 (2014)
3. David Pinto, Darnes Vilariño, Carlos Balderas, Mireya Tovar and Beatriz Beltran, Evaluating n-gram Models for a Bilingual Word Sense Disambiguation Task, Computación y Sistemas 15(2), 2011
4. Darnes Vilariño Ayala, David Pinto, Mireya Tovar, Carlos Balderas, Beatriz Beltrán: A Probabilistic Model Based on n-Grams for Bilingual Word Sense Disambiguation. MICAI (1) 2010: 82-91
5. David Pinto, Darnes Vilariño, Carlos Balderas, Mireya Tovar, and Beatriz Beltrán: A Naive Bayes Approach to Cross-Lingual Word Sense Disambiguation and Lexical Substitution. MCPR 2010. [LNCS 6256, Springer-Verlag](http://www.springerlink.com/content/) : 352-361

**8.8 Información complementaria:**

**Marque con una “X” las respuestas positivas**

Este proyecto ha sido presentado en su totalidad, o parcialmente, a otro organismo (indicar los recursos obtenidos):

Este proyecto ha sido aprobado por: durante el período comprendido de: hasta:

Este proyecto no ha sido presentado a ningún otro organismo, ni siquiera de manera parcial.

1. El CD o USB debe contener la misma propuesta presentada en papel, organizado de la siguiente manera:

   1. Archivo con la propuesta en PDF (Formato debidamente cumplimentado), máximo 5 MB

   2. Carpeta con los CVs del equipo mexicano en PDF (Responsable, integrantes y doctorando), máximo 5 MB

   3. Carpeta con los CVs del equipo francés en PDF (Responsable, integrantes y doctorando), máximo 5 MG

   4. Carpeta que contenga la propuesta y los CVs de ambos equipos en formato Word, máximo 8 MB [↑](#footnote-ref-1)
2. <http://unoporuno.limsi.fr/unoporuno> [↑](#footnote-ref-2)
3. Construcción de Incubadoras para las Diásporas del Saber en América Latina: <http://bit.ly/1nLmhzo> [↑](#footnote-ref-3)
4. [http://www.e-diasporas.fr](http://www.e-diasporas.fr/)/ [↑](#footnote-ref-4)
5. <http://namesorts.com> [↑](#footnote-ref-5)
6. <http://www.cytoscape.org/> [↑](#footnote-ref-6)
7. <http://cartodb.com> [↑](#footnote-ref-7)