El Procesamiento del Lenguaje Natural para extraer conocimiento de la Web, Documentos y Redes sociales

Dr. José Luis Ochoa Hernández

Departamento de Ingeniería Industrial
Universidad de Sonora
Hermosillo, Sonora, México

Reunión de otoño CUDI 2012 CHIHUAHUA 29 y 30 DE OCTUBRE

29 y 30 DE OCTUBRE 2012 ● CHIHUAHUA

Contenido

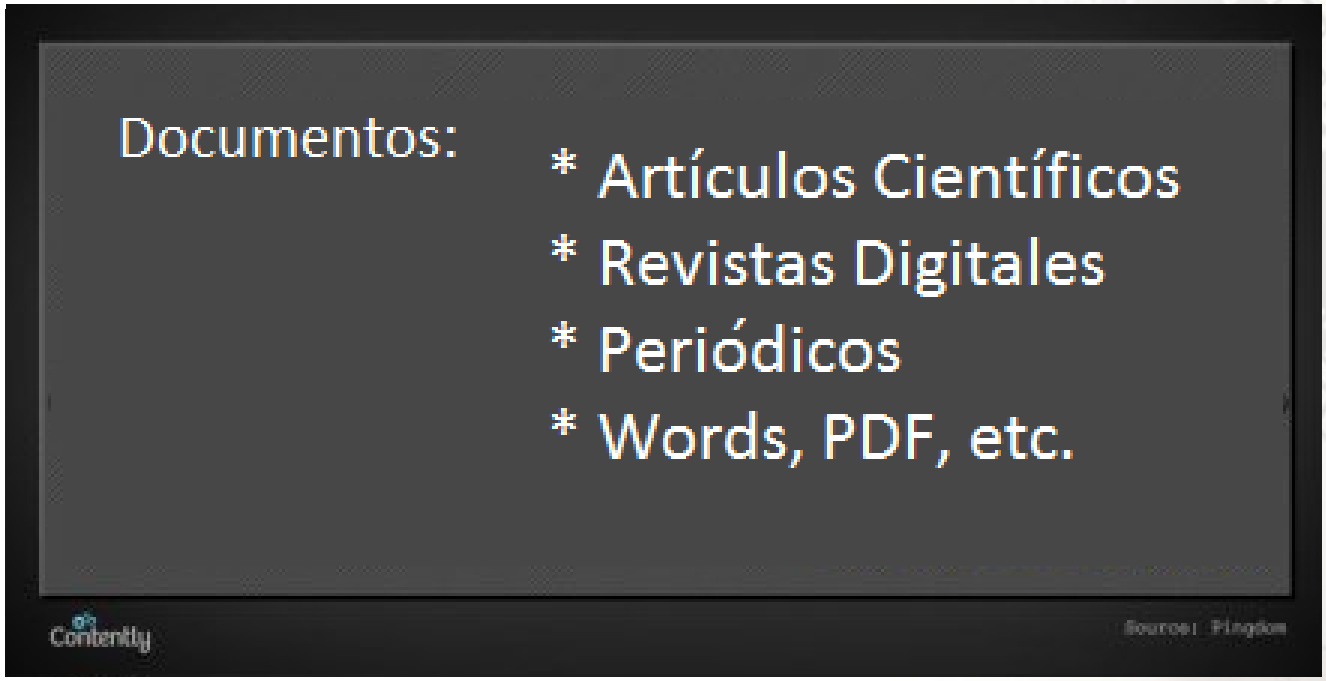
- Introducción
- Objetivos
- Metodología
- Herramienta de SW
- Resultados

INTRODUCCIÓN

Problema Actual

- Obtener información válida, es una tarea fundamental en diversos contextos.
- Actualmente la información digital se expande exponencialmente.
- Internet cuenta con más de **644.275.754** páginas Web.
- Con mucha frecuencia, las páginas duplican la información y no siempre es correcta, consistente o completa.
- El idioma y la polisemia constituyen algunos de los principales problemas en recuperación de información.
- La Web Semántica se plantea como una solución, pero ….. hay que crearla.

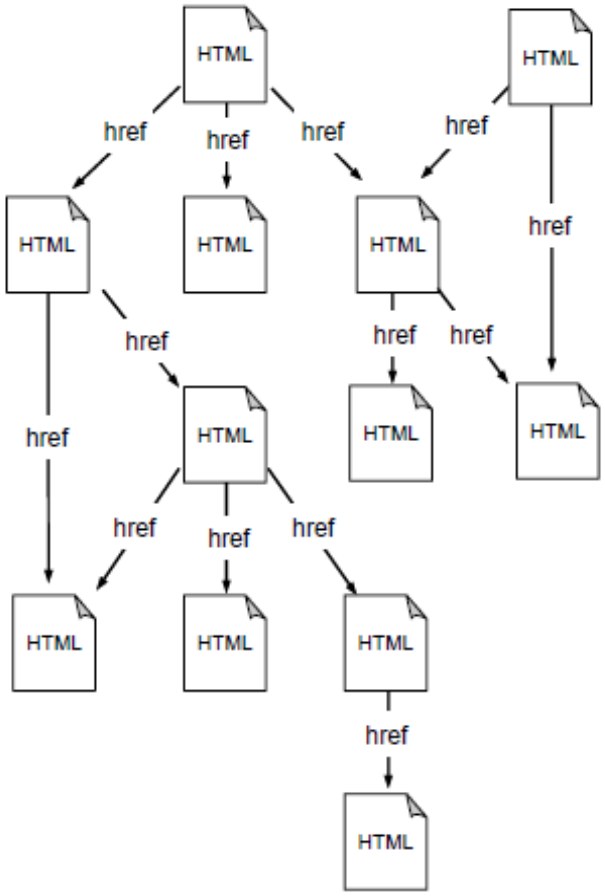
¿Qué es el Contenido Digital?



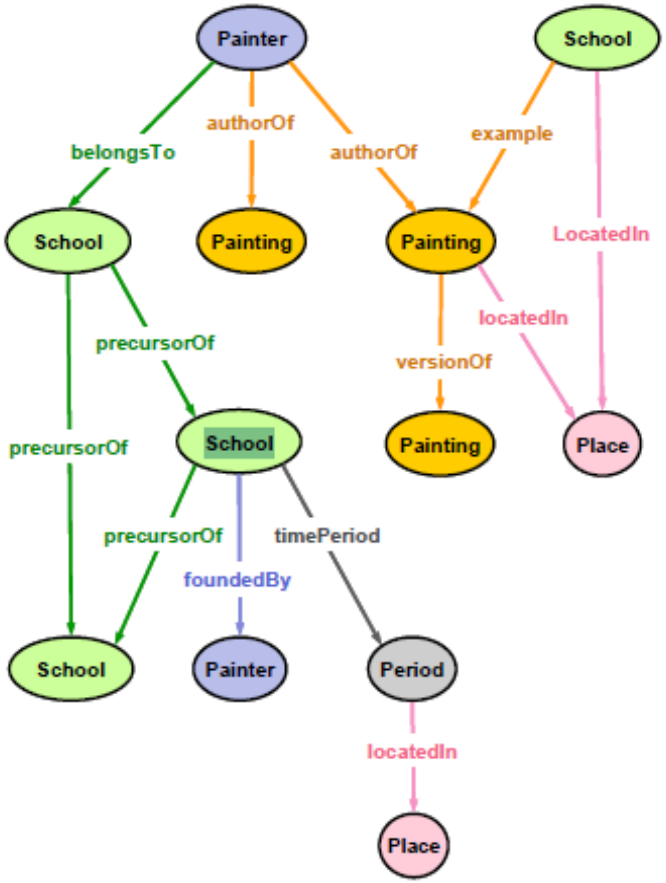
Posibles Soluciones

- La Web Semántica
  - Ontologías
    - Metodologías para construir Ontologías
      - Metodología para construir ontologías a partir de cero
      - Metodologías para reingeniería ontológica
      - Metodologías de extracción de ontologías a partir de texto en lenguaje natural

Web semántica



a) Web actual



b) Web semántica

a) Web actual

b) Web semántica

Metodologías Actuales

Ontology Learning	
En Ingles	**[Abascal-Mena, 2009]** Towards a semantic web: ontology development based on the extraction of semantic concepts from digital documents.
En Chino	**[Lee et al., 2007]** Automated ontology construction for unstructured text documents.
En Tailandés	**[Kawtrakul et al., 2004]** Automatic Thai Ontology Construction and Maintenance System.
En Persa	**[Khosravi and Vazifedoost, 2007]** Creating a Persian Ontology through Thesaurus Reengineering for Organizing the Digital Library of the National Library of Iran.
En Alemán	**[Bontas et al., 2005]** Creating ontologies for content representation — the OntoSeed suite.
En Francés	**[Passant, 2007]** Using Ontologies to Strengthen Folksonomies and Enrich Information Retrieval in Weblogs.
En Español	**[Valencia-Garcia et al., 2006]** A Methodology for Extracting Ontological Knowledge from Spanish Documents.

Objetivos

Objetivos General

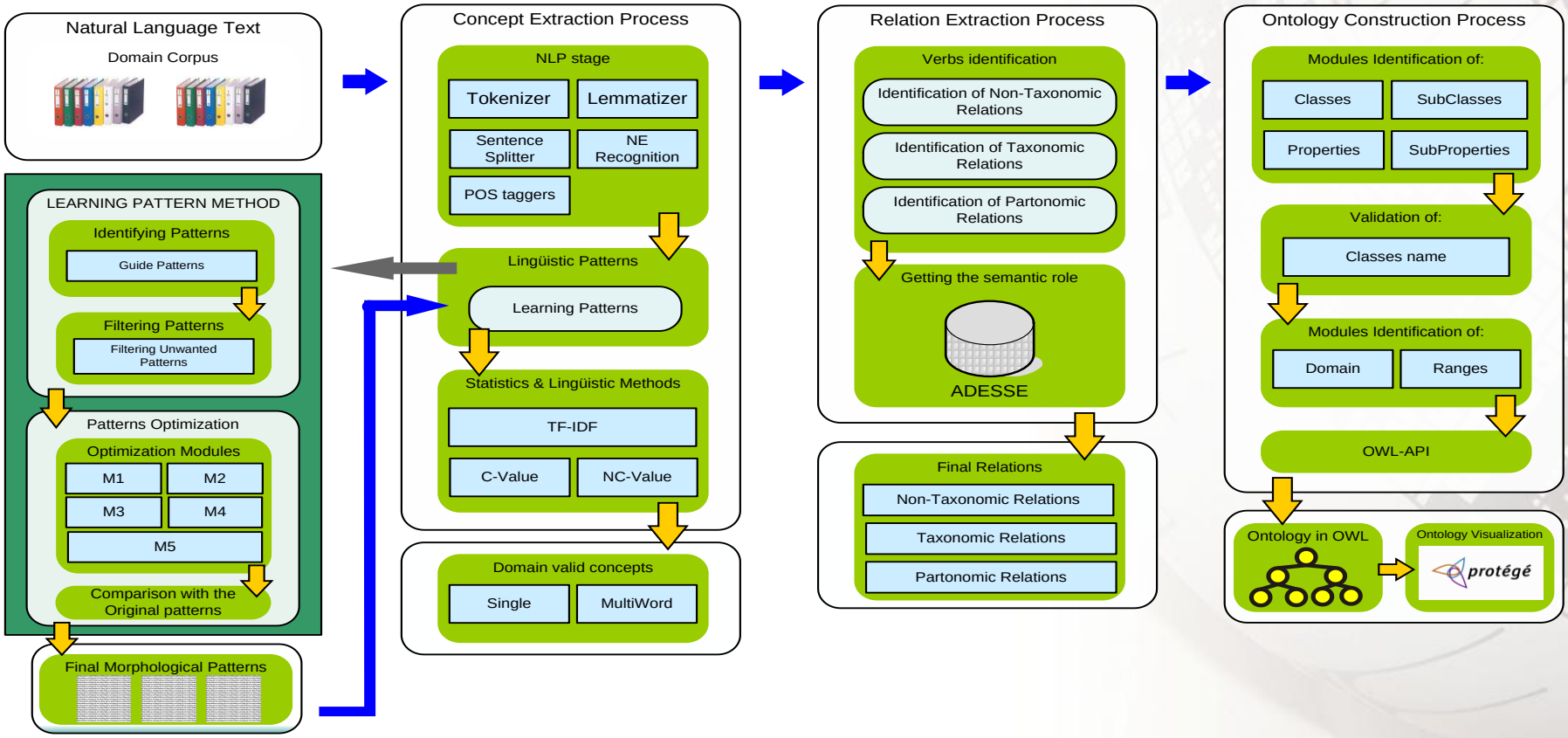
- Desarrollo de un método para la construcción automática de ontologías a partir de textos escritos en Lenguaje Natural, que tenga en cuenta un amplio conjunto de relaciones semánticas entre conceptos, de forma independiente del dominio y en el lenguaje español.

A tomar en cuenta…..

- **Independencia del dominio:** Que sea posible generar ontologías en cualquier dominio.
- **Adaptabilidad y parametrización:** El sistema debe permitir la configuración de diversos aspectos del sistema.
- **Eficiencia y flexibilidad:** El sistema podrá intercambiar las herramientas externas fácilmente.

Metodología

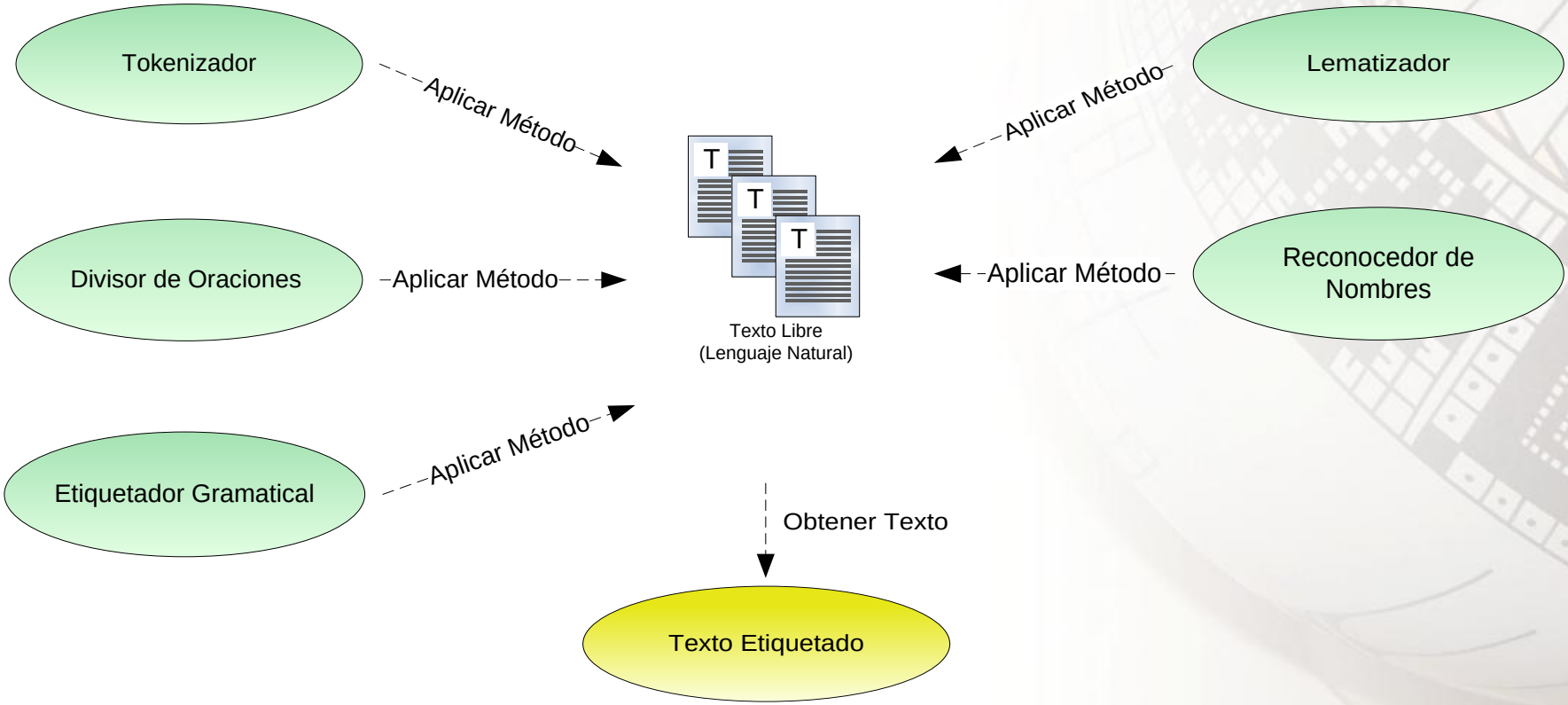
Arquitectura Desarrollada



Módulo - Extracción de conceptos

- **Fase de PLN**
- Fase de patrones lingüísticos
- Fase de extracción de conceptos compuestos
- Fase de extracción de conceptos simples

Fase de Procesamiento del Lenguaje Natural



Fase de Procesamiento del Lenguaje Natural

(Ejemplo)

El procedimiento para designar al nuevo Rector es establecido en el presente estatuto.

el•el•DA0MS0	*es•ser•VSIP3S0*
procedimiento•procedimiento•NCMS000	*establecido•establecer•VMP00SM*
para•para•SPS00	*en•en•SPS00*
designar•designar•VMN0000	*el•el•DA0MS0*
a•a•SPS00	*presente•presente•AQ0CS0*
el•el•DA0MS0	*estatuto•estatuto•NCMS000*
nuevo•nuevo•AQ0MS0	*. •. •Fp*
rector•rector•NCMS000	

Módulo - Extracción de conceptos

- Fase de PLN
- **Fase de patrones lingüísticos**
- Fase de extracción de conceptos compuestos
- Fase de extracción de conceptos simples

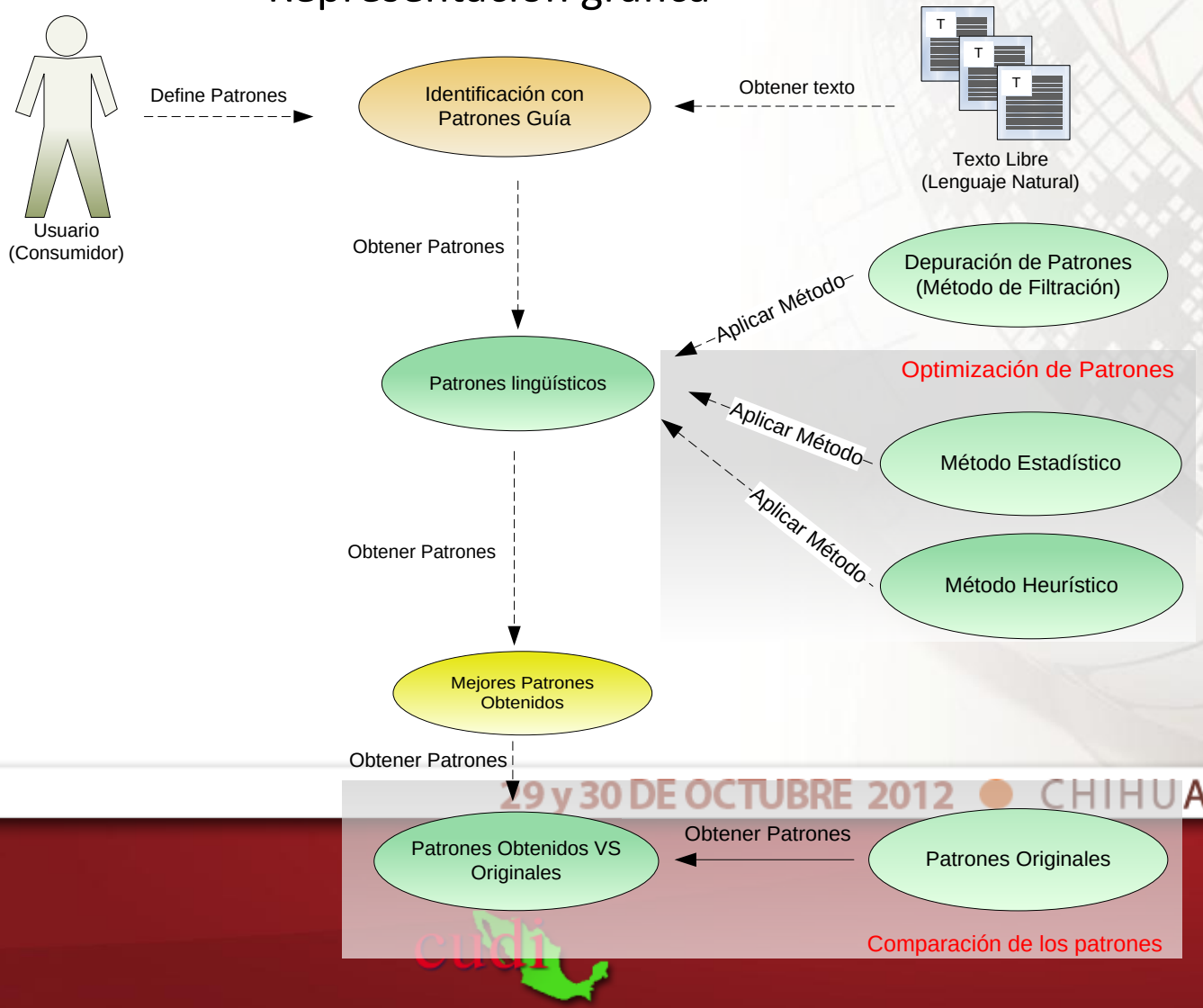
Fase de patrones lingüísticos

- Aprendizaje automático de patrones
  - Definir la longitud máxima del patrón
  - Definir la especificación de cada patrón
  - Creación de una lista de patrones guía
  - Identificación de patrones en el corpus
  - Optimización de patrones

- Proporcionados por el usuario
  - Añadirlos al sistema

Aprendizaje automático de patrones

Representación gráfica



Patrones lingüísticos obtenidos

(Ejemplo)

Longitud del Patrón	*Patrones Guía*	*Patrones Obtenidos*	*Términos*
1	xx	NC AQ	*secretario* *académico*
2	xx·xx	NC + AQ	*contrato colectivo*
3	xx·xx·xx	NC + AQ + AQ NC + SP + NP	*secretario general académico* *programa de doctorado*
4	xx·xx·xx·x	NC + SP + NC + A	*institución de educación superior*
6	xx·x·xx·xx·xx·xx	NC+A+SP+NC+SP+NC	*título profesional a nivel de licenciatura*

Módulo - Extracción de conceptos

- Fase de PLN
- Fase de patrones lingüísticos
- **Fase de extracción de conceptos compuestos**
- Fase de extracción de conceptos simples

Extracción de conceptos compuestos

- Obtención de términos candidatos basado en patrones
  - Método de Clasificación C-Value [Frantzi et al., 2000]

$$C-Value = \begin{cases} \log_2 |a| * f(a) \\ \log_2 |a| * \left(f(a) - \frac{1}{P(T_a)} \sum_{b \in T_a} f(b) \right) \end{cases}$$

  - Donde:
    - a es el término candidato.
    - $|a|$ es la longitud del término candidato.
    - $f(a)$ es la frecuencia del término candidato a en el corpus.
    - T_a es el conjunto de candidatos de mayor longitud que contienen a a.
    - $P(T_a)$ es el número de los candidatos de mayor longitud que contienen a a.
    - $\sum f(b)$ es la ocurrencia total de a como subtérmino de cualquier término candidato b tal que $|a| < |b|$.

Extracción de conceptos compuestos

- Factor de ponderación de contexto
  - Método de clasificación NC-Value [Frantzi et al., 2000]

$$weight(w) = \frac{t(w)}{n}$$

  - Donde:
    - w es la palabra de contexto.
    - t(w) es el número de veces que aparece la palabra de contexto con el término.
    - *n* es el número total de veces que se considero.
    - *weight (w)* es el factor de ponderación de contexto.

Extracción de conceptos compuestos

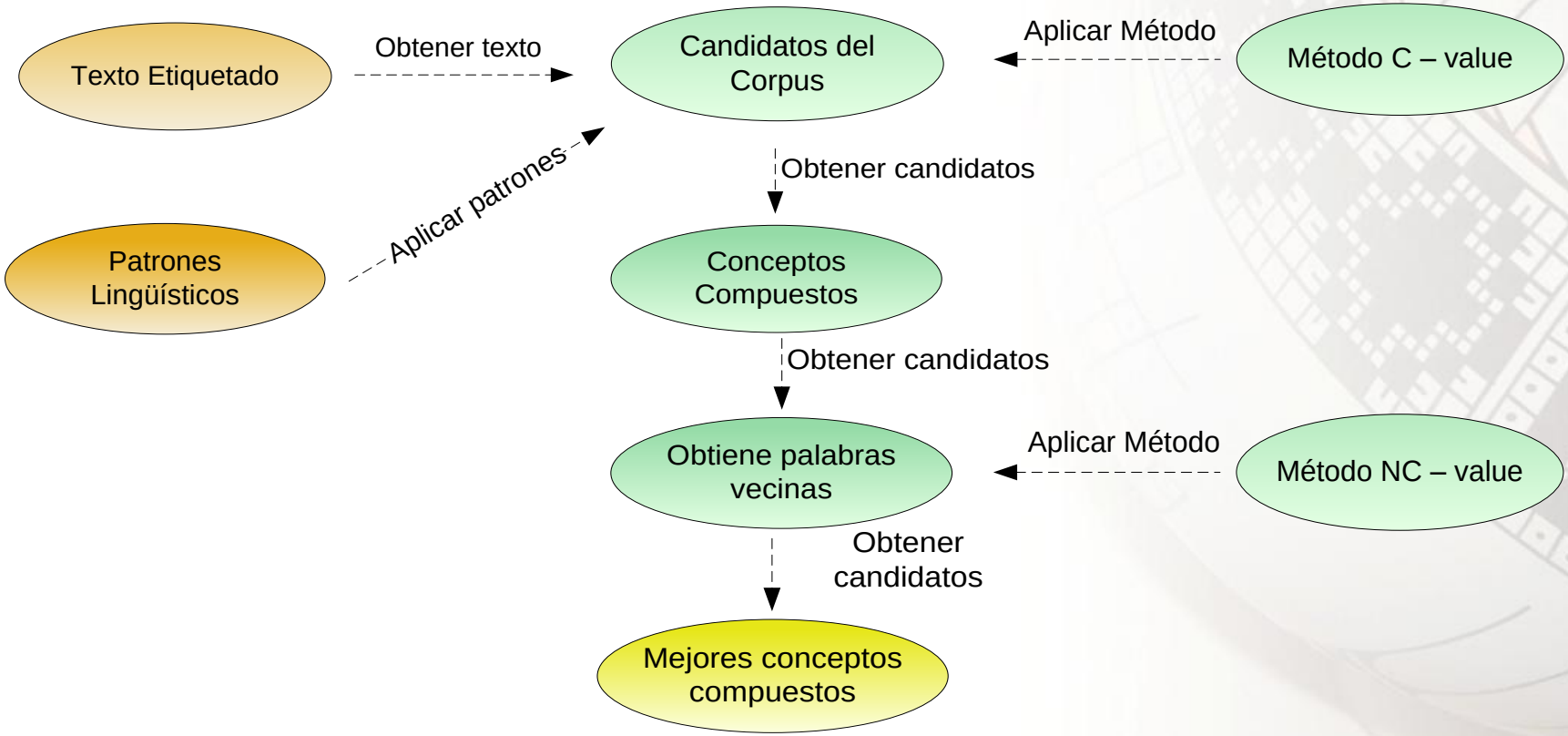
- Obtención de términos candidatos basado en patrones

$$NC-value(a) = 0.8C-value(a) + 0.2\sum_{b \in C_a} f_a(b)weight(b)$$

  - Donde:
    - a es el término candidato.
    - C_a es el conjunto de palabras de contexto de a.
    - b es una palabra de C_a.
    - f(a) es la frecuencia de *b* como palabra de contexto de *a*.
    - weight (b) es el peso de *b* como palabra de contexto.

Extracción de conceptos compuestos

Representación gráfica



Módulo - Extracción de conceptos

- Fase de PLN
- Fase de patrones lingüísticos
- Fase de extracción de conceptos compuestos
- **Fase de extracción de conceptos simples**

Extracción de conceptos simples

- Obtención de términos candidatos basado en patrones
  - Método TF-IDF de Clasificación [(Salton, 1991); (Knoth et al., 2009)]

$$(tf - idf)_{i,j} = tf_{i,j} \times idf_i$$

Donde:

- $tf_{i,j}$ *representa la frecuencia del término*
- idf_i *que representa la frecuencia del documento inversa*

Extracción de conceptos simples

- Método TF- IDF de clasificación

$$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}}$$

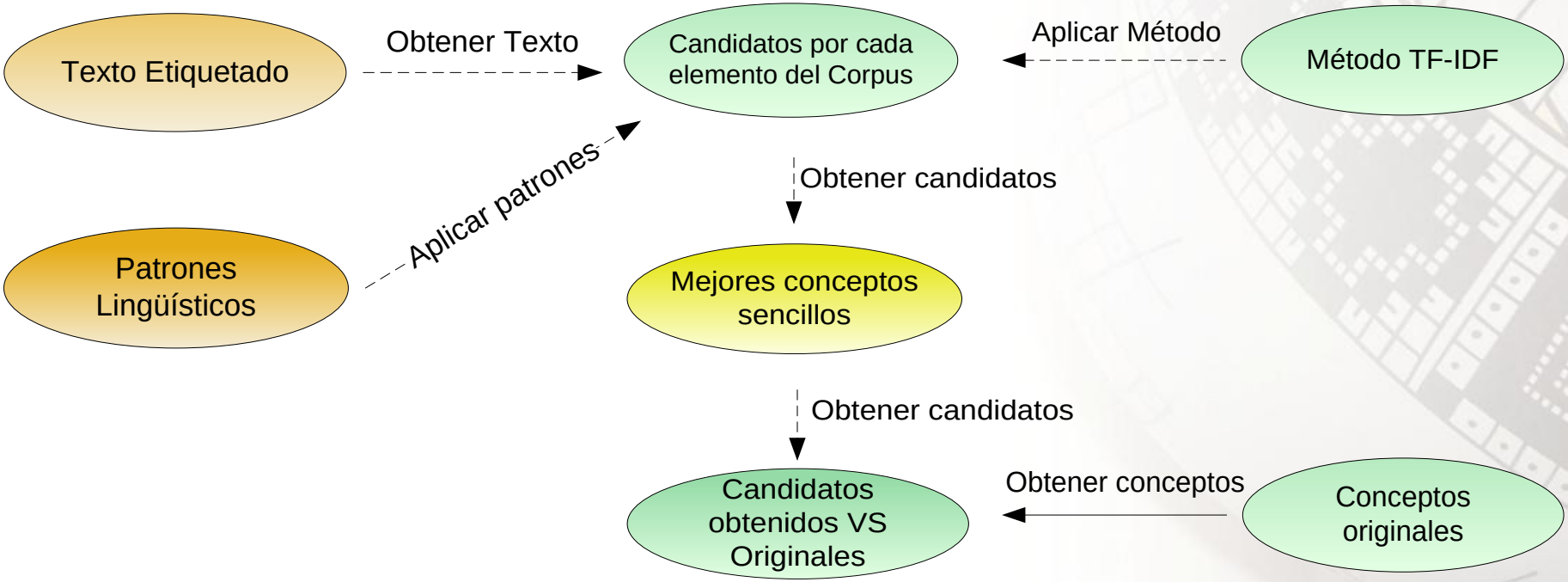
- Donde:
- *El **numerador** $n_{i,j}$: es el número de ocurrencias del término considerado (ti) en el documento dj.*
- *El **denominador** es la suma del número de ocurrencias de los términos en el documento d_j, es decir, el tamaño del documento $|d_j|$.*

$$idf_i = \log \frac{|D|}{|\{d : t_i \in d\}|}$$

- Donde:
- **|D|** : es el número total de documentos en el corpus.
- $|\{d : t_i \in d\}|$ es el número de documentos en los que aparece el término t_i.
- En el supuesto de que el término no este en el corpus, se producirá una división por cero. Por lo tanto es común utilizar $1 + |\{d : t_i \in d\}|$.

Extracción de conceptos simples

Representación gráfica



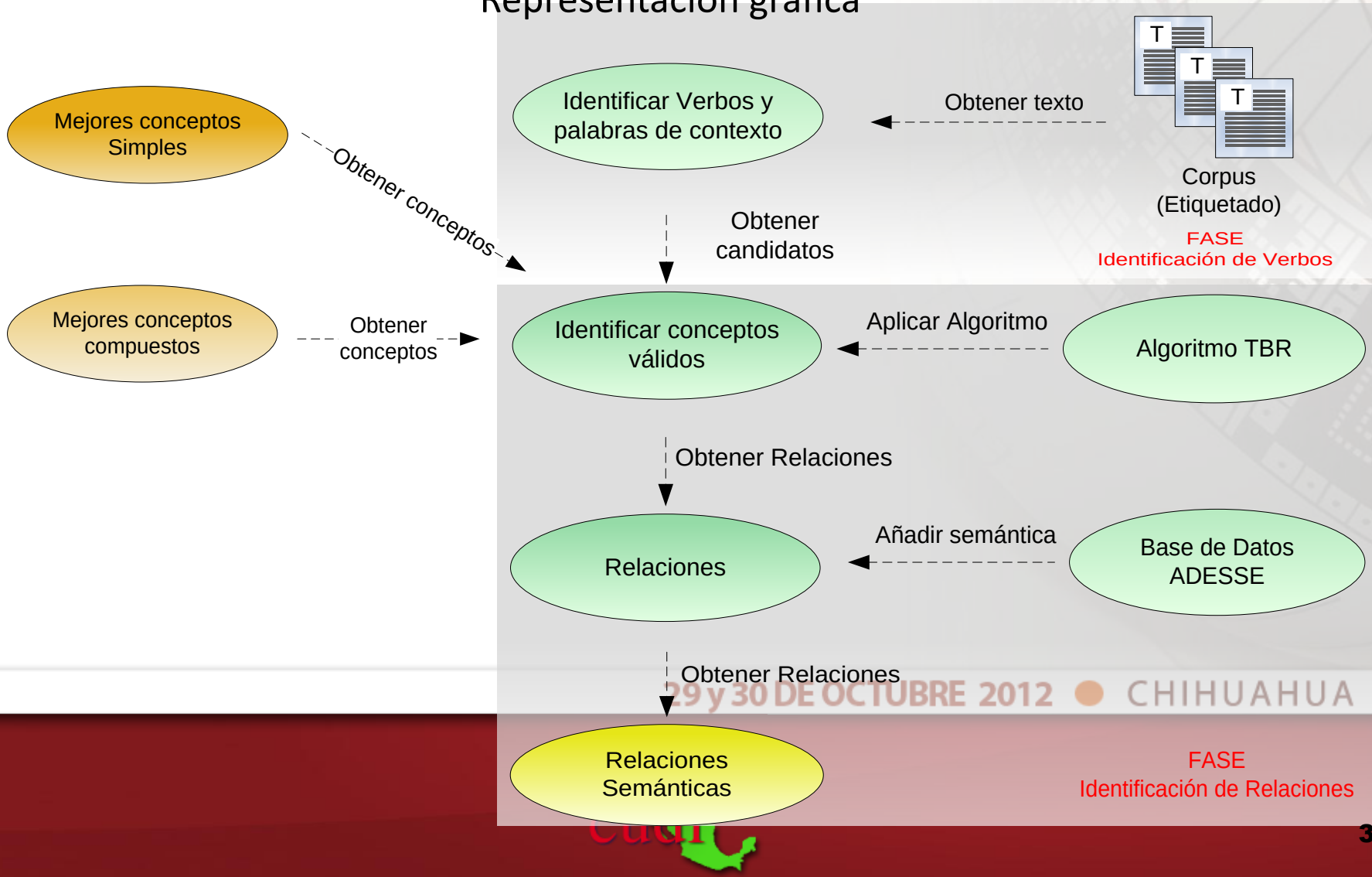
Ejemplo de los conceptos extraídos

NC-value	*Conceptos*	*TF-IDF*	*Conceptos*
60.23	*dirección de servicio escolar*	48.58	*Docencia*
60.00	*plan de estudio*	48.57	*Programa*
57.06	*personal académico*	45.99	*Asignatura*
55.47	*jefe de departamento*	44.75	*Nivel*
50.00	*institución de educación superior*	40.68	*Conclusión*
44.38	*programa de doctorado*	35.87	*Posgrado*

Módulo - Extracción de relaciones

Módulo - Extracción de relaciones

Representación gráfica



Identificación y extracción de relaciones

- Identificación de verbos

Oración	*Relaciones*
...y difundir las convocatorias de movilidad estudiantil proporcionadas por la dirección de movilidad, intercambio y...	*movilidad estudiantil* ***proporcionar*** *dirección de movilidad*

- Búsqueda de conceptos alrededor del Verbo

CONCEPTOS IDENTIFICADOS IZQUIERDA	*VERBO*	*CONCEPTOS IDENTIFICADOS DERECHA*
movilidad estudiantil	***proporcionar***	*Dirección*
convocatorias		*Movilidad*
movilidad		*Intercambio*
estudiantil		*dirección de movilidad*

Asignación de roles semánticos

- Roles Semánticos = significado no implícito

```
<predicate lemma="sustituir">
<roleset id=”sustitución”>
<roles>
  <role n="0" descr="holder"/>
  <role n="1" descr="sustituir"/>
</roles>
</roleset>
```

El **[$_0$rector]** es sustituido por el **[$_1$Secretario General Académico]**.

Relaciones Taxonómicas

- Para identificar las relaciones taxonómicas, se busca la unión del verbo “ser” en tercera persona “es” con un determinante “un ó una”,

Oración	Relación taxonómica	Explicación
“*La Junta Directiva es un órgano colegiado con facultades de nombrar al Rector.*”	junta directiva **es un** órgano colegiado	Nos dice que la junta directiva perteneces a la categoría de órganos colegiados.
La Universidad de Sonora es una institución autónoma de servicio público.	Universidad de Sonora **es una** Institución autónoma	Nos dice que a la Universidad de sonora se le considera una institución independiente.

Relaciones Partonómicas

Oración	Relación	Relación semántica	Explicación
"Nos fue turnada para su estudio y dictamen la Iniciativa de Ley Orgánica de la Universidad de Sonora."	Ley Orgánica **de la** Universidad de Sonora	Universidad de Sonora **tiene una** Ley Orgánica.	De la cual obtenemos la relación semántica asociada a la conjunción "de la", que puede ser "tener" o "tiene un".
La obligatoriedad del Estado de sostener el "Fondo Universidad de Sonora", administrado por el Instituto de Crédito Educativo del Estado de Sonora…."	a) obligatoriedad **del** Estado = obligatoriedad **de el** Estado b) Instituto de crédito educativo **del** Estado de Sonora = Instituto de crédito educativo **de el** Estado de Sonora	a) el estado **tiene una** obligatoriedad, es decir, obligaciones b) el estado de Sonora **tiene un** Instituto de Crédito Educativo	a) De la cual obtenemos la relación semántica asociada a la conjunción "de el", que puede ser "tener" o "tiene". b) De la cual obtenemos la relación semántica asociada a la conjunción "de el", que puede ser "tener" o "tiene un".

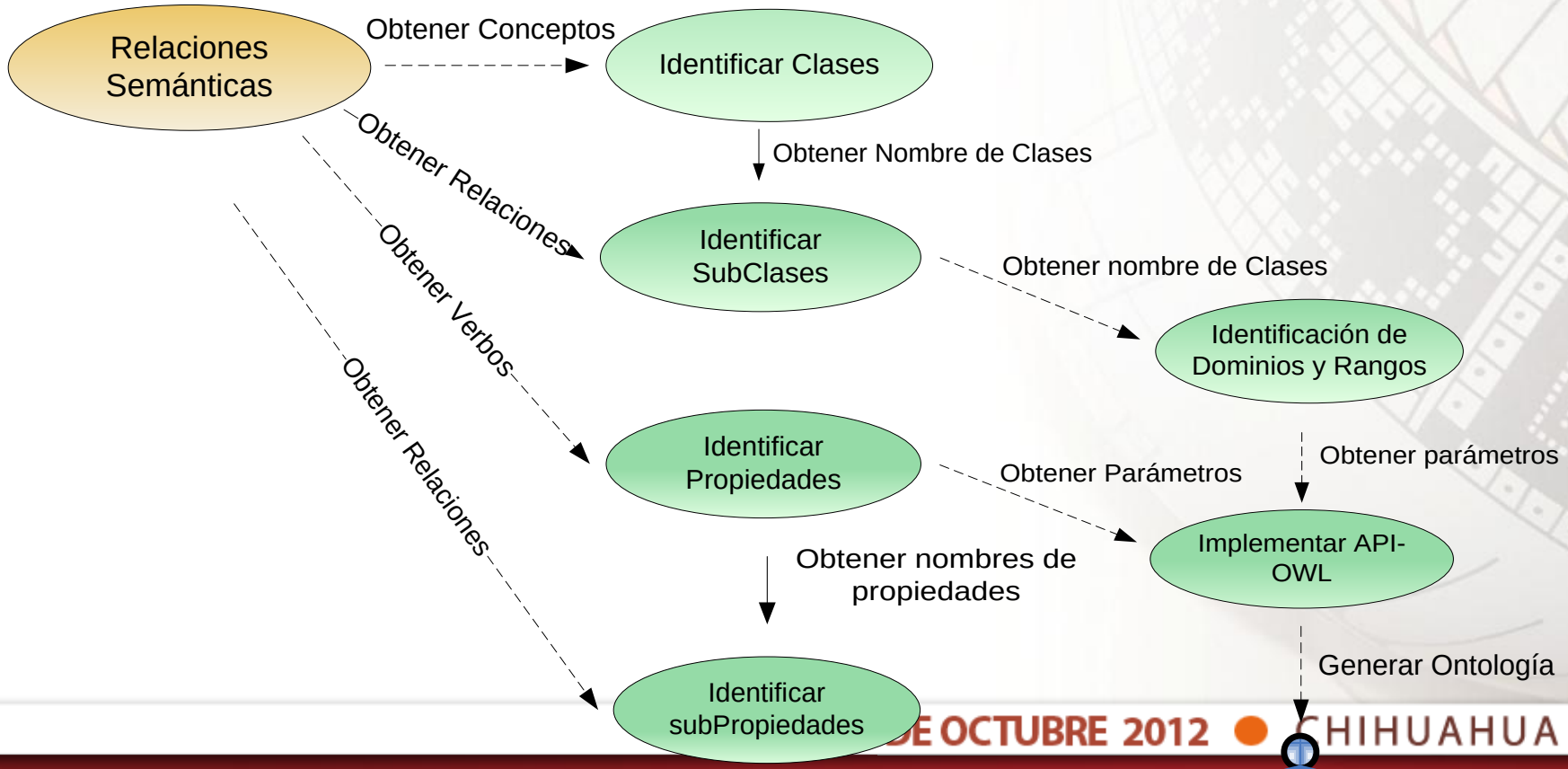
Módulo - Construcción de la ontología

Módulo - Construcción de la ontología

- **Fase de identificación de nombres de clases y subclases**
  - Fase de identificación de nombres de propiedades y subpropiedades
  - Fase de identificación de algunas restricciones

Módulo - Construcción de la ontología

Representación gráfica



Módulo - Construcción de la ontología

- Identificación de nombres de clases y subclases

Conceptos compuestos

- La categoría gramatical nombre = nombre de clase
- Las subclases son los conceptos

Conceptos simples

- Las clases son los conceptos

Módulo - Construcción de la ontología

- Identificación de propiedades y subpropiedades

Para conceptos compuestos y simples

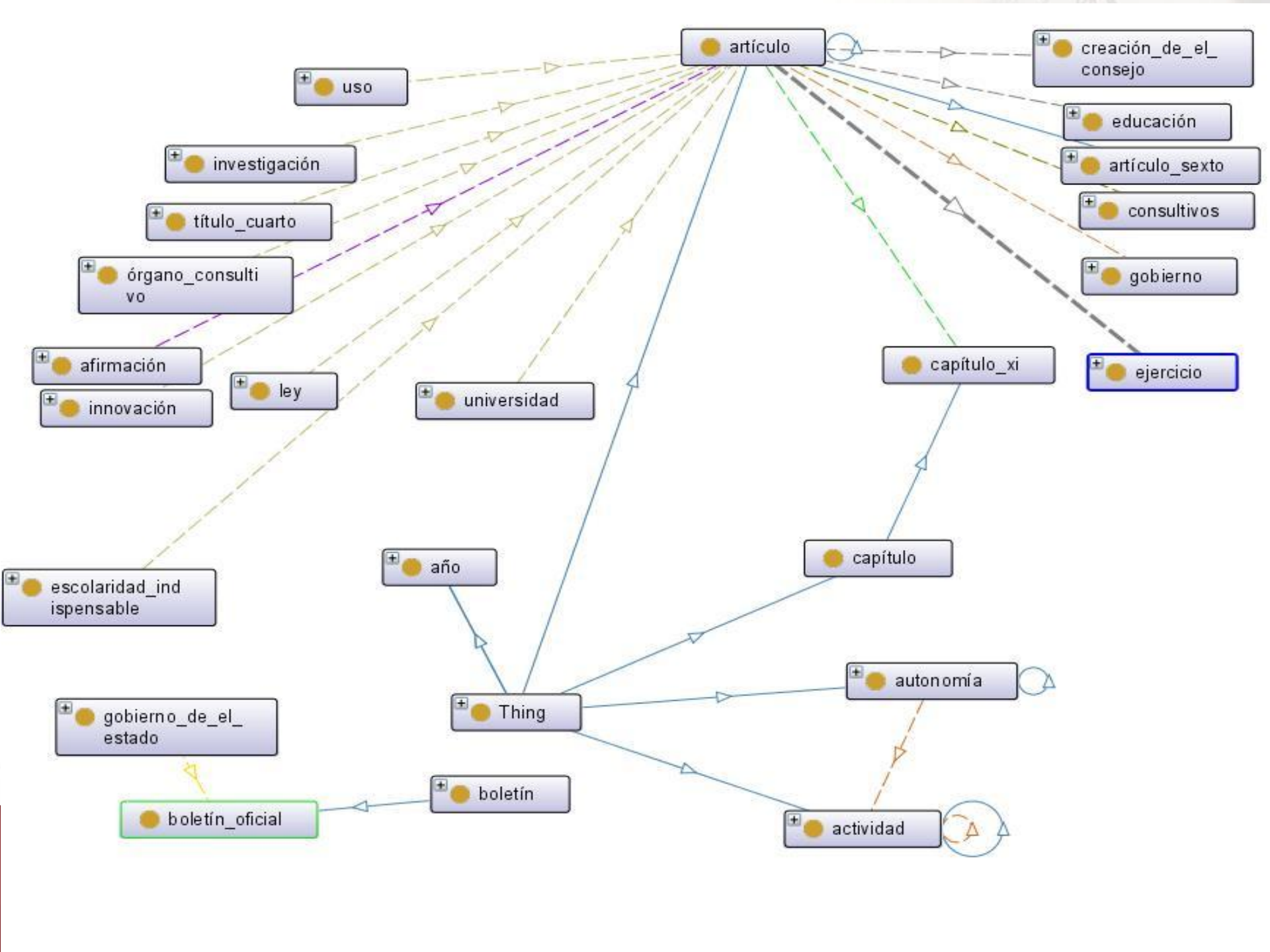
  - Las propiedades son obtenidas a partir del rol semántico.
  - Las subpropiedades son el verbo original en su forma lematizada.

Módulo - Construcción de la ontología

- Identificación de algunas restricciones

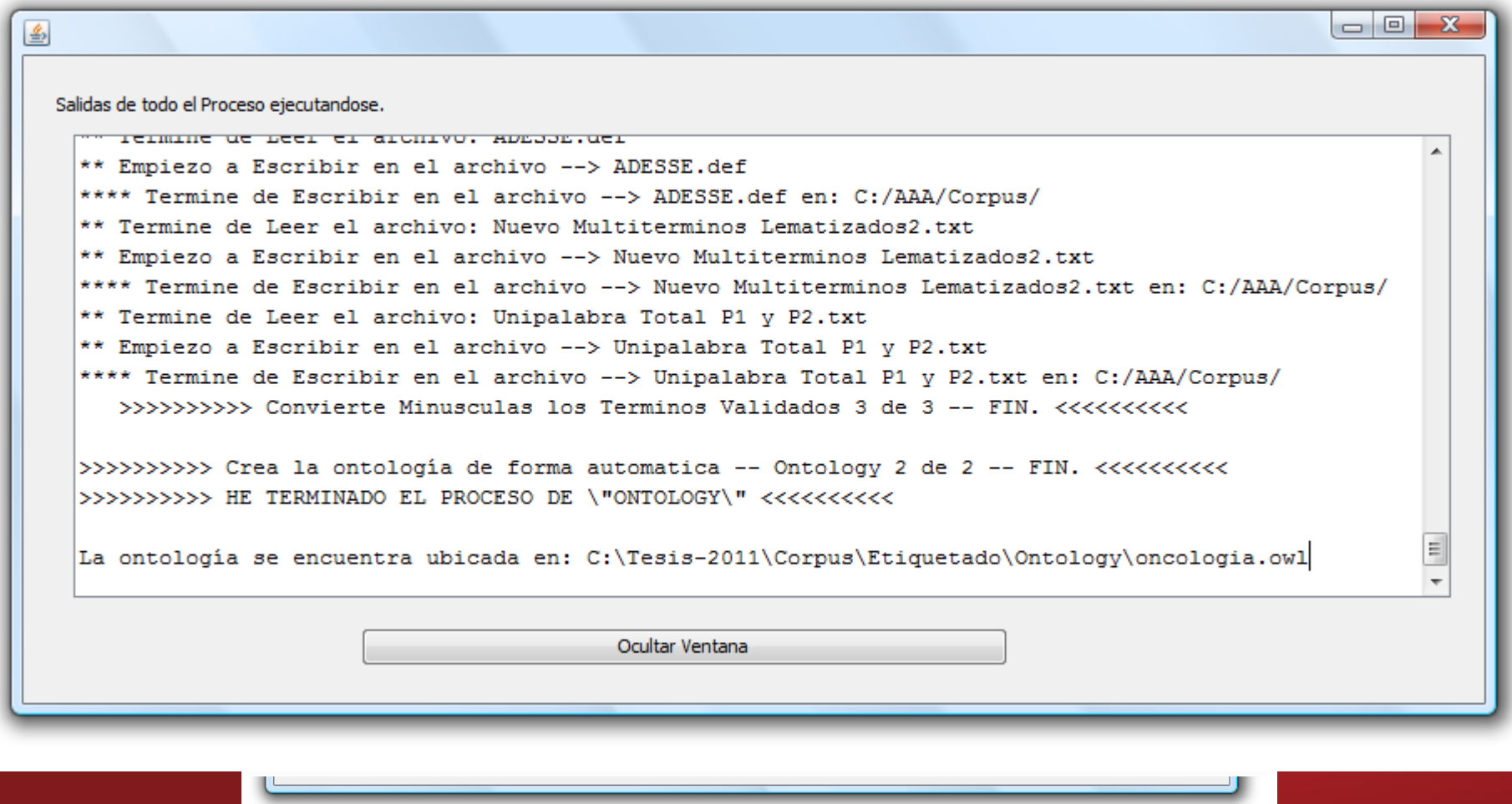
Dominio y rango

- Las clases a la izquierda de cada relación determina el dominio.
- Las clases a la derecha de cada relación determina el rango.


artículo
creación_de_el_consejo
uso
educación
investigación
artículo_sexto
título_cuarto
consultivos
órgano_consultivo
gobierno
afirmación
capítulo_xi
ejercicio
innovación
ley
universidad
escolaridad_indispensable
año
capítulo
autonomía
Thing
gobierno_de_el_estado
boletín
boletín_oficial
actividad

Herramienta de Software

Herramienta de SW



Resultados

- Los resultados han sido cuantificados y valorados de acuerdo a las siguientes métricas de evaluación.

  - Precisión:

$$precisión = \frac{|entidades\ \ extraídas| \cap |entidades\ \ relevantes|}{|entidades\ \ extraídas|}$$

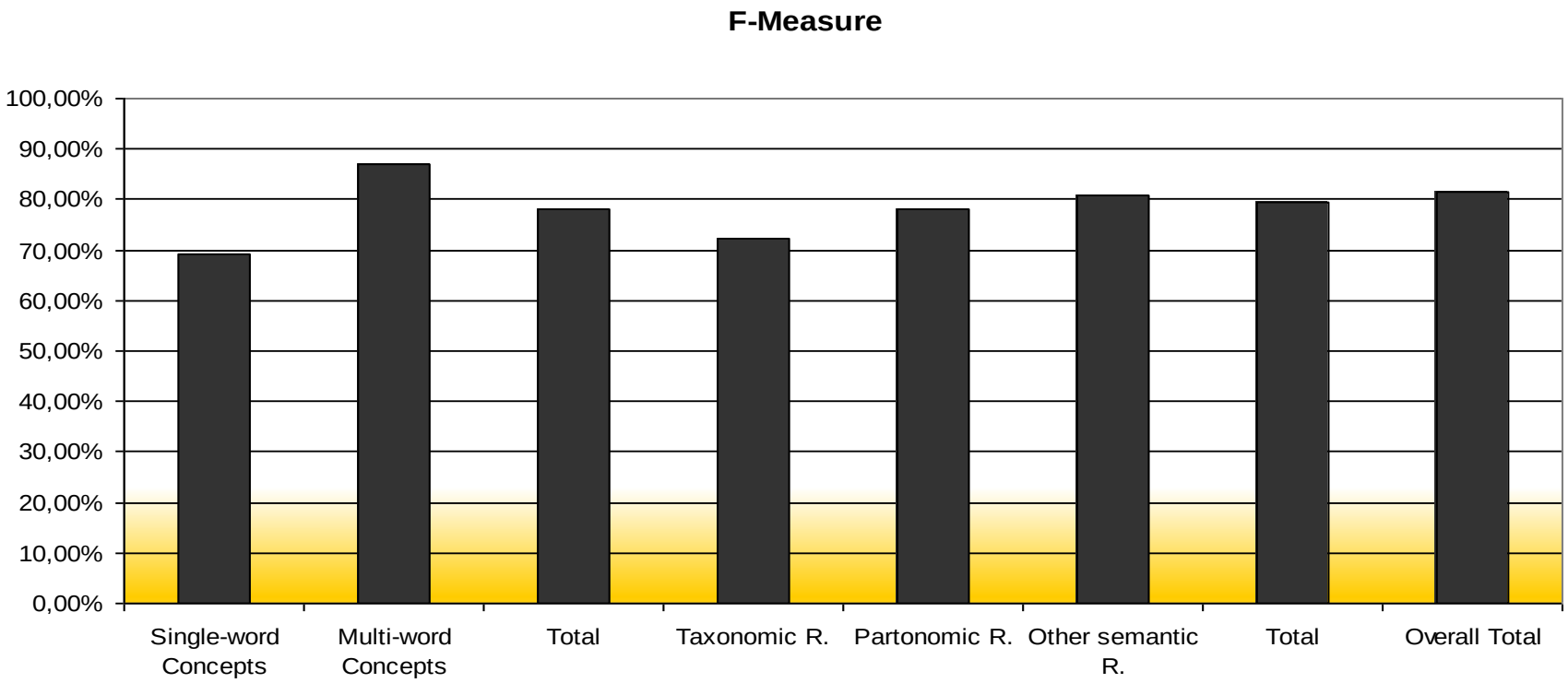
  - Recall:

$$recall = \frac{|entidades\ \ extraídas| \cap |entidades\ \ relevantes|}{|entidades\ \ relevantes|}$$

  - Medida F:

$$F - measure = 2 * \frac{precision * recall}{precision + recall}$$

Resultados obtenidos en el dominio universitario



Muchas Gracias

¿Preguntas?

Mail de contacto: joseluis.ochoa@industrial.uson.mx
Departamento de Ingeniería Industrial,
Universidad de Sonora
Hermosillo, Sonora, México.

Se agradece al gobierno mexicano
y al programa de consolidación Institucional
de repatriación de conacyt (168341)