

Minería semántica

Jose Aguilar
CEMISID, Escuela de Sistemas
Facultad de Ingeniería
Universidad de Los Andes
Mérida, Venezuela

Algunas clases de Minería en Datos

- Minería de Datos Espaciales
- Minería espacio-temporal y Objetivos en movimiento
- Minería Cyber-físico de datos del sistema: salud, control de tráfico aéreo, simulación de inundaciones
- Minería de datos multimedia
- Minería de texto
- Minería de la Web
- Minería de streams de datos



Minería de Secuencia de Datos

- Buscar Similitud en serie temporal de datos
- Regresión y Análisis de Tendencias en series temporales de datos
- Minería Patrón Secuencial en secuencias simbólicas
- Clasificación de Secuencia
- Alineación de secuencias biológicas

Minería Semántica



Determinar relaciones semánticas

- Minería de Datos Semánticos (*Semantic Data Mining*)
- Minería de la Web Semántica (*Semantic Web Mining*)
- Minería Ontológica (*Ontology Mining*).
- Minería de Texto

Análisis de las redes de aprendizaje
Análisis del contenido de aprendizaje
Análisis del contexto de aprendizaje

Minería Semántica

La Minería de Datos es un área bastante madura en las Ciencias Computacionales, cuyo **principal objetivo es la extracción de conocimiento.**

La Minería de Datos ha requerido ser enriquecido estos últimos años, debido a la necesidad de incorporar **contenido semántico.**

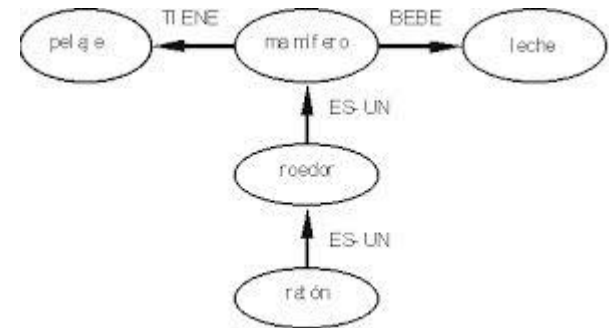


- Ese proceso para determinar relaciones semánticas es conocido como Minería Semántica,**

Minería Semántica



	A	B	C	D	E
1	NOMBRES	CARGO	TELEFONOS	LOCALIDAD	SUELDO
2	Daniela Cárdenas	Chef	3168294789-2574986	ENGATIVA	\$ 1.700.000
3	Gabriela Reyes	Subchef	327459836-4354822	SAN CRISTOBAL	\$ 110.000
4	Carmen Vanegas	Enologo	3154689857-2157458	KENEDDY	\$ 950.000
5	Cristina Porras	Chef Pastelera	3146874953-6874235	BOSA	\$ 150.000
6	Liliana Cruz	Chef Panadera	3201478951-7451825	SUBA	\$ 1.500.000
7	Paola Cristancho	Soucier	3157489614-4785126	CHAPINERO	\$ 800.000
8	Camila Davalos	Cajera	3214675961-7584621	TEUSQUILLO	\$ 700.000
9	Lina Bohorquez	Mesera	3012574816-2245783	CANDELARIA	\$ 600.000
10	Pamela Carrasco	Mesera	3157485912-2485796	CANDELARIA	\$ 600.000
11	Lorena Valencia	Mesera	3204578963-2487512	ENGATIVA	\$ 600.000
12	Jairo Arevalo	Parquedero	3002157459-2861459	BOSA	\$ 489.500
13				TOTAL	\$ 8.199.500

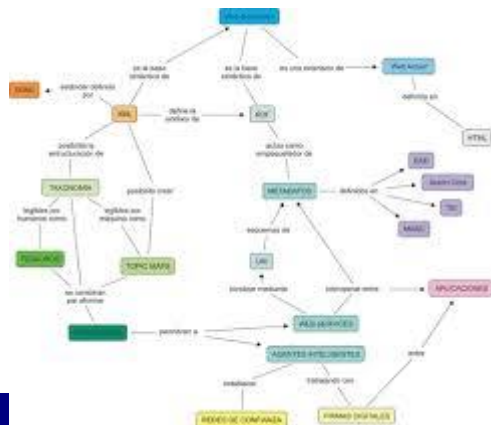
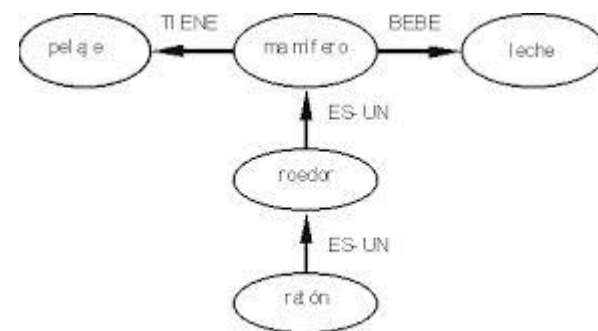


Minería Semántica



	A	B	C	D	E
1	NOMBRES	CARGO	TELEFONOS	LOCALIDAD	SUELDO
2	Daniela Cárdenas	Chief	3168294789-2574986	ENGATIVA	\$ 1.700.000
3	Gabriela Reyes	Subchief	327459836-4354822	SAN CRISTOBAL	\$ 110.000
4	Carmen Vanegas	Enologo	3154689857-2157458	KENNEDY	\$ 950.000
5	Cristina Porras	Chief Pastelera	3146874953-6874225	BOSA	\$ 150.000
6	Liliana Cruz	Chief Panadera	3201478951-7451825	SUBA	\$ 1.500.000
7	Paola Cristancho	Soucier	3157489614-4785126	CHAPINERO	\$ 800.000
8	Camila Davalos	Cajera	3214875961-7584621	TEUSQUILLO	\$ 700.000
9	Lina Bohorquez	Mesera	3012574816-2245783	CANDELARIA	\$ 600.000
10	Pamela Carrasco	Mesera	3157485912-2485796	CANDELARIA	\$ 600.000
11	Lorena Valencia	Mesera	3204578963-2487512	ENGATIVA	\$ 600.000
12	Jairo Arevalo	Parquero	3002157459-2861459	BOSA	\$ 489.500
13				TOTAL	\$ 8.199.500

Minería



Minería Semántica



La minería semántica se encarga de extraer conocimiento semántico desde **diferentes fuentes semánticas**,

- Páginas web,
- **Contenido sin estructura en la web,**
- Contenido estructurado en la web,
- **Grafos anotados,**
- Ontologías,
- **Tabla de Datos, entre otros**

Minería de Texto

DM



Dato estructurado

HomeLoan (
 Loatee: Frank Rizzo
 Lender: MWF
 Agency: Lake View
 Amount: \$200,000
 Term: 15 years
)

Multimedia



texto libre

Frank Rizzo bought
his home from Lake
View Real Estate in
1992.
He paid \$200,000
under a 15-year loan
from MW Financial.

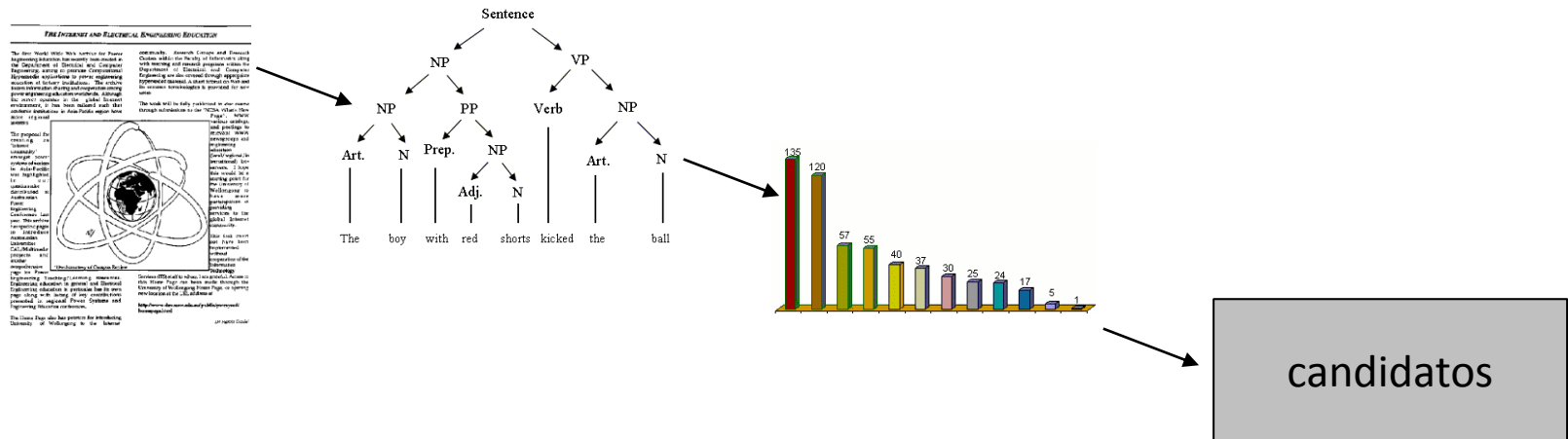
Hypertexto

[Frank Rizzo](#)
Bought
[this home](#)
from [Lake](#)
View Real Estate
In **1992**.
...

Minería de Textos

Consiste en **generar conocimiento nuevo**, a partir de grandes cantidades de **texto**, el cual no está literalmente escrito en los documentos

- Desarrollo y explotación de corpus lingüísticos.
- Reconocimiento de patrones lingüísticos.
- Caracterización de recursos lingüístico estadísticamente.
- ...



Etapas de la minería de texto

1. **Selección de documentos:** implica la **identificación y recuperación** de los documentos potencialmente relevantes de un conjunto grande (por ejemplo, Internet).
2. **Pre-tratamiento documento:** incluya la **limpieza y la preparación** de los documentos, por ejemplo, eliminación de información extraña, corrección de errores, la normalización ortográfica, tokenización, etiquetado, etc.
3. **Procesamiento de documentos:** consiste principalmente en la **extracción de información/conocimiento**. En la web semántica se basa en la extracción de metadatos

Técnicas de la minería de texto

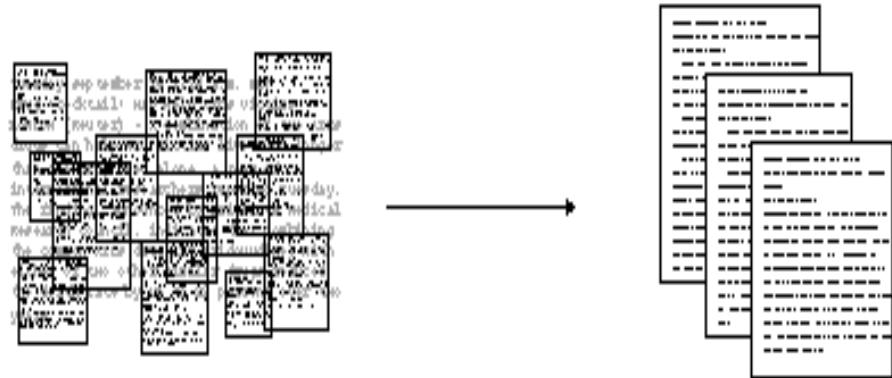
1. Selección y filtrado de documentos
técnicas de Recuperación de Información (RI)
2. Pre-procesamiento de documentos
técnicas de PLN
3. Procesamiento de documentos
técnicas PNL/estadísticas/IA

Extracción de metadatos

- **Extracción de metadatos explícita:** involucra información que describe el documento, como la información que contiene la cabecera de los documentos HTML (títulos, resúmenes, autores, fecha de creación, etc.), XML, RDF, etc.
- **Extracción de metadatos implícita:** implica información semántica que se deduce, es decir, información endógena como los nombres de las entidades y las relaciones contenidas en el texto.
 - Esto implica esencialmente las técnicas de extracción de información, a menudo con la ayuda de una ontología.
 - Linked data es de gran ayuda

Extracción de Inform. (EI) no es Recuperación de Inform. (RI)

RI busca documentos a partir de grandes colecciones de textos (por lo general, la Web) en respuesta a palabras clave o consultas específicas



Se recuperan los documentos

EI saca hechos e información estructurada de los contenidos de las grandes colecciones de textos



Se analizan los hechos

Minería de Texto

Abridged Declaration of Independence

A Declaration By the Representatives of the United States of America, in General Congress Assembled.

When in the course of human events it becomes necessary for a people to advance from that subordination in which they have hitherto remained, and to assume among powers of the earth the equal and independent station to which the laws of nature and of nature's god entitle them, a decent respect to the opinions of mankind requires that they should declare the causes which impel them to the change.

We hold these truths to be self-evident; that all men are created equal and independent; that from that equal creation they derive rights inherent and inalienable, among which are the preservation of life, and liberty, and the pursuit of happiness; that to secure these ends, governments are instituted among men, deriving their just power from the consent of the governed; that whenever any form of government shall become destructive of these ends, it is the right of the people to alter or to abolish it, and to institute new government, laying it's foundation on such principles and organizing it's power in such form, as to them shall seem most likely to effect their safety and happiness. Prudence indeed will dictate that governments long established should not be changed for light and transient causes: and accordingly all experience hath shewn that mankind are more disposed to suffer while evils are sufferable, than to right themselves by abolishing the forms to which they are accustomed. But when a long train of abuses and usurpations, begun at a distinguished period, and pursuing invariably the same object, evinces a design to reduce them to arbitrary power, it is their right, it is their duty, to throw off such government and to provide new guards for future security. Such has been the patient sufferings of the colonies; and such is now the necessity which constrains them to expunge their former systems of government. the history of his present majesty is a history of unremitting injuries and usurpations, among which no one fact stands single or solitary to contradict the uniform tenor of the rest, all of which have in direct object the establishment of an absolute tyranny over these states. To prove this, let facts be submitted to a candid world, for the truth of which we pledge a faith yet unsullied by falsehood.

¿Cuántas palabras grandes, pequeñas y medianas están en el texto?

Histograma de longitud de palabras

- Mucho (amarillo)= 10 + letras
- Medio (rojo)= 5 a 9 letras
- Poco (azul)= 2 a 4 letras
- Morado= 1 letra

Abridged Declaration of Independence

A Declaration By the Representatives of the United States of America, in General Congress Assembled.

When in the course of human events it becomes necessary for a people to advance from that subordination in which they have hitherto remained, and to assume among powers of the earth the equal and independent station to which the laws of nature and of nature's god entitle them, a decent respect to the opinions of mankind requires that they should declare the causes which impel them to the change.

We hold these truths to be self-evident; that all men are created equal and independent; that from that equal creation they derive rights inherent and inalienable, among which are the preservation of life, and liberty, and the pursuit of happiness; that to secure these ends, governments are instituted among men, deriving their just power from the consent of the governed; that whenever any form of government shall become destructive of these ends, it is the right of the people to alter or to abolish it, and to institute new government, laying it's foundation on such principles and organizing it's power in such form, as to them shall seem most likely to effect their safety and happiness. Prudence indeed will

dictate that governments long established should not be changed for light and transient causes: and accordingly all experience hath shewn that mankind are more disposed to suffer while evils are sufferable, than to right themselves by abolishing the forms to which they are accustomed. But when a long train of abuses and usurpations, begun at a distinguished period, and pursuing invariably the same object, evinces a design to reduce them to arbitrary power, it is their right, it is their duty, to throw off such government and to provide new guards for future security. Such has been the patient sufferings of the colonies; and such is now the necessity which constrains them to expunge their former systems of government. the history of his present majesty is a history of unremitting injuries and usurpations, among which no one fact stands single or solitary to contradict the uniform tenor of the rest, all of which have in direct object the establishment of an absolute tyranny over these states. To prove this, let facts be submitted to a candid world, for the truth of which we pledge a faith yet unsullied by falsehood.

Histograma de longitud de palabras

Mapa 1
204 palabras

Abridged Declaration of Independence

A Declaration By the Representatives of the United States of America, in General Congress Assembled.

When in the course of human events it becomes necessary for a people to advance from that subordination in which they have hitherto remained, and to assume among powers of the earth the equal and independent station to which the laws of nature and of nature's god entitle them, a decent respect to the opinions of mankind requires that they should declare the causes which impel them to the change.

We hold these truths to be self-evident; that all men are created equal and independent; that from that equal creation they derive rights inherent and inalienable, among which are the preservation of life, and liberty, and the pursuit of happiness; that to secure these ends, governments are instituted among men, deriving their just power from the consent of the governed; that whenever any form of government shall become destructive of these ends, it is the right of the people to alter or to abolish it, and to institute new government, laying it's foundation on such principles and organizing it's power in such form, as to them shall seem most likely to effect their safety and happiness. Prudence indeed will

- Amarillo, 17
- Rojo, 17
- Azul, 107
- Morado, 3

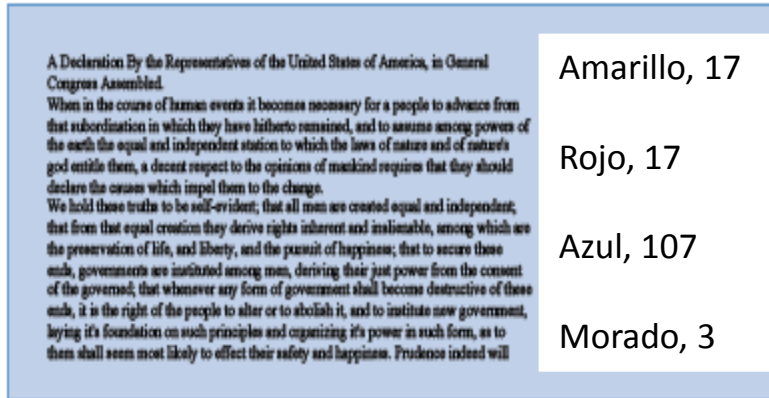
Mapa 2
190 palabras

dictate that governments long established should not be changed for light and transient causes: and accordingly all experience hath shewn that mankind are more disposed to suffer while evils are sufferable, than to right themselves by abolishing the forms to which they are accustomed. But when a long train of abuses and usurpations, begun at a distinguished period, and pursuing invariably the same object, evinces a design to reduce them to arbitrary power, it is their right, it is their duty, to throw off such government and to provide new guards for future security. Such has been the patient sufferings of the colonies; and such is now the necessity which constrains them to expunge their former systems of government. the history of his present majesty is a history of unremitting injuries and usurpations, among which no one fact stands single or solitary to contradict the uniform tenor of the rest, all of which have in direct object the establishment of an absolute tyranny over these states. To prove this, let facts be submitted to a candid world, for the truth of which we pledge a faith yet unsullied by falsehood.

- Amarillo, 20
- Rojo, 71
- Azul, 93
- Morado, 6

Histograma de longitud de palabras

Mapa 1

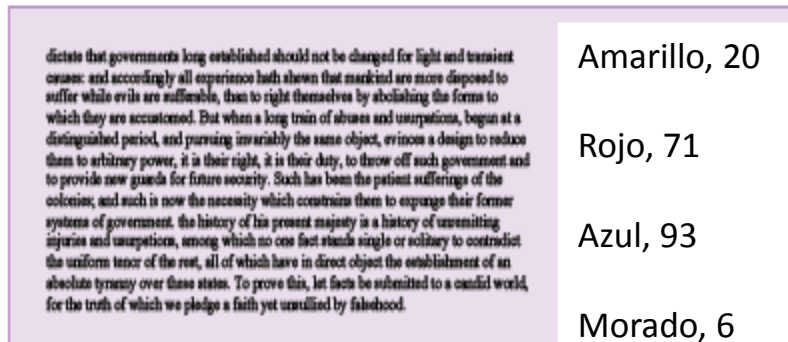


Combinar

• Amarillo, 37

• Rojo, 88

Mapa 2



• Azul, 200

• Morado, 9

Simple análisis redes sociales: Contar amigos

Entrada

Jim,Sue

Sue,Jim

Lin,Joe

Joe,Lin

Jim,Kai

Kai,Jim

Jim,Lin

Lin,Jim

mapeo

Jim,1

Sue,1

Lin,1

Joe,1

Jim,1

Kai,1

Jim,1

Lin,1

combino

Jim,(1,1,1)

Lin,(1,1)

Soe,(1)

Joe,(1)

Kai,(1)

reduzco

Salida

Jim, 3

Lin,2

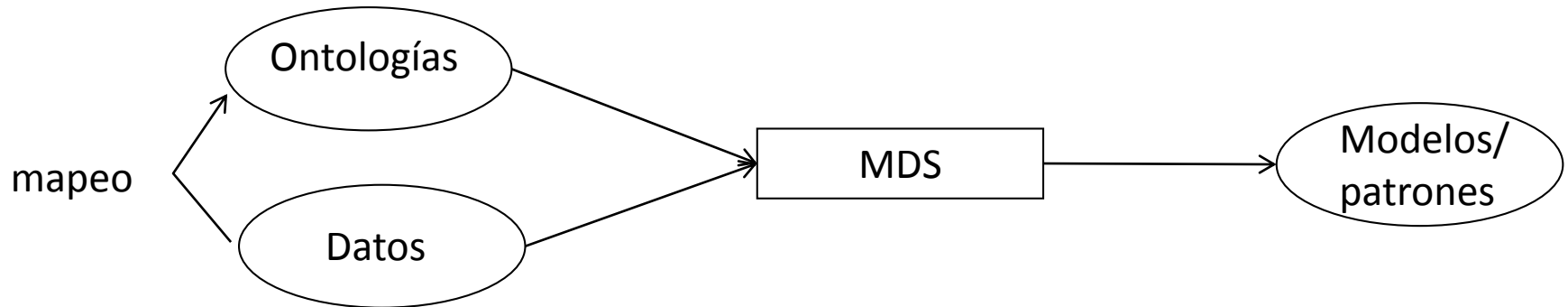
Soe,1

Joe,1

Kai,1

Minería de Datos Semánticos (MDS)

- **Incorporar conocimiento de un dominio** a los datos.
Minar **recursos anotados semánticamente**, como ontologías para enriquecer semánticamente los datos
- **Añadir contenido semántico a/desde los datos usando técnicas de MD** para la extracción de ese conocimiento (en este caso, la fuente es contenido semántico).



Minería de Datos Semánticos (MDS)

- El proceso de MDS se da en dos pasos,
 1. **Identificación del enriquecimiento semántico,**
 2. **Aplicación de técnicas de MD como tal en él.**
- En el primer paso se usan ontologías, o cualquier contenido semántico, y **se realiza un mapeo** con la data que se va a trabajar, almacenada normalmente en bases de datos.
- En el segundo paso se aplican técnicas de MD para **buscar patrones, relaciones**, y en general, cualquier operación que **explote el enriquecimiento semántico**.

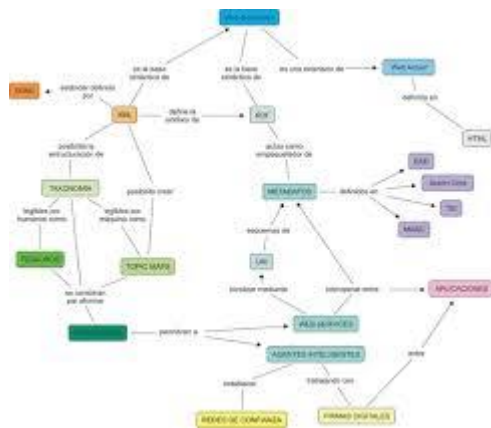
conocimiento

Minería de Datos Semánticos

Dado: tabla de datos de transacciones, bases de datos relacionales, documentos de texto, páginas Web, ... una o más ontologías de dominio, etc.

	A	B	C	D	E
1	NOMBRES	CARGO	TELEFONOS	LOCALIDAD	SUELDO
2	Daniela Cárdenas	Chief	3168294789-2574986	ENGATIVA	\$ 1.700.000
3	Gabriela Reyes	Subchief	327459836-4354822	SAN CRISTOBAL	\$ 110.000
4	Carmen Vanegas	Enologo	3154689857-2157458	KENEDDY	\$ 950.000
5	Cristine Porras	Chief Pastelera	3146874953-0874225	BOSA	\$ 150.000
6	Liliana Cruz	Chief Panadera	3201478951-7451825	SUBA	\$ 1.500.000
7	Paola Cristancho	Soucier	3157489614-4785126	CHAPINERO	\$ 800.000
8	Camila Davalos	Cajera	3214875961-7946421	TEUSQUILLO	\$ 700.000
9	Lina Bohorquez	Mesera	3012574816-2245783	CANDELARIA	\$ 600.000
10	Pamela Carrasco	Mesera	3157485912-2485796	CANDELARIA	\$ 600.000
11	Lorena Valencia	Mesera	3204578963-2487512	ENGATIVA	\$ 600.000
12	Jairo Arevalo	Parquesidero	3002157439-2861459	BOSA	\$ 489.500
13				TOTAL	\$ 8.199.500

Minería



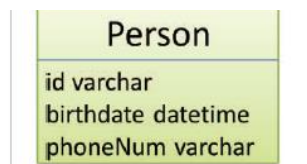
Encontrar: un modelo de clasificación, un conjunto de patrones, etc.



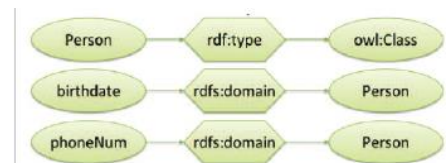
Minería de Datos Semánticos

- **Actual escenario de la MDS:** Minería de datos **empíricos** con ontologías como conocimiento de fondo
 - Abundantes datos empíricos,
 - Escaso conocimiento de fondo
- **Futuro escenario de MDS:**
 - Volumen creciente de ontologías y colecciones de datos semánticamente anotados
 - más de 6 billones de tripletas RDF
 - más de 200 millones de enlaces

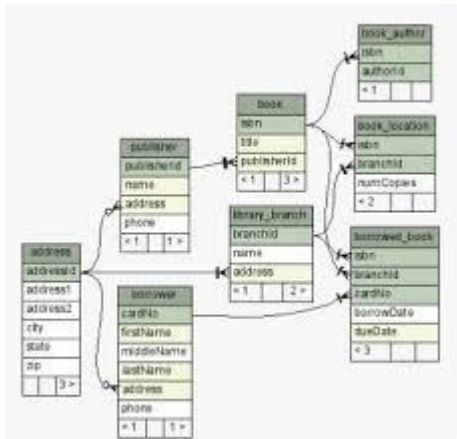
Definición relacional



Ontología



Minería de Datos Semánticos



Paso 1

1. Definir características Relacionales
2. construir una tabla proposicional

	f1	f2	f3	f4	f5	f6	...	...	...	...	fn
g1	1	0	0	1	1	1	0	0	1	0	1
g2	0	1	1	0	1	1	0	0	0	1	0
g3	0	1	1	1	0	0	1	0	0	0	1
g4	1	1	1	0	1	1	0	0	1	1	0
g5	1	1	1	0	0	1	0	1	1	0	1
g1	0	0	1	1	0	0	0	1	0	0	1
g2	1	1	0	0	1	1	0	1	0	1	1
g3	0	0	0	0	1	0	0	1	1	0	0
g4	1	0	1	1	1	0	1	0	0	1	0

	f1	f2	f3	f4	f5	f6	...	...	...	...	fn
g1	1	0	0	1	1	1	0	0	1	0	1
g2	0	1	1	0	1	1	0	0	0	1	0
g3	0	1	1	1	0	0	1	1	0	0	1
g4	1	1	1	0	1	1	0	0	1	1	0
g5	1	1	1	0	0	1	0	1	1	0	1
g1	0	0	1	1	0	0	0	1	0	0	1
g2	1	1	0	0	1	1	0	1	0	1	1
g3	0	0	0	0	1	0	0	1	1	1	0
g4	1	0	1	1	1	0	1	0	0	1	0

Paso 2

Descubrir reglas

```

If sepal width < 2.55 and petal length < 4.95 and
petal width < 1.55 then Iris versicolor
If petal length ≥ 2.45 and petal length < 4.95 and
petal width < 1.55 then Iris versicolor
If sepal length ≥ 6.55 and petal length < 5.05 then Iris versicolor
If sepal width < 2.75 and petal width < 1.65 and
sepal length < 6.05 then Iris versicolor
If sepal length ≥ 5.85 and sepal length < 5.95 and
petal length < 4.85 then Iris versicolor
If petal length ≥ 5.15 then Iris virginica
If petal width ≥ 1.85 then Iris virginica
If sepal width ≥ 1.75 and sepal width < 3.05 then Iris virginica
If petal length ≥ 4.95 and petal width < 1.55 then Iris virginica
    
```

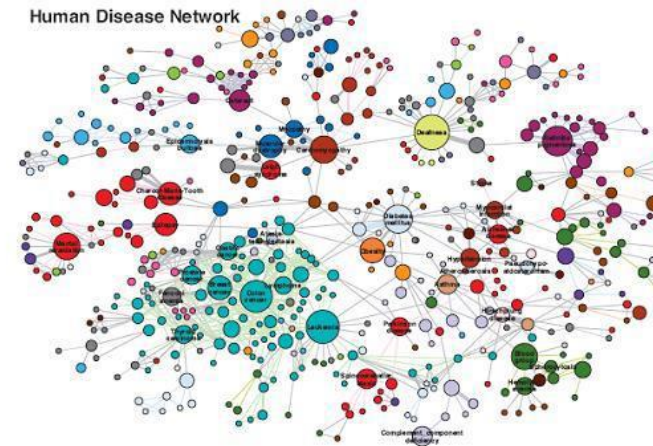
MDS en Clusters de Signaling Pathways Networks

Una red de vías de señales, o signaling pathway, es el conjunto de **reacciones implicadas** en la reacción de una célula a un estímulo externo.

Los clusters no dan mucha información per se, pero al identificar las funciones biológicas que identifican cada cluster se pueden definir **familias**.

La **activación del receptor** provocada por la unión a un ligando se asocia directamente a la respuesta de la célula.

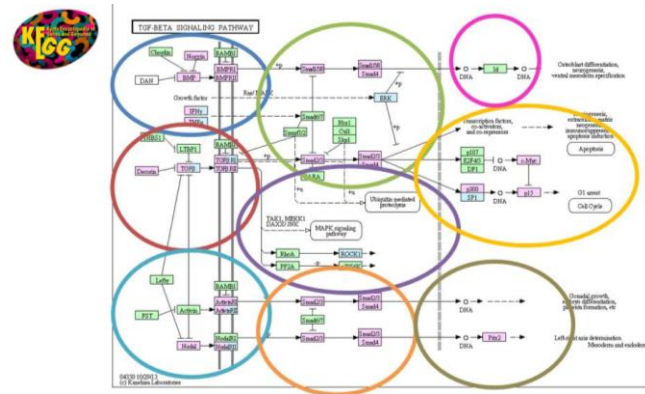
Por ejemplo, el **neurotransmisor GABA** puede activar un receptor de la superficie celular que es parte de un canal iónico.



Macro-Algoritmo SeMiC

Macro algoritmo que permite detectar los clusters dentro de una red signaling pathway, y enriquecerlos con GO.

1. Recibir como entrada una **signaling pathway network**
2. Llevarla a un **formato de red** (las proteínas serán tratados como nodos y las reacciones como relaciones)
3. Calcular la **modularidad** para cada nodo en la red
4. Realizar el **cluster jerárquico**,
5. Calcular **los centroides** de cada cluster, usando técnicas de centralidad de redes.
6. **Enriquecer** cada centroide **semánticamente** con GO



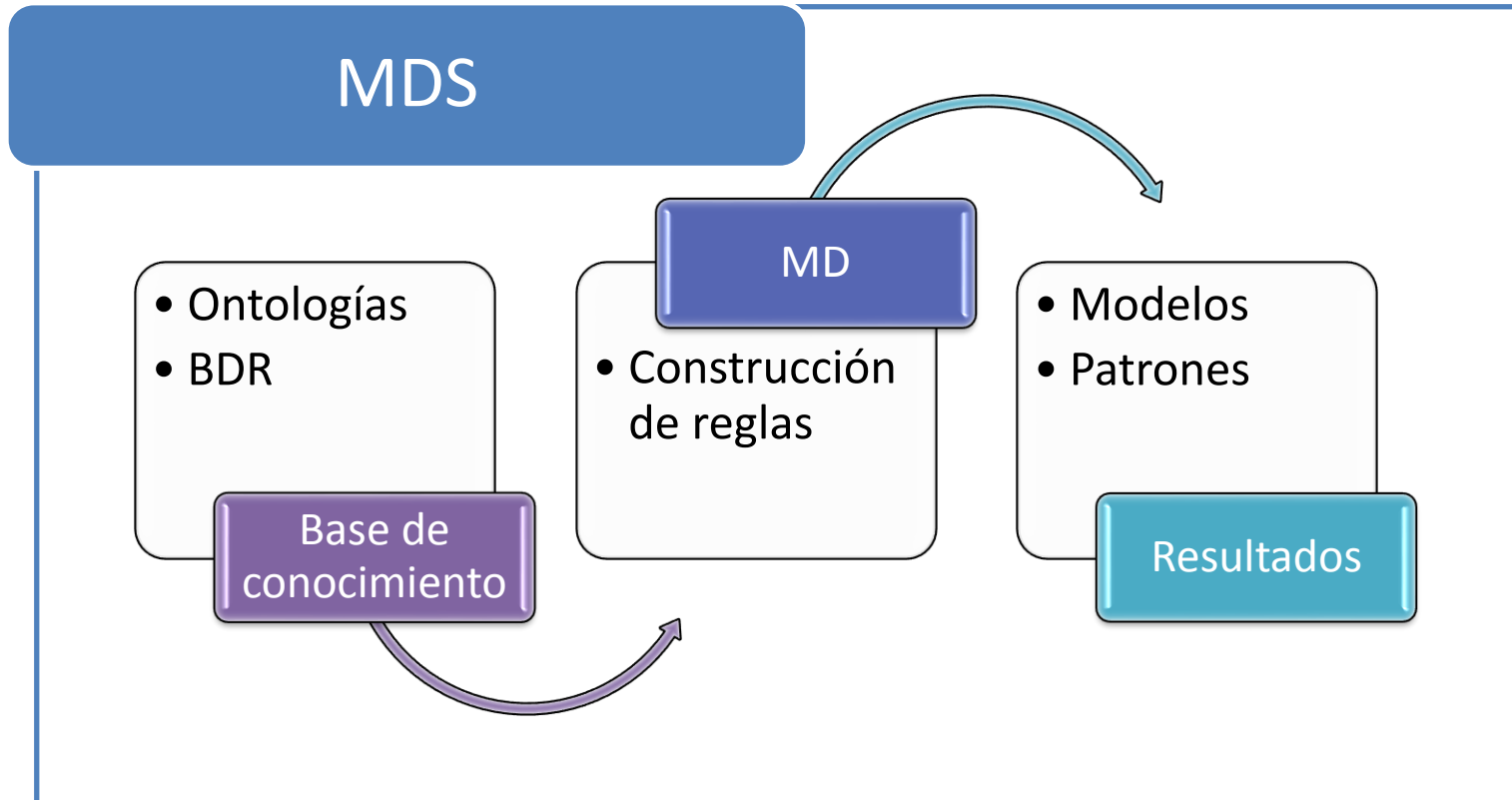
Después se calculan los clústeres **(paso 4)**. A continuación se extraen los centroides **(paso 5)**

Los centroides pasan a un enriquecimiento semántico (**paso 6**) usando Gene Ontology (GO)

herramienta usada: Gephi (para análisis de redes sociales)

Minería de Datos Semánticos

Recapitulando



SDM no solo es para construir reglas, otras formas de conocimiento se pueden generar (grupos, etc.)

Minería de la Web Semántica (MWS)

- Es la integración de dos áreas de conocimiento,
 - **Web Semántica (Semantic Web)**
 - **Minería en la Web (Web Mining)**

La **Web Semántica** es usada para darle significado a los datos que se encuentran en la Web.

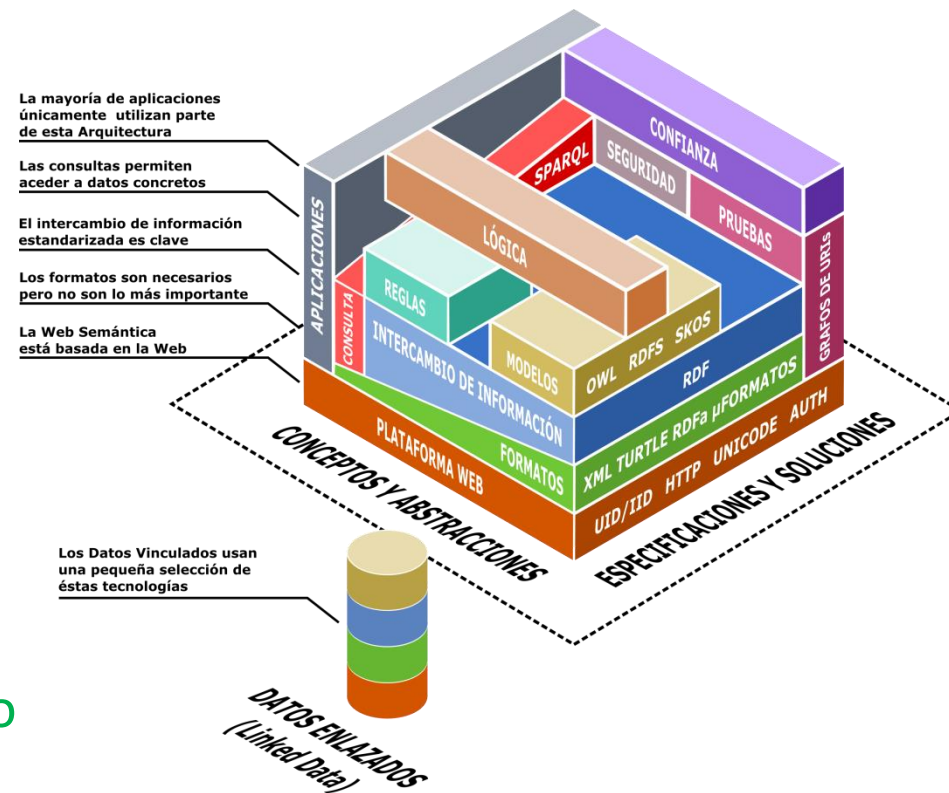
La **Minería en la Web** se usa para extraer patrones de comportamiento en la Web.

Minería de la Web Semántica

La diferencia de MWS con MDS es el propósito y lo que se está minando.

MWS mina datos de la Web, y los resultados son usados en la Web.

- La web semántica es expresada en formatos como OWL, RDF, XML,
- Son los recursos que van a ser minados para extraer conocimiento de la web semántica



Minería Web

Existen varios tipos :

- El contenido de la web,
- La estructura de la web
- El uso que se hace de la web.

Resultados de la Búsqueda

Contenido de la Página Web

Enlaces

Patrones generales de uso

Patrones personales de acceso



- Microformatos
- RDFa
- FOAF
- SEO Semántico

Web mining

Minería del contenido de la Web

Es el **descubrimiento de información** útil desde los **contenidos textuales y gráficos** de los documentos Web, y tiene sus orígenes en el procesamiento del lenguaje natural y en la recuperación de la información.

Minería de la estructura de la Web

Es el proceso de **descubrir el modelo** subyacente a la **estructura de enlaces de la Web** y analiza, fundamentalmente, la **topología de los hipervínculos** (con o sin descripción de los enlaces)

Minería del uso de la Web

Es la aplicación de técnicas de minería de datos para descubrir **patrones de acceso** (o hábitos) **a los sitios Web**.

Web mining

El minado de contenido, es una forma de *Text Mining*, que se aplica al contenido en la Web.

Por ejemplo, identificar en una página términos similares.

El minado de la estructura estudia el **esqueleto que forman los enlaces entre las páginas de la Web**, se mina un conjunto de enlaces.

El minado del uso de la web, se enfoca en minar un **historial de uso de usuarios**

Por ejemplo, consultas que hacen en una página, movimientos que los usuarios hacen entre páginas, etc.

Web mining

Es usual que en el minado del uso, se usen las otras dos técnicas (de estructura y de contenido), haciendo al minado de uso la más completa forma de **Minería Web** y la más usada

Minería Web

Microformatos

XFN (XHTML Friends Network)

XFN quick reference	
relationship category	XFN values
friendship (at most one):	friend acquaintance contact
physical:	met
professional:	co-worker colleague
geographical (at most one):	co-resident neighbor
family (at most one):	child parent sibling spouse kin
romantic:	muse crush date sweetheart
identity:	me

hCard

hcalendar

FOAF

```
<?xml version="1.0" standalone="yes"?>
```

```
<rdf:RDF
```

```
  xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
  xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
  xmlns:foaf="http://xmlns.com/foaf/0.1"/>
```

```
<foaf:Person>
```

```
  <foaf:name>
```

```
    Taniana Josefina Rodríguez de Paredes
```

```
  </foaf:name>
```

```
  <foaf:mbox rdf:resource="mailto:taniana@ula.ve"/>
```

```
  <foaf:knows>
```

```
    <foaf:Person>
```

```
      <foaf:name> Jose Aguilar </foaf:name>
```

```
      <foaf:mbox rdf:resource="mailto:aguilar@ula.ve"/>
```

```
    </foaf:Person>
```

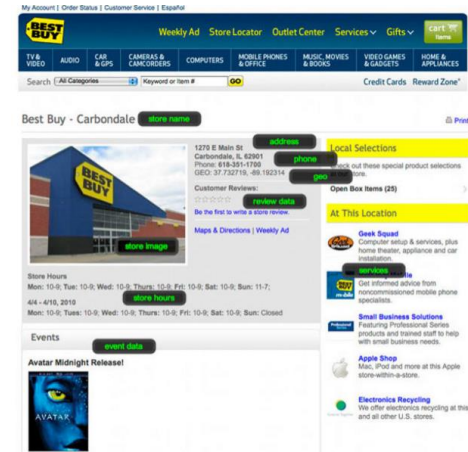
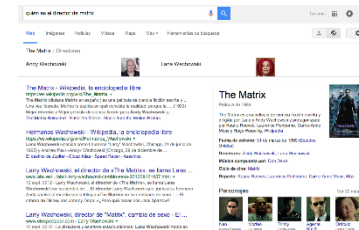
```
  </foaf:Knows>
```

```
</foaf:Person>
```

```
</rdf:RDF>
```

Social Data Analytics
Social Network Analytics
Linked Data

SEO Semántico



Best Buy employees entered information into the blogs every day, using online forms that output RDFa. Myers told us that the use of RDFa makes 'human input from our store employees more visible on the Web.'

Best Buy is using Good Relations, a Semantic Web vocabulary for e-commerce that describes product, price, and company data.

Knowledge Graph



Minería Ontológica (*MO*)

Actualmente, con el gran crecimiento en las cantidades de ontologías disponibles sobre un dominio de conocimiento dado, ha llevado a la MO a explorar técnicas que puedan **extraer conocimiento adicional de un conjunto de ontologías**, para lograr un dominio de conocimiento más amplio.

1. Extracción de patrones de conocimiento,
2. Construir o enriquecer ontologías.

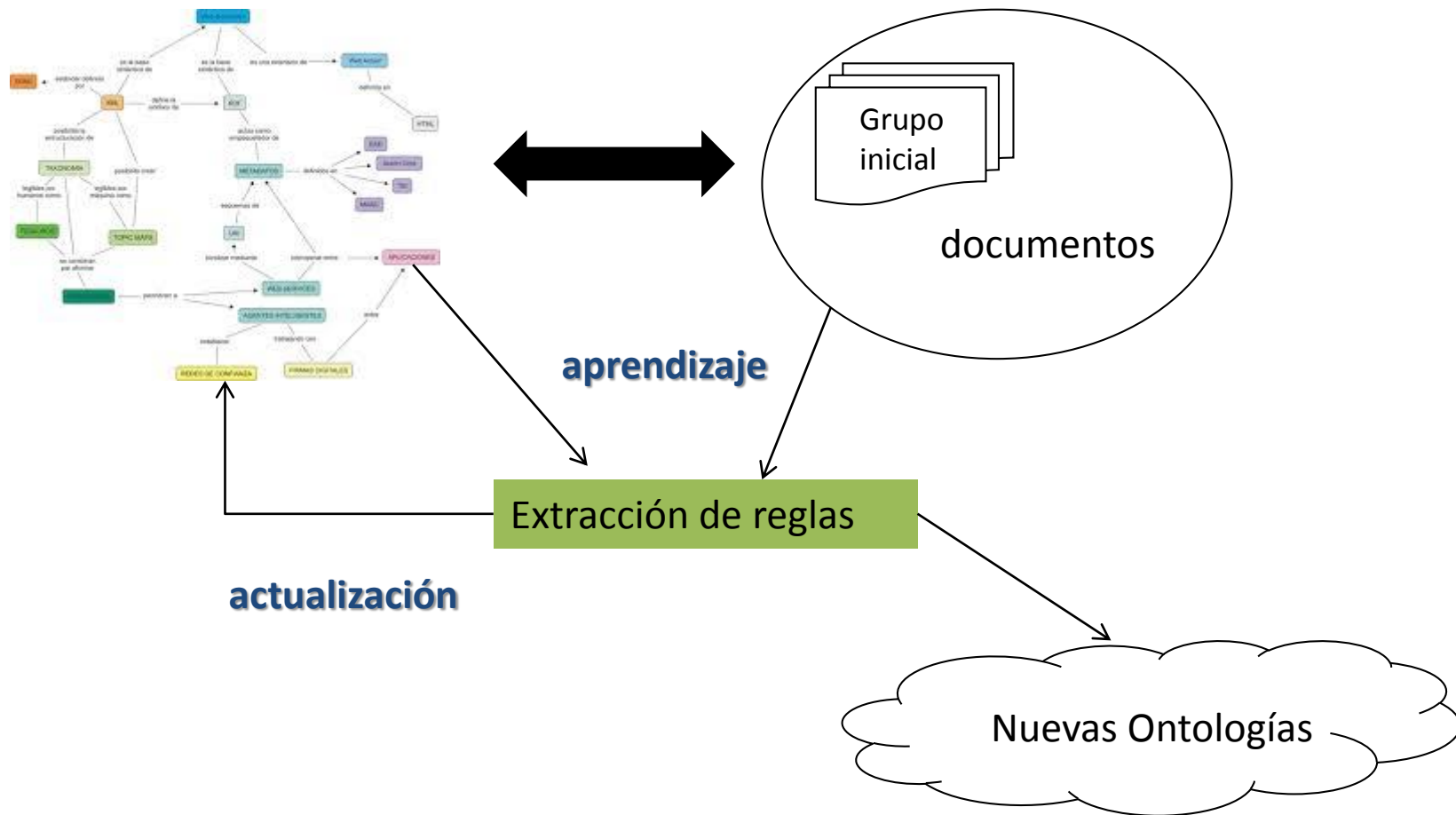
Minería Ontológica

- **Extracción de Reglas:** extrae reglas de un conjunto de ontologías.
- **Integración de Ontologías:** busca el vocabulario compartido entre varias ontologías.
- **Enlazado de Ontologías:** encuentra relaciones entre entidades de distintas ontologías.
- **Mezcla de Ontologías:** mezcla la información de varias ontologías con el fin de estandarizar conocimiento.
- **Alineación de Ontologías:** Identifica conceptos semejantes entre ontologías.

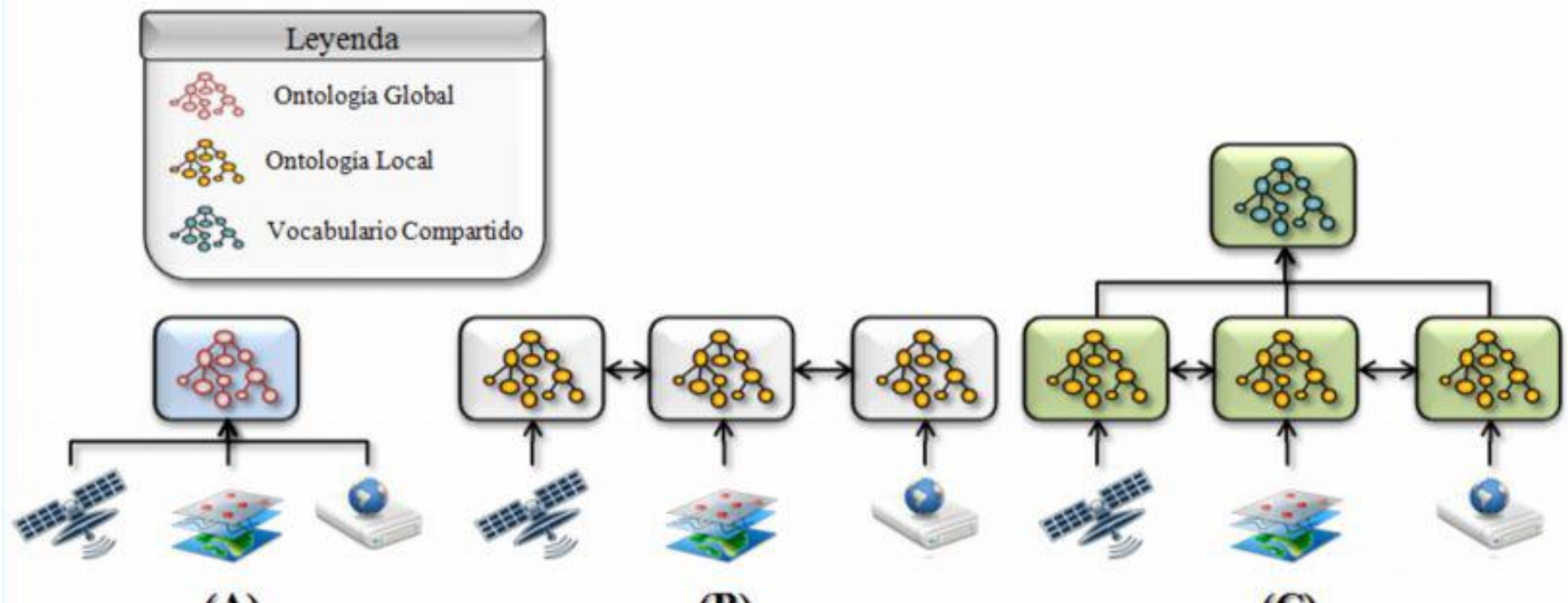
Ontologías
Emergentes

Extracción de Reglas

Ontología



Integración de ontologías



Alineación de ontologías

Identificar conceptos de una ontología que sean semejantes en las otras ontologías

Distancia semántica entre cada par de conceptos en ontologías distintas

Existen varios métodos y herramientas para realizar la alineación de ontologías

Alineación de ontologías

Técnicas de alineación de ontologías

- Basado en similitud lingüística (*linguistic matching*)
- Basado en similitud de grafos (*graph matching*)

Alineación de ontologías

alineación

Esta compuesto por los siguientes elementos:

- dos ontologías $O1$ y $O2$,
- un conjunto p de parámetros,
- un conjunto r de recursos para la alineación, y
- una función f de alineación, que retorna un conjunto de correspondencias A'

La función f integra diversos recursos para encontrar correspondencias entre dos conceptos.

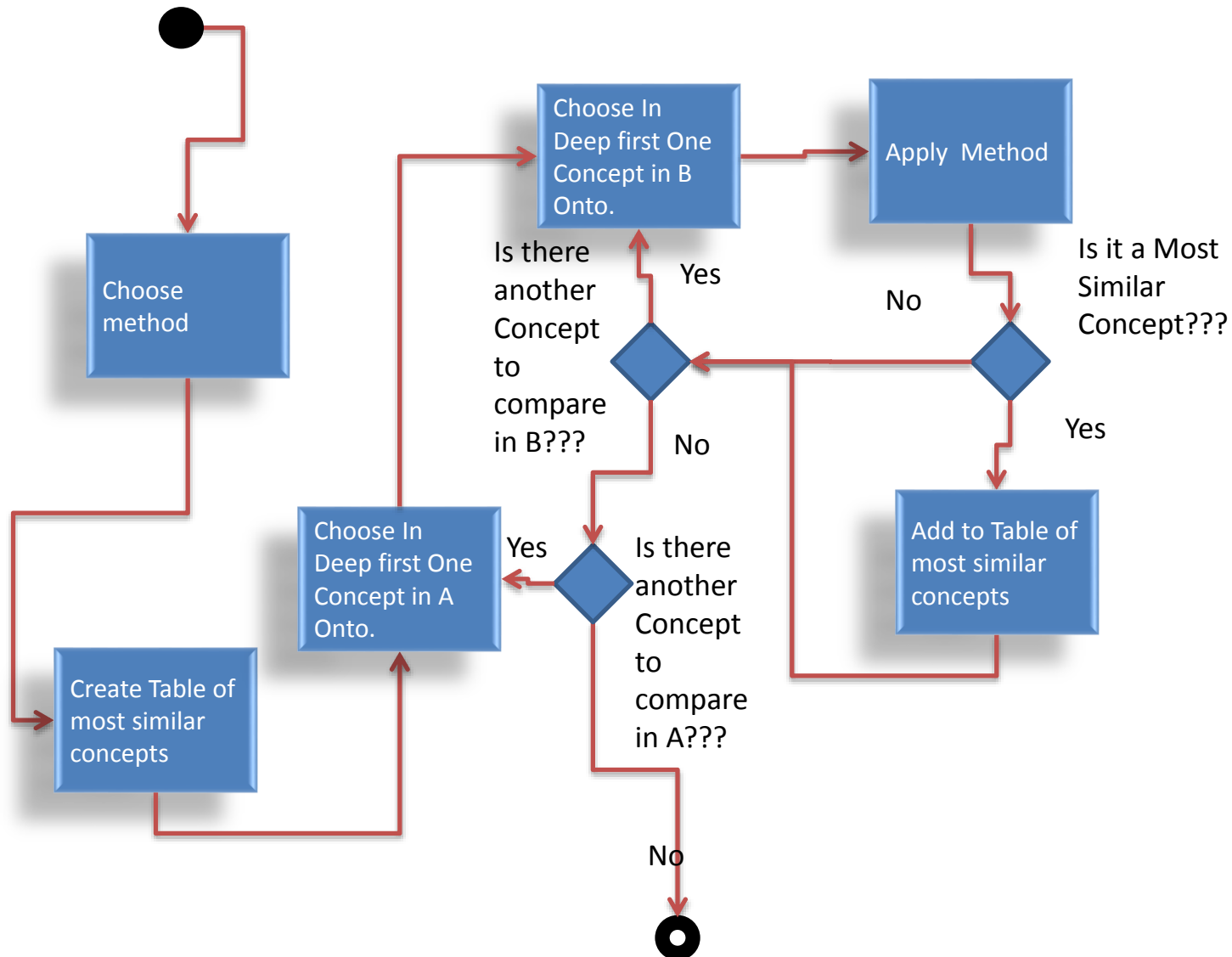
En cada $O1$ y $O2$ se analizan parte de sus elementos como: conceptos, propiedades de conceptos y jerarquía de conceptos.

- El conjunto p representa los requisitos para realizar la alineación; $p = \{\text{lenguaje de diseño OWL, número de elementos, vocabulario del idioma, no inferencias}\}$.
- El conjunto de recursos se refiere a los elementos empleados para obtener el conjunto de correspondencias $r = \{\text{conjunto medidas de similitud, algoritmo AdaBoost, algoritmo de clasificación K-Vecinos}\}$.
- El conjunto A' simboliza todas las correspondencias semánticas.

Alineación de ontologías

1. Verificar el lenguaje de diseño de las dos ontologías de entrada.
2. Obtener la jerarquía de clases de ambas ontologías.
3. Extraer información de cada clase.
4. Verificar si las propiedades de las clases son adecuadas para realizar el proceso de alineación.

Alineación



Métodos para calcular conceptos cercano

Suponen que CA sea un concepto o nodo en la ontología A y PA su predecesor. COM busca encontrar el concepto más parecido CB a CA en la ontología B, y PB (predecesor del concepto CB) a PA que aún no se ha encontrado.

Cuatro casos para calcular la similitud:

Caso A: El concepto CA coincide con CB en B y los predecesores PA y PB

Dados CA y PA, se busca en B por dos conceptos, CB y PB, de manera que la definición de PB coincide con la mayoría de las palabras que definen PA, y la mayoría de las palabras que definen CB coinciden con la definición de CA. En ese caso retorna CB (conocido como CMS).

Métodos para calcular conceptos cercano

Caso B: PA coincide con PB, pero no hay coincidencia entre CA y CB.

PB se encuentra, pero no CB.

- En este caso, de forma recursiva con PA como parámetro, se pasa a confirmar que PB es un predecesor de CA.
- Si un primo de PB (PB') pasa a ser la raíz de la ontología (OBRoot), entonces el algoritmo termina sin éxito. Si eso no sucede, entonces CA se busca en B a través de cada hijo de PB (el hijo debe coincidir con la mayoría de sus propiedades de CA).
- Si el candidato CB tiene hijos, se verifica que coincidan con los hijos de CA. Si un CB' se encuentra con las propiedades esperadas de CA, el algoritmo termina devolviendo con éxito CB'. De lo contrario, COM intenta encontrar CB' entre los hermanos de PB.
- Si eso no sucede, se busca entre los nietos de PB. Si CB' no se encuentra, entonces el valor más cercano a la CA es un hijo (desconocido, no está presente) de PB, por lo tanto, COM devuelve "hijo de PB" (lo que significa que un hijo PB que no existe todavía en B es el concepto más similar a CA).

Métodos para calcular conceptos cercano

Caso C: CA coincide con CB, pero no hay ninguna coincidencia entre el PA y PB.

- Si CB se encuentra, pero no PB, entonces se comprueba si el abuelo de CB en B es similar a la PA, o si el bisabuelo de CB en B es similar a PA.
- Si este es el caso, entonces el concepto más similar de PA en B es el abuelo o bisabuelo de CB y se termina el algoritmo.
- Si no se encuentra, entonces se verifica si la mayoría de las relaciones y los valores de CA coinciden con los de CB y si la mayoría de los hijos CA coinciden con la mayoría de los hijos de CB.
- Si las propiedades y los hijos coinciden, entonces la respuesta es CB y el algoritmo termina, a pesar de que PB no se ha encontrado en B.
- Si sólo una parte de las propiedades y los hijos coincide entonces la respuesta es "Probablemente CB" y se termina el algoritmo.
- Si no hay propiedades ni los hijos son iguales, entonces la respuesta es "no existe" y el algoritmo concluye.

Métodos para calcular conceptos cercano

Caso D: CA no coincide con el CB y PA no coincide con PB.

Si CB no existe y tampoco PB, entonces la respuesta de la COM es "no existe " y termina el algoritmo.

Detección de correspondencias semánticas

Las medidas de similitud seleccionadas para el proceso de alineación se dividen en dos grupos:

Similitud en base a términos: Se enfoca en el nombre de las entidades en las ontologías, principalmente en el nombre de las clases.

Similitud semántica: Su alcance va más allá de los nombres de las entidades, se enfoca en los componentes que definen la semántica de una clase:

- **Similitud entre propiedades de clases:** Considera las coincidencias existentes entre las propiedades de dos clases.
- **Similitud entre superclases:** Se refiere al par de superclases con mayor similitud respecto a dos clases comparadas.

Similitud léxica

- La Distancia de Levenshtein o distancia de edición (edit distance), fue creada en 1965 por el científico ruso Vladimir Levenshtein.
- La idea consiste en **determinar el número mínimo de operaciones requeridas para transformar una cadena de caracteres en otra,**
- Estas operaciones son: **inserción, eliminación o sustitución de un carácter.**

Por ejemplo, la distancia de Levenshtein entre los términos "hotel" y "hostal" es de dos, porque se necesitan al menos dos operaciones elementales para cambiar un término en el otro término.

Similitud entre propiedades

- Es la **verificación de si cada propiedad p_n del conjunto de propiedades P de una clase $C1$, coincide con otra propiedad p'_m del conjunto de propiedades de P' de otra clase $C2$.**
- Para realizar dicha comparación entre cada propiedad, se utilizan **las etiquetas de ambas propiedades (o nombres) como entrada a una medida de similitud léxica** (por ejemplo, la distancia de Levenshtein)

Similitud entre Superclases

También conocida como similitud entre conceptos ascendentes, a partir de dos clases iniciales.

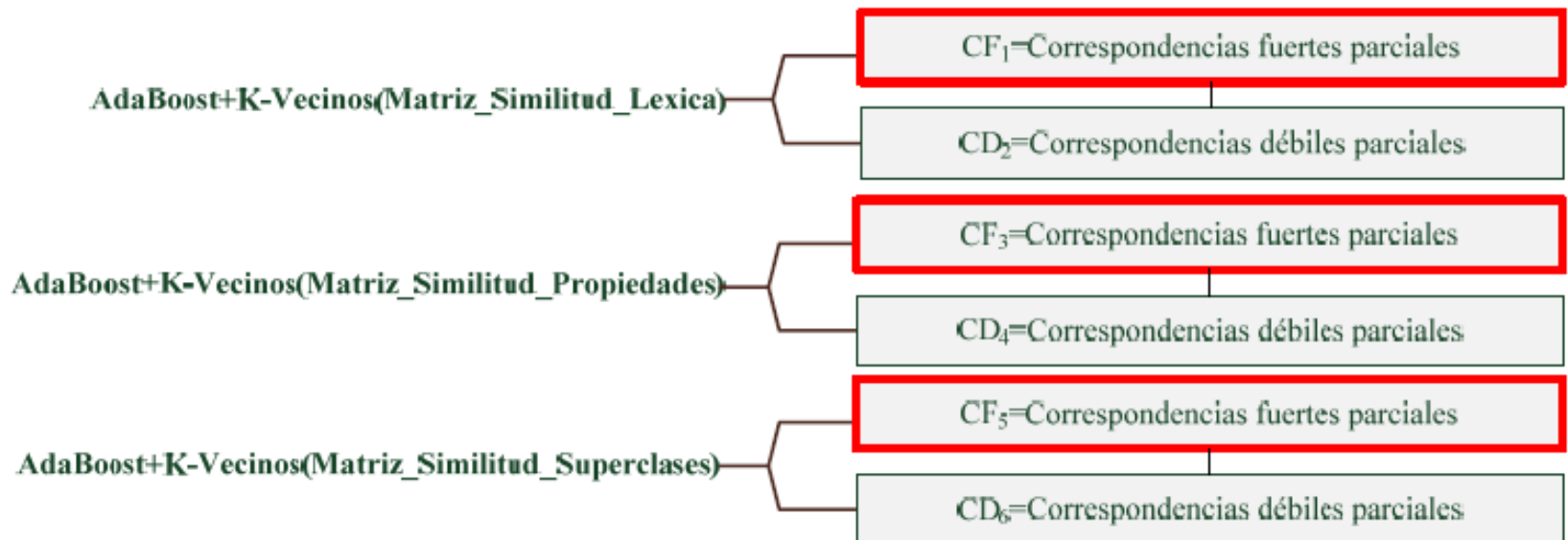
La similitud entre algún par de superclases puede influir para establecer una correspondencia entre dos clases.

Para calcular esta medida de similitud es necesario **calcular previamente la similitud léxica y de propiedades de todas las parejas de clases.**

El objetivo es **comparar parejas de superclases y seleccionar la pareja de superclases con mayor similitud entre sí**, a fin de encontrar de manera aproximada la pareja de superclases en donde coincidan o converjan las clases C1 y C2.

Correspondencias semánticas

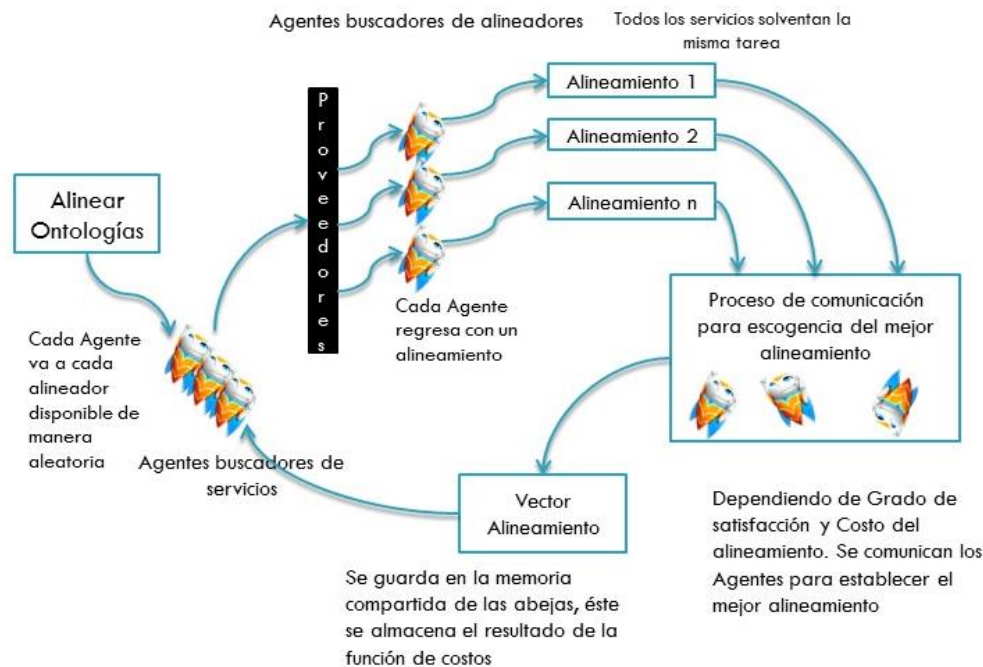
- Para clasificar si hay o no una correspondencia fuerte, se usa el algoritmo general de AdaBoost mas el de KVecinos (AdaBoost+K-vecinos)
- Como entrada recibe valores de una matriz de similitud.
- La salida es la clasificación de estos valores en dos posibles clases: una clase que agrupa los valores de similitud más altos y otra clase que agrupa los valores de similitud más bajos.



Criterios para determinar la correspondencia semántica fuerte

Caso	“Correspondencia fuerte parcial” SIMILITUD LÉXICA	“Correspondencia fuerte parcial” SIMILITUD ENTRE PROPIEDADES	“Correspondencia fuerte parcial” SIMILITUD ENTRE SUPERCLASES	“Correspondencia semántica fuerte”
1	SI	SI	SI	SI
2	SI	SI	NO	SI
3	SI	NO	SI	NO
4	NO	SI	SI	SI
5	SI	NO	NO	NO
6	NO	SI	NO	NO
7	NO	NO	SI	NO
8	NO	NO	NO	NO

Sistema de recomendación de Alineamiento de Ontologías usando ABC



- Los agentes abejas tienen la tarea de escoger una técnica de alineamiento
- Las abejas recolectoras se dirigen cada una a una fuente de alimento
- Cada fuente de alimento es una técnica de alineamiento
- Cada abeja recolectora regresa con la información de la calidad del néctar
- La calidad del néctar es transmitida a las abejas en espera
- Estas abejas deciden a que técnica de alineamiento ir basándose en la calidad
- Los resultados de la calidad de los alineamientos se van guardando en un vector que es la memoria global y compartida de las abejas.
- Todo este proceso se repite hasta que se alcancen los objetivos o se llegue a una condición de parada.

Sistema de recomendación de Alineamiento de Ontologías usando ABC

La ganancia $G(S_i)$ es calculada de la siguiente manera:

$$G(S_i) = \frac{S_a(S_i)}{CA(S_i)} \times P_c$$

- **S_i :** Servicio que se puede realizar para solventar una actividad solicitada, en nuestro caso corresponde a **una técnica de alineamiento**, y es equivalente a una fuente de néctar en el algoritmo ABC.
- **$G(S_i)$:** Ganancia obtenida por el servicio S_i (técnica de alineamiento i , ver fórmula 3.1), y es equivalente a la **calidad del néctar** en el caso del algoritmo ABC.
- **$S_a(S_i)$:** **Número de nodos alineados** por la técnica de alineación S_i . Es usado para calcular $G(S_i)$ que es la calidad del néctar.
- **$CA(S_i)$:** **Tiempo** que tarda una abeja en ir al **Servicio S_i** y regresar con resultados (en nuestro caso, este es el tiempo de cálculo empleado por la técnica de alineamiento i , que repercute también con la calidad del néctar $G(S_i)$).
- **P_c :** Probabilidad de conservar la opinión. Valor **pseudo-aleatorio**, con una **distribución normal**, dentro del rango de 0 y 1, el cual modifica el valor de $G(S_i)$.

Enlazado de Ontologías

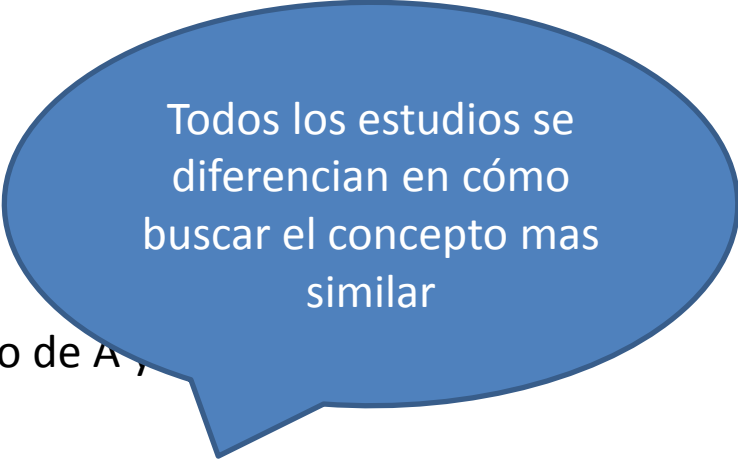
Enlazado

es el proceso para encontrar relaciones entre entidades que pertenecen a diferentes ontologías.

Enlazado débil de Ontologías: es una **correspondencia entre conceptos idénticos**. En este caso, básicamente lo que se realiza es la **intersección de las ontologías**, a partir de la cual se podrían hacer inferencias específicas en cada ontología.

Enlazado Fuerte de Ontologías: Es realizado de **manera semiautomático**, con la ayuda de un **experto del conocimiento** global que se está enlazando, el cual puede definir nuevos conceptos, así como **enlaces que relacionan conceptos** de ontologías distintas, creando así una **Meta-Ontología** con partes de conocimiento de las ontologías enlazadas.

Los resultados pueden ser usados para visualizar correspondencias, transformar una fuente en otra, crear un conjunto de relaciones o reglas entre las ontologías, generar consultas para extraer información desde las dos ontologías, etc.



Todos los estudios se
diferencian en cómo
buscar el concepto mas
similar

Dadas dos ontologías A y B

Para cada par de conceptos, uno de A y

- Buscar similitud léxica

- Buscar similitud semántica

Calcular una matriz de similitud para todos los pares

Escoger conceptos similares

Si existen conceptos similares

- copiar los conceptos similares

- Copiar los hijos y sus relaciones de los conceptos similares

- Detectar inconsistencias

- Resolver inconsistencias

Si no, no se puede realizar el enlazado



Enlazado
débil

Enlazado débil

Mezclado de ontologías

es el proceso donde varias ontologías dentro de un mismo dominio se unen para estandarizar el conocimiento, hacer crecer el conocimiento y tener el conocimiento total de manera local.

Mezclado

Los mezcladores unen ontologías que manejan el **mismo conocimiento**, pero con diferentes representaciones, o que poseen representaciones parciales de dicho conocimiento, tal que las ontologías pueden coincidir en ciertos conceptos y en otros no.

Mezcla Débil de Ontologías: se toma una ontología A, la copian como resultado C, y la van enriqueciendo con la otra B, comparando todos los conceptos de la ontología C (que son los mismos de A en este momento) con los de la ontología B, enriqueciendo los conceptos de C con sus conceptos semejantes de B. Dejando por fuera parte del conocimiento de B.

Mezcla Fuerte de Ontologías: Es una mezcla débil, pero incorporándole el conocimiento dejado por fuera de B,

Macro-algoritmo para la Mezcla Débil de Ontologías

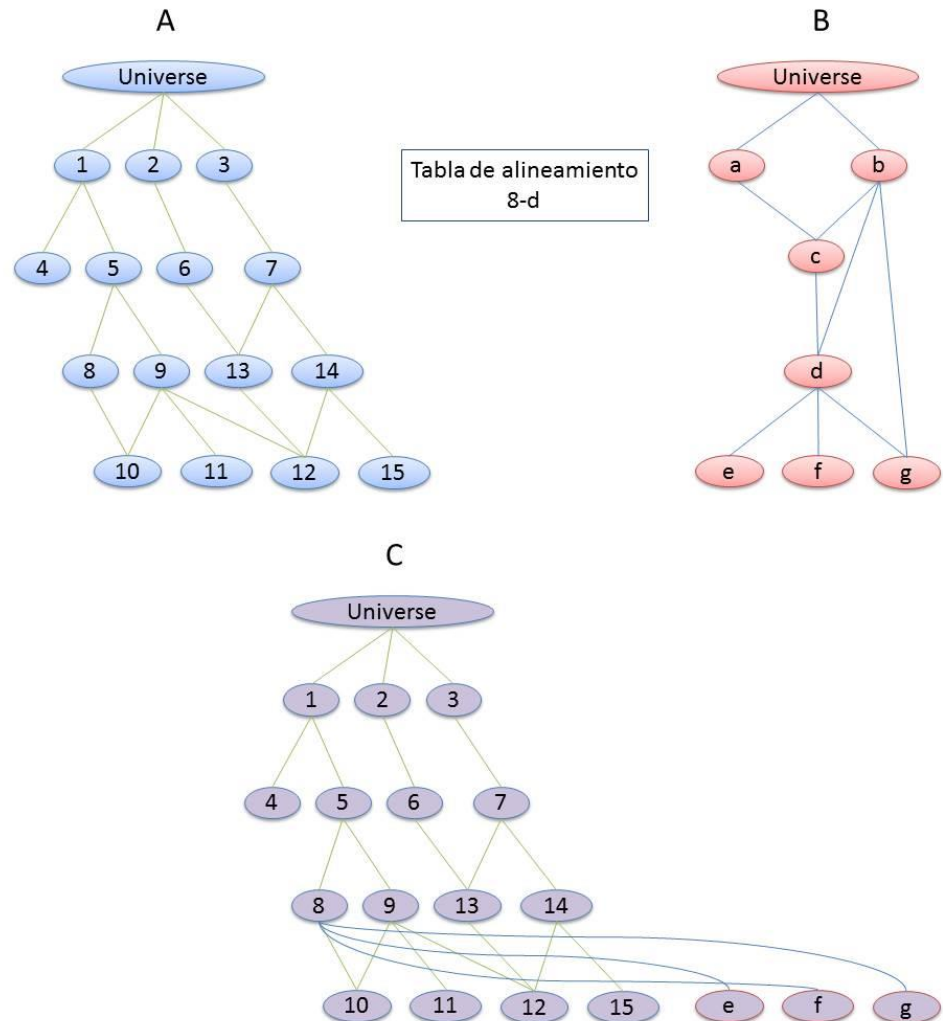
- **Copia** ontología de A a C.
- A partir del concepto raíz en C, se sigue un **recorrido en profundidad**:
 - **Buscar** en B cada concepto Cc de C (utilizando algún método (como CMS) para **calcular el concepto más cercano** en otra ontología (CMS).
 - Si hay un **CMS** en B, se seleccionan las relaciones del CMS en B que se pueden añadir a C, así como los nuevos conceptos, de la siguiente manera :
 - Se **enriquecen las relaciones** de Cc que son sinónimas con las del CMS;
 - Se **añaden las relaciones** nuevas del CMS a Cc.
 - Se **copian los conceptos** que no tenga C que se encuentren en las nuevas relaciones.
 - Se **detectan inconsistencias** entre las relaciones de Cc y CMS
 - Se **mantienen sólo las relaciones que sean consistentes** en Cc
 - Si no se logra resolver las inconsistencias, **permanecen en Cc relaciones originales** de A.
 - Se **copian los conceptos hijos** de CMS a C.
 - Si no hay un CMS, se toma el siguiente concepto Cc en profundidad

se diferencian en
cómo buscar el
concepto mas
similar

Se dejan
conceptos de B
por fuera

Algoritmos de Mezcla de Ontologías

Mezcla de A y B



El problema de **la mezcla tradicional** de ontologías (que nosotros hemos llamado **mezcla débil**), es que deja conocimiento sin ser incorporado en la ontología resultante.

Figure 1

1. Se realiza la mezcla débil,
2. Se incorporan los conceptos y relaciones dejadas por fuera.

The diagram illustrates the first case of ancestral alignment (1° caso) between two phylogenetic trees, A and B, to find a common ancestor for alignment.

Tree A (Left): A rooted tree with a root node labeled "Universe" (blue oval). It has three children: 1, 2, and 3 (blue ovals). Node 1 has children 4, 5, and 6; node 2 has child 7. Node 4 has children 8 and 9; node 5 has child 13; node 6 has child 14. Node 8 has child 10; node 9 has child 11; node 13 has child 12; node 14 has child 15. Node 10 has child 16; node 11 has child 17; node 12 has child 18. Node 15 has child 18. All nodes are blue ovals.

Tree B (Right): A rooted tree with a root node labeled "Universe" (red oval). It has two children: a and b (red ovals). Node a has child c; node b has child c and child d. Node c has child d. Node d has children e, f, and g. Node e has child h; node f has child i; node g has child i. All nodes are red ovals.

Alignment Table (Center): A box labeled "Tabla de alineamiento 1-a" (Alignment Table 1-a).

Alignment Process (Bottom): A large blue curved arrow points from the alignment table to the aligned trees. The aligned trees show the common ancestor "Universe" (blue oval) at the root. The left subtree is identical to Tree A. The right subtree is identical to Tree B. A line connects node 15 of Tree A to node c of Tree B, indicating their common ancestor. The nodes are colored blue for Tree A and red for Tree B.

Algoritmos de Mezcla de Ontologías

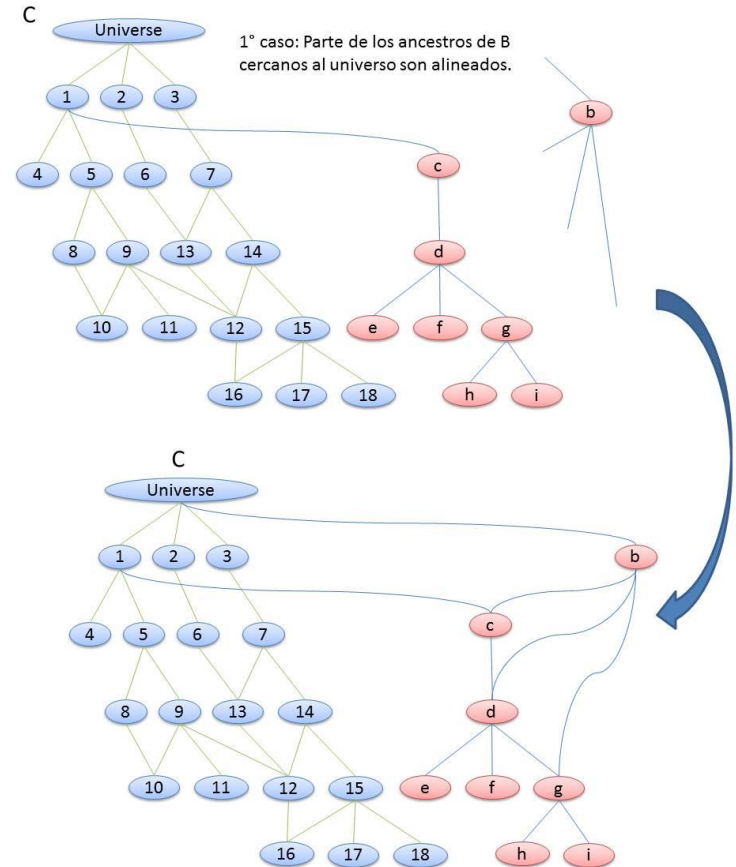
Segunda Parte: si de la ontología a la cual se le está extrayendo el conocimiento para ser agregado a la primera **queda aún conocimiento sin ser agregado**, se analizan los siguientes casos:

Caso 1:

- Se alinearon **parcialmente** los conceptos de B
- los nodos no alineados no se copian en la ontología resultado.

Solo bastaría con:

- **agregar los nodos no alineados a C**
- **copiar las relaciones** que no fueron copiadas o alineadas.



Algoritmos de Mezcla de Ontologías

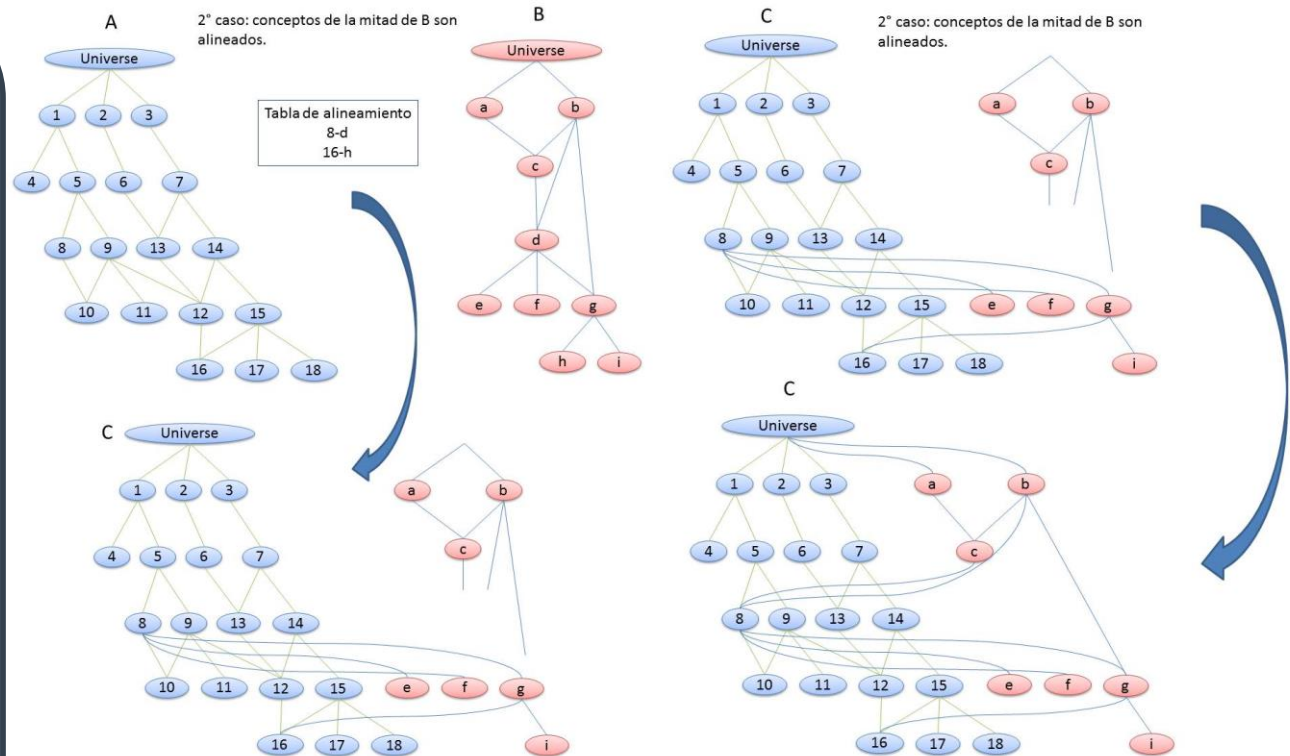
Caso 2:

- Se **alinearon** ciertos **nodos intermedios**
- Se deja a los ancestros sin ser copiados

La mezcla fuerte debe:

- agregar estos conceptos al universo de C
- Agregar las relaciones donde ellos participan

(como ocurre con **b-g**, **c-8** y **b-8**)



Algoritmos de Mezcla de Ontologías

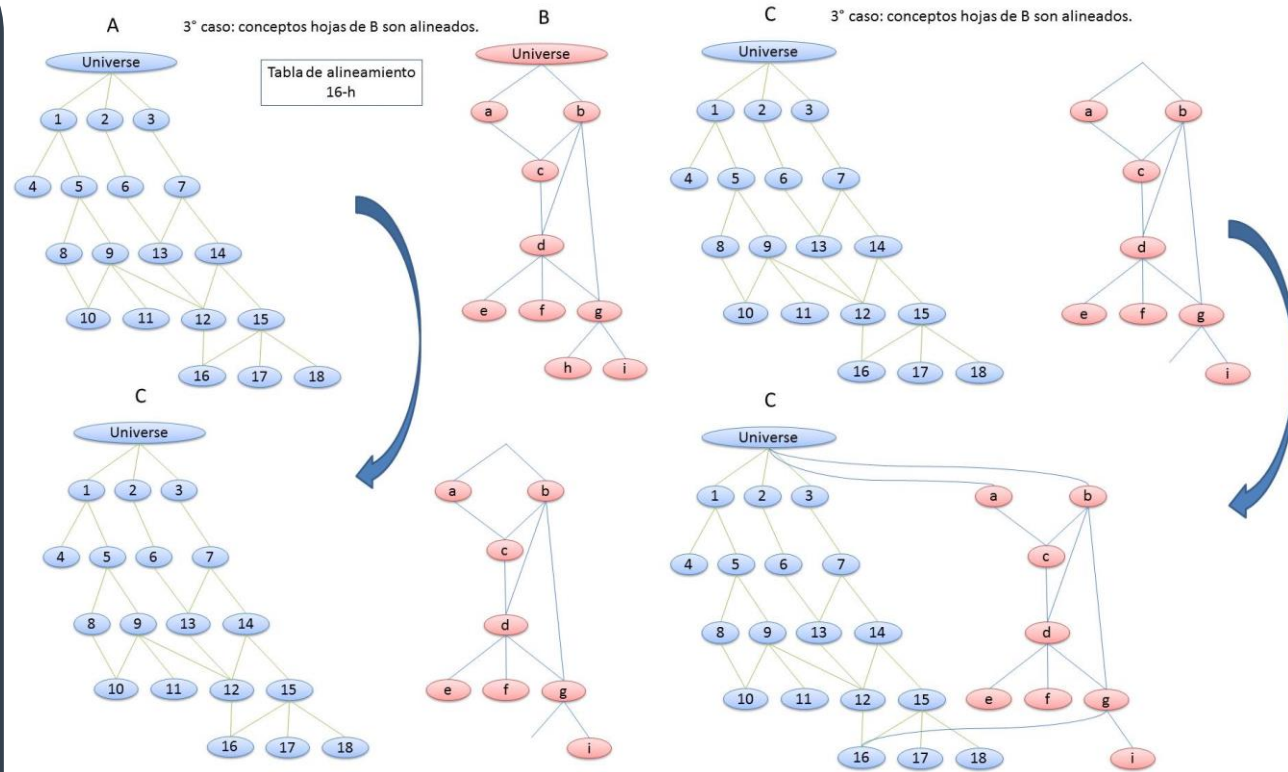
Caso 3:

- Solo los **nodos hojas se alinearon**
- Se deja por fuera de C un conjunto de conocimiento grande de B.

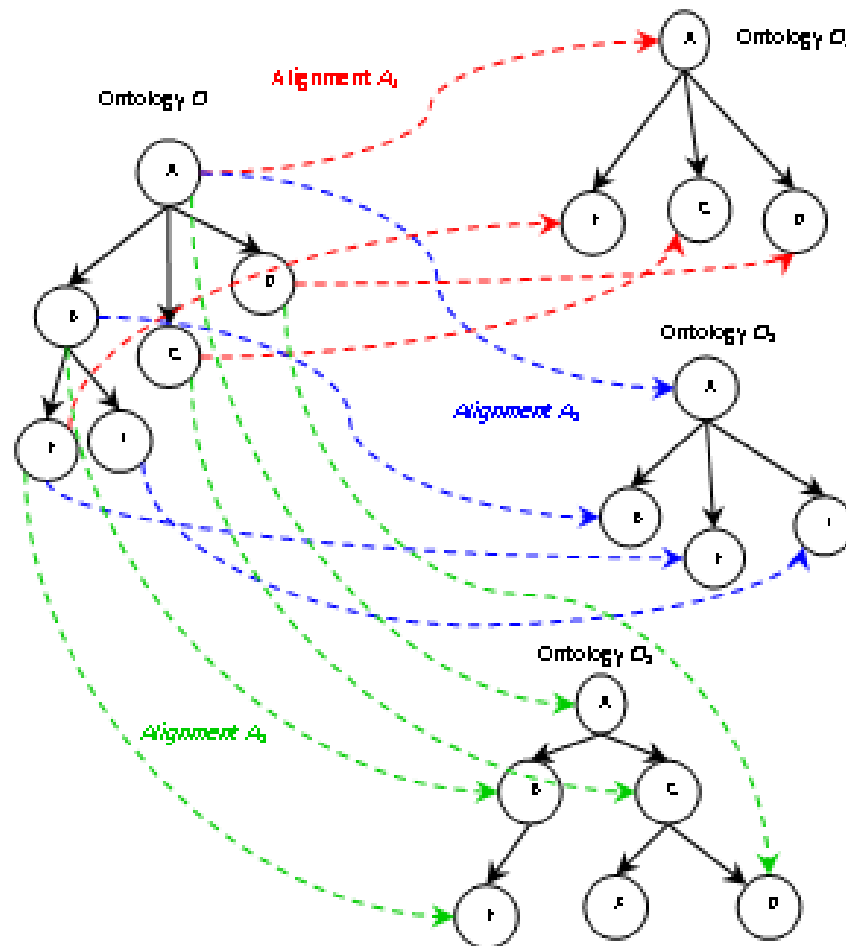
La mezcla fuerte debe

- **agregar** todos los conceptos **no copiados a C**
- buscar las relaciones que estos conceptos ya tenían con otros en B
- copiar también esas relaciones con los conceptos que fueron copiados o con los que fueron alineados

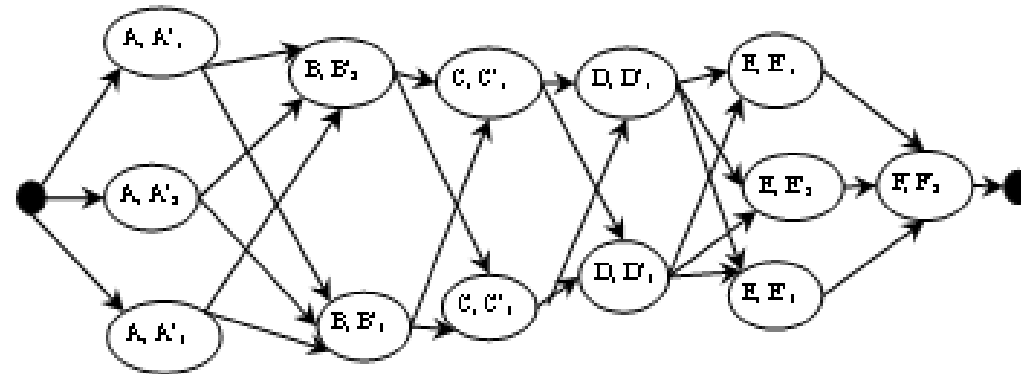
(g-16)



Un enfoque para la combinación múltiple de ontologías



Un enfoque para la combinación múltiple de ontologías



La similitud de los conceptos C y C' será proporcional a la similitud de los conceptos ancestrales

$$SA(C, C') = PC_A \times Sim(C, C') + \frac{2(1-PC_A)}{n(n+1)} \sum_{j=1}^m \sum_{i=1}^n (n+1-i) Sim(Anc_i(C), Anc_j(C'))$$

La similitud de los conceptos C y C' será proporcional a la similitud con los hermanos

$$SS(C, C') = PC_S \times Sim(C, C') + \frac{1-PC_H}{n} \sum_{i=1}^n \max(Sim(S_i, S'_1), \dots, Sim(S_i, S'_m))$$

La similitud entre los dos conceptos C y C' también será proporcional a la similitud de los descendientes directos

$$SD(C, C') = PC_D \times Sim(C, C') + \frac{1-PC_D}{n} \sum_{i=1}^n \max(Sim(H_i, H'_1), \dots, Sim(H_i, H'_m))$$

Medida de Similitud

$$MS(C, C') = \frac{SA(C, C') + SD(C, C') + SS(C, C')}{3}$$

Un enfoque para la combinación múltiple de ontologías

"Grado de enriquecimiento" (GE)

indicador de la cantidad de nuevos conceptos obtenidos por la ontología fuente después de seleccionar una alineación para un concepto.

GE de la ontología después de seleccionar la alineación de un concepto C con C 'de los nuevos conceptos que se pueden añadir a los corresponsales de la ontología:

- Niños de conceptos C 'no alineados (Nueva hipónimos) y sus descendientes
- Los hermanos de los conceptos C 'no alineados con el ancestro alineado inmediata (padre) (Nueva Cohiponímias) y sus descendientes.
- Conceptos antepasados de C 'no alineados (Nueva hiperónimos).

$$(C, C') = CHildren_Non_Aligned(C') \\ + Siblings_Non_Aligned(C') \\ + Ancestors_Non_Aligned(C')$$

CRITERIOS DE CALIDAD PARA LA FUSIÓN DE ONTOLOGÍA

- **Cobertura (o completitud):** grado de preservación de la información en la nueva ontología. Existen varios sub-casos,
 - cobertura con respecto a las ontologías fuente o objetivo,
 - con respecto a los conceptos retenidos de las ontologías (cobertura general), o
 - cobertura foliar definida como el grado en que se conservan los conceptos de las hojas de entrada.
- **Compacidad:** tamaño relativo de la fusión obtenida. Se define como la suma del número de conceptos de ambas ontologías de entrada menos el número de conceptos alineados. El tamaño relativo debe ser 1, como máximo.
 - Los valores inferiores a uno indican una compacidad mejorada.
 - Tamaño de resultado absoluto: número de conceptos en la nueva ontología resultante.

CRITERIOS DE CALIDAD PARA LA FUSIÓN DE ONTOLOGÍA

- **Redundancia:** Las medidas se basan en la siguiente ecuación:

$$LPB = LPS + LPT - ML$$

LPB es el número mínimo de rutas a las hojas, sin ninguna redundancia; LPS y LPT es el número de rutas a las hojas, en cada una de las dos ontologías de entrada, respectivamente, y ML es el número de conceptos de hoja que están alineados.

- La redundancia relativa en una nueva ontología resultante es la relación entre el número de trayectorias foliares en él y LPB.
 - Los valores superiores a 1 indican la introducción de rutas redundantes para los conceptos fusionados.
 - Un valor de 1 es típicamente óptimo, ya que indica el éxito en evitar rutas redundantes.
 - Valores inferiores a 1 implican que algunos conceptos en las hojas no están cubiertos con la fusión resultante, pérdida de información y la reducción de cobertura (en este caso, las hojas).