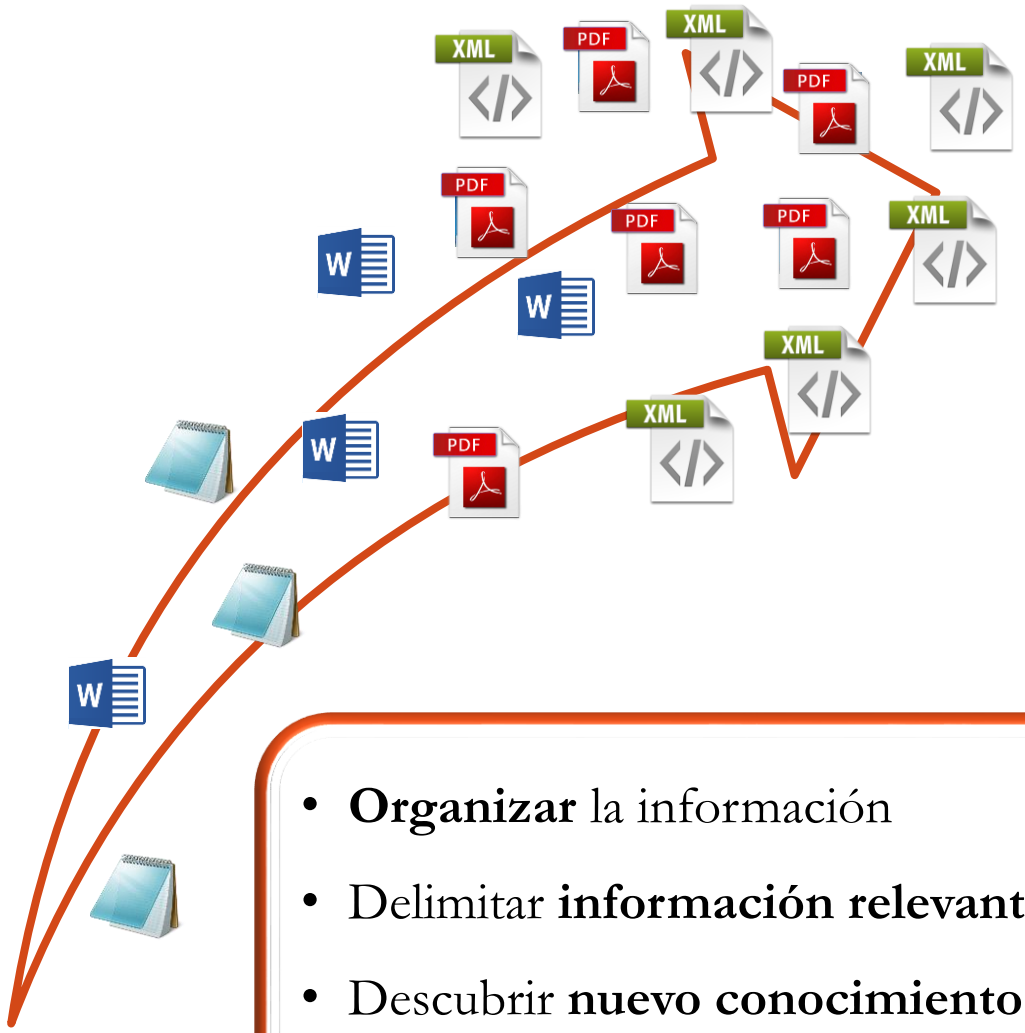
SemClustDML: algoritmo para agrupar artículos científicos basado en la información brindada por las referencias bibliográficas

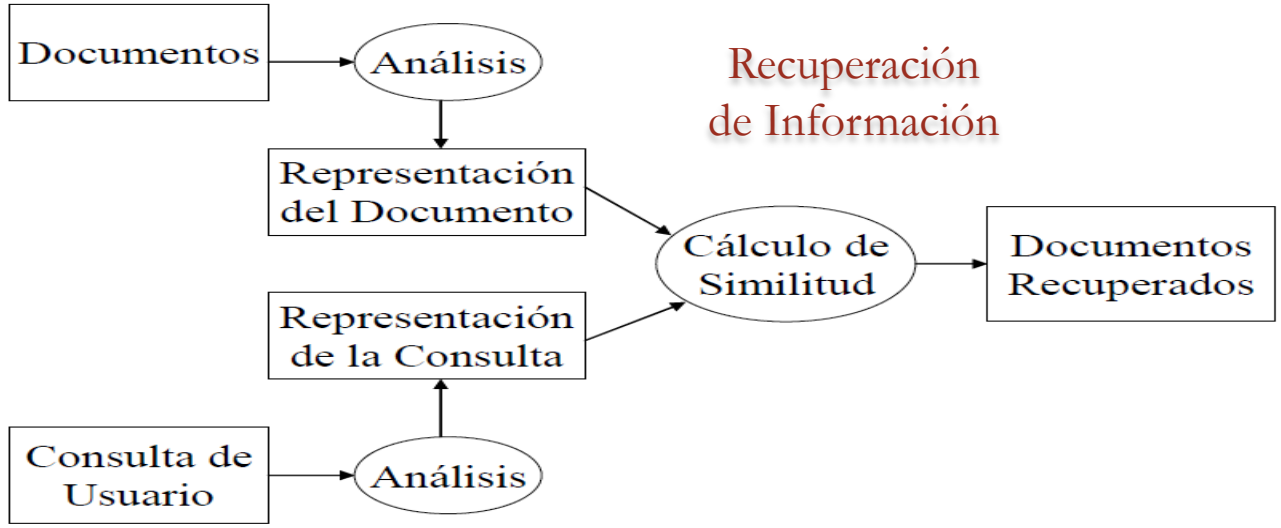
SemClustDML: algorithm to clustering scientific papers based on information provided by bibliographic references

Autores: Lisvandy Amador[1], María M. García[2], Daniel Gálvez Lío[2,3], Damny Magdaleno[2, 3]

[1]Instituto de Biotecnología de las Plantas, Villa Clara, Cuba.
[2]Universidad Central "Marta Abreu de Las Villas", Villa Clara, Cuba.
[3]Universidad Metropolitana del Ecuador (UMET), Quito, Ecuador.

Antecedentes





Continuo crecimiento de los datos

Necesidad de descubrir conocimiento en la información recuperada

Categorización, clasificación y **agrupamiento**

- **Organizar** la información
- Delimitar **información relevante**
- Descubrir **nuevo conocimiento** a partir de la información

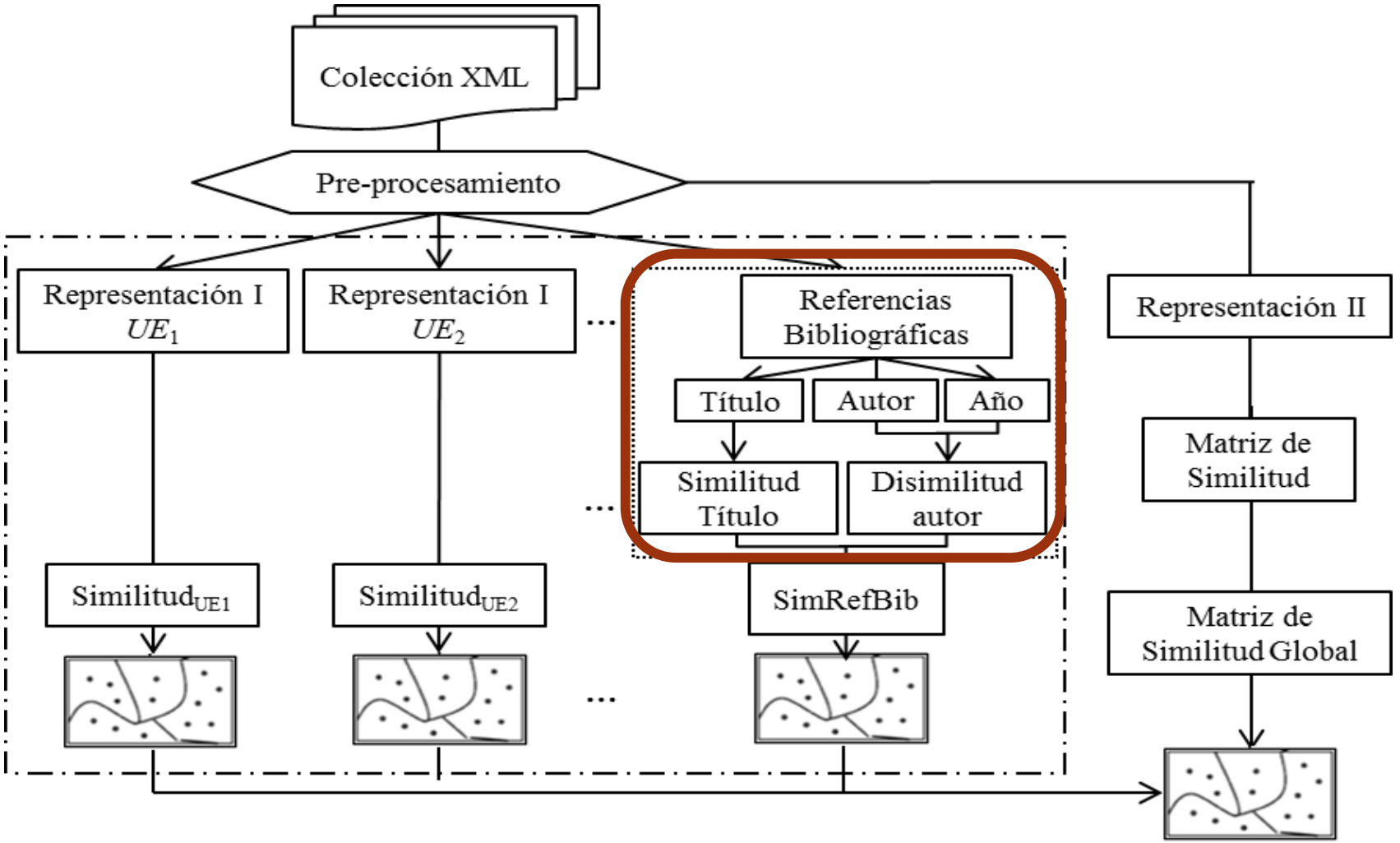
Los algoritmos de agrupamiento se pueden clasificar siguiendo diversos criterios, como pueden ser:

- Tipo de los datos de entrada,
- Criterios para definir la similitud entre los objetos, conceptos en los cuales se basa el análisis y forma de representación de los datos.

Si la participación del usuario influye en el agrupamiento, se tienen otras dos clasificaciones:

- algoritmos de agrupamiento automático y
- algoritmos de agrupamiento semiautomático.

Metodología para el agrupamiento



Magdaleno, D., et al (2015). "Clustering XML Documents Using Structure and Content based on a New Similarity Function OverallSimSUX". Computación y Sistemas Vol.19 Nro.1, 2015.

Subunidad *Título*

- En el preprocesamiento se tienen en cuenta posibles frases y no palabras independientes (*fuzzy set*).
- Representación VSM.
- Cálculo de la similitud *Título*.(ST)

Subunidades *Autor* y *Año*

- **VSM'(MAutor, NDoc)** = {Cij, NPij, AIij, AFij}
  - **Cij**: autor *i* en el documento *j*
  - **NPij**: autor *i* no es referenciado como principal en el documento *j*
  - **AI*ij***: menor año en que es referenciado el autor *i* en el documento *j*
  - **AFij**: mayor año en que es referenciado el autor i en el documento *j*
- Cálculo de la disimilitud *Autor*.(DisAut)

Medida general de semejanza

$$SimRefBib(i,j) = \begin{cases} ST(i,j)^{\,DisAut(i,j)} & si \quad ST(i,j) > 0 \\ 1 - DisAut(i,j) & si \quad ST(i,j) = 0 \end{cases}$$

$ST(i,j)$: Similitud título entre los documentos i, j

Planteamiento del problema

En general los algoritmos de agrupamiento asumen que documentos considerados similares presenten valores de similitud altos y los que no lo son presenten valores bajos; pero en muy pocos casos el valor de similitud es cero.

Al aplicar algunos de estos algoritmos usando como entrada la matriz de similitud obtenida con la función *SimRefBib*, no se garantiza obtener siempre buenos resultados en el agrupamiento, debido en gran medida, por la forma en que internamente cada uno obtiene los grupos.

Esto no significa que la función *SimRefBib* no sea capaz de discernir de manera correcta entre los elementos que deben pertenecer a cada grupo, pero garantiza que el diseño de un algoritmo que se adapte a estas características especiales favorecerá considerablemente el resultado del agrupamiento de artículos científicos.

Objetivo

Es por ello que se propone como objetivo de este trabajo: Desarrollar un algoritmo de agrupamiento que haga uso de las características especiales de la matriz de similitud obtenida con la función *SimRefBib* para mejorar los resultados del agrupamiento de artículos científicos basado en las referencias bibliográficas.

Algoritmo de agrupamiento *SemClustDML*

Algoritmo 1. Algoritmo de agrupamiento *SemClustDML*
Entrada: Matriz de similitud *matriz*, Conjunto de n Objetos $(O = \{o_1, o_2, \dots, o_n\})$, umbral de similitud γ, longitud mínima de cada clúster l, cantidad de elementos aleatorios a seleccionar para comprobar si los clústers son agrupables v.
Salida: Lista de clústers formados (C)
Inicio:

1. Búsqueda de los centroides iniciales: $C = \{o_1, o_2, \dots, o_k\}$, $k \leq n$, donde cada o_i se considera un nuevo clúster c.
2. Asignación de cada objeto $o_i \notin C$ al c_j correspondiente.
3. Si $cup \leftarrow \bigcup_{g,h=1}^{k} cap(c_g, c_h) = \emptyset$, donde $cap = c_g \cap c_h$, entonces ir al paso 5.
4. Determinar para cada $o_i \in cap(c_g, c_h) \neq \emptyset$, el c_j correspondiente, donde $sim \leftarrow \frac{\sum_{r=1}^{m_j} matriz\,(o_i, c_{jr})}{m_j}$ es máxima, m_j cantidad de elementos en c_j. Ir paso 3.
5. $\forall o_i \notin C, c_j \leftarrow (c_j \cup o_i)$, donde $sim \leftarrow \frac{\sum_{r=1}^{m_j} matriz\,(o_i, c_{jr})}{m_j}$ es máxima.

Fin

Pasos para refinar el resultado del agrupamiento

1. Para cada c_s formado seleccionar (si es posible) 2 nuevos centroides aplicar los pasos del 2 al 5 del algoritmo *SemClustDML*.

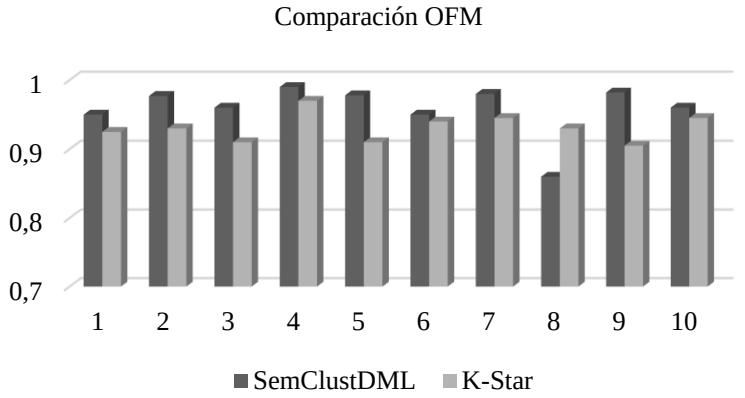
2. $\forall o_i \in c_s$ y longitud de c_s menor que l ; $c_j \leftarrow (c_j \cup o_i)$, donde $sim \leftarrow \frac{\sum_{r=1}^{m} matriz(o_i, c_{jr})}{m}$ es máxima.

3. $\forall C_i, C_j \, con \, i, j = \overrightarrow{1..k}$, k cantidad de clúster, verificar si C_i, C_j son agrupables.

División de cluster
Tamaño del clúster
Clústeres agrupables

Resultados y discusión

No. Corpus	Cantidad de documentos	Cantidad de clases	Temas que trata
Conjuntos de documentos XML confeccionados a partir de documentos recuperados del sitio del ICT del Centro de Estudios de Informática de la Universidad Central "Marta Abreu" de Las Villas http://ict.cei.uclv.edu.cu			
1	32	2	Fuzzy Logic, SVM
2	25	2	Rough Set, Association Rules
3	32	2	Rough Set, SVM
4	28	2	Association Rules, Fuzzy Logic
5	32	2	Association Rules, SVM
Recopilación de documentos del repositorio IDE-Alliance, internacionalmente utilizados para evaluar agrupamiento. Proporcionados por la Universidad de Granada. España.			
6	28	3	Copula, Belief Propagation, CL
7	19	2	Copula, Belief Propagation
Documentos pertenecientes al sitio ICT y al repositorio IDE-Alliance			
8	41	4	Rough Set, Copula, Belief Propagation, CL
9	29	2	Copula, SVM
10	38	3	Copula, SVM, Belief Propagation



Experiment			SemClustDML-INEXK_STAR
Z			-2.293
Aymp. Sig (2-tailed)			0.022
Monte Carlo Sig (2-tailed)	Sig.		0.021
	95% Confidence Interval	Lower Bound	0.018
		Upper Bound	0.026
Monte Carlo Sig (2-tailed)	Sig.		0.010
	95% Confidence Interval	Lower Bound	0.008
		Upper Bound	0.012

El algoritmo se comporta de manera estable sobre el caso promedio, el cual tiene una complejidad computacional O(nlog(kn)), sin consideran el refinamiento y una complejidad de O(n2) considerando el refinamiento.

Conclusiones

La función de similitud *SimRefBib*, especialmente diseñada para el agrupamiento de artículos científicos permite discernir de manera correcta entre los grupos que deben formarse para una colección de documentos dada, sin embargo, surge la necesidad de diseñar un algoritmo de agrupamiento que sea capaz de adaptarse a las características especiales de la matriz resultante del cálculo de esta función para lograr buenos resultados en el agrupamiento de este tipo de documentos.

Se implementó el algoritmo de agrupamiento para artículos científicos *SemClustDML* el cual hace uso de las características especiales de la matriz *SimRefBib* para mejorar el desempeño en el agrupamiento de este tipo de documentos. Este algoritmo cuenta con dos etapas: la etapa del agrupamiento propiamente dicha y una segunda etapa que consta de tres fases en las cuales se refina el resultado del agrupamiento.

La comparación del algoritmo *K-Star* con el algoritmo *SemClustDML* propuesto en esta investigación arrojó que existen diferencias significativas para la medida *OFM*, obteniéndose mejores resultados para el algoritmo *SemClustDML*.