5as Jornadas "Matemáticas Everywhere"

Una nueva taxonomía de colecciones y de funciones de similitud para su comparación

A new taxonomy of collections and similarity functions for comparing them

Alfredo Sánchez Alberca

Organizan:

Centro Internacional de Encuentros Matemáticos
Castro Urdiales, España

18, 19 y 20 de junio de 2018

Resumen

Las colecciones de objetos, entendidas como agrupaciones de objetos con entidad propia, están presentes en todos los ámbitos de nuestro mundo. Aunque no existe aún una definición matemática formal de colección, las colecciones se usan en muchas aplicaciones científicas, y especialmente en Ciencias de la Computación, donde se utilizan distintas estructuras de datos y con distintas propiedades para representarlas.

En este artículo se presenta una nueva clasificación taxonómica de los tipos de colecciones más comunes organizada de acuerdo a cuatro propiedades estructurales: homogeneidad, unicidad, orden y cardinalidad. Sobre la base de esta taxonomía se presenta un catálogo de funciones de similitud para comparar los distintos tipos de colecciones. Este catálogo resulta útil para identificar las funciones de similitud más apropiadas para comparar dos colecciones dadas y aplicarlas automáticamente.

Palabras Clave: Colección, Similitud, Disimilitud, Distancia, Taxonomía.

Abstract

Collections of objects, understood as groups of objects with its own right, are present all over the world. Despite of the fact that there is no formal mathematical definition of collection, collections are used in many applications of Science. In Computer Sciences, in particular, collections are represented by different data structures with different properties.

This paper presents a new taxonomic classification of the most common types of collections organized according to four estructural properties: homogeneity, unity, order and cardinality. Based on this taxonomy we also present a catalog of similarity functions for comparing the different types of collections. This catalog is helpful to identify the most suitable similarity functions to compare two given collections and apply them automatically.

Keywords: Collection, Similarity, Dissimilarity, Distance, Taxonomy.

1. Introducción

Las colecciones de objetos, entendidas como agrupaciones o agregaciones de elementos u objetos de diversa índole, están presentes en prácticamente todos los ámbitos de nuestro mundo: los libros de una biblioteca, los alumnos de un curso, los productos de una tienda, los eventos de una agenda, los genes de un individuo, los elementos químicos que componen una sustancia, los trenes que hacen un determinado recorrido, los pixels de una foto y las fotos de un álbum, los jugadores de un equipo y los equipos de una liga, el ranking en una competición, los cartas que llegan a una oficina de correos, los platos de un menú, los menús de un restaurante y los restaurantes de una ciudad, etc. Cualquier aplicación o sistema inteligente que pretenda abordar un problema en el que aparezcan colecciones deberá disponer de un sistema de modelado y representación de estas. Así, en el ámbito de las ciencias de la computación, las colecciones de objetos están presentes de manera explícita en prácticamente todos los sistemas de información, desde los simples arrays en los lenguajes de programación, hasta los registros en las bases de datos, pasando por las listas de procesos, las colas de impresión, o estructuras de datos como los conjuntos, las bolsas, etc., y en última instancia, todo se reduce a colecciones de bits.

Sin embargo, no hay una correspondencia clara entre las colecciones del mundo real y el modelo o estructura de datos más apropiada para representarla, porque esto depende del nivel de abstracción que se haga y de las primitivas de representación de que disponga un determinado sistema. A veces se utiliza una misma estructura de datos para representar colecciones con características distintas y otras veces una misma colección se representa de manera distinta en diferentes sistemas, lo que suele provocar confusiones, incongruencias y errores.

Otro problema más importante, derivado de la falta de sistemas de representación adecuados para las colecciones de objetos, es su comparación. Aunque en la literatura hay abundantes ejemplos de funciones de similitud o disimilitud, no existe un criterio claro de cuál es la más apropiada para cada caso, y ello es debido, en parte, a que a menudo las representaciones de las colecciones no reflejan las características estructurales que condicionan su semántica, y por tanto, estas características no pueden ser explotadas por las funciones de similitud utilizadas en su comparación, obteniendo muchas veces pobres estimaciones o incluso erróneas. Esto dificulta considerablemente la comparación de colecciones, y por tanto todas tareas que requieren la comparación de colecciones, como las búsquedas semánticas [16, 36], las correspondencias entre ontologías [9] o el enlazado de datos [2].

En este artículo se presenta una nueva clasificación taxonómica de los tipos de colecciones más comunes organizadas de acuerdo a cuatro propiedades estructurales: homogeneidad, unicidad, orden y cardinalidad. Esta taxonomía introduce, además de los tipos de colecciones clásicos de la teoría de conjuntos, nuevos tipos de colecciones que hasta ahora no habían sido modelizados por ninguna ontología de colecciones, lo que permite caracterizar mucho mejor cada tipo de colección y elegir la representación más adecuada en función de sus características semánticas. Sobre la base de esta taxonomía se define un catálogo de funciones de similitud para comparar los distintos tipos de colecciones y se muestran ejemplos de uso.

2. Taxonomía de colecciones

Tradicionalmente las colecciones, entendidas como agrupaciones o agregaciones de elementos u objetos, han sido estudiadas por la Teoría de Conjuntos dentro de las Matemáticas [11, 1, 3]. En las últimas décadas también se han estudiado como tipos abstractos de datos dentro del ámbito de las bases de datos y la representación del conocimiento [20, 21, 15]. Tanto en un caso como en otro, existen dos aspectos que deben tenerse en cuenta a la hora de definir una colección: el contenido, es decir, los elementos que componen la colección, y la estructura, es decir la

forma en que los elementos se organizan dentro de la colección y las restricciones que se aplican sobre ellos.

Dependiendo de las características estructurales de la colección y de las restricciones que se apliquen sobre sus elementos, se obtienen diferentes tipos de colecciones, con distintas semánticas y operaciones lógicas. La clasificación que se propone en este artículo se basa en cuatro propiedades estructurales de las colecciones:

Homogeneidad Indica si todos los elementos de la colección deben ser del mismo tipo (*colección homogénea*) o no (*colección heterogénea*).

Unicidad Indica si todos los elementos de la colección deben ser diferentes (*colección con unicidad*) o pueden repetirse (*colección sin unicidad*).

Orden Indica si existe un orden establecido entre los elementos de la colección (*colección ordenada*) o no (*colección no ordenada*). Este orden no tiene nada que ver con cualquier orden subyacente en los tipos de datos de los elementos que componen la colección, sino con el orden de los elementos en la propia estructura de la agregación.

Cardinalidad Indica si el número de elementos que forman la colección es fijo (*colección con cardinalidad fija*) o no (*colección con cardinalidad variable*).

Realizando un emparrillado sobre estas propiedades, surgen un total de 16 tipos de colecciones que se presentan en la tabla 1.

Tabla 1. Clasificación taxonómica de colecciones de acuerdo a sus propiedades de homogeneidad, unicidad, orden y cardinalidad.

		Cardinalidad variable		Cardinalidad fija	
		Sin unicidad	Con unicidad	Sin unicidad	Con unicidad
Heterogéneo	Sin orden	*Multiheteroconjunto*	*Heteroconjunto*	*Caja*	*Heterocombinación*
	Con orden	*Lista*	*Heteroranking*	*Tupla*	*Heterovariación*
Homogéneo	Sin orden	*Multiconjunto*	*Conjunto*	*Multicombinación*	*Combinación*
	Con orden	*Secuencia*	*Ranking*	*Vector*	*Variación*

De acuerdo a las características estructurales que tengan, estos tipos de colecciones se pueden organizar en la clasificación taxonómica que aparece en la figura 1.

En la raíz de la taxonomía tenemos los *multiheteroconjuntos*, que son las colecciones con menos restricciones estructurales, ya que son agrupaciones de elementos sin homogeneidad, sin unicidad, sin orden y sin cardinalidad fija. Un ejemplo podría ser la bolsa de la compra en un supermercado ya que en la bolsa podemos poner elementos u objetos de distintos tipos (heterogeneidad), podemos poner objetos repetidos (sin unicidad), el orden de los objetos no importa (sin orden) y podemos poner cualquier número de objetos (cardinalidad variable).

En el extremo opuesto de la taxonomía están las *variaciones*, que son las colecciones con más restricciones estructurales, ya que son homogéneas, con unicidad, con orden y con cardinalidad fija. Este tipo de colecciones es bastante conocido en área de la combinatoria.

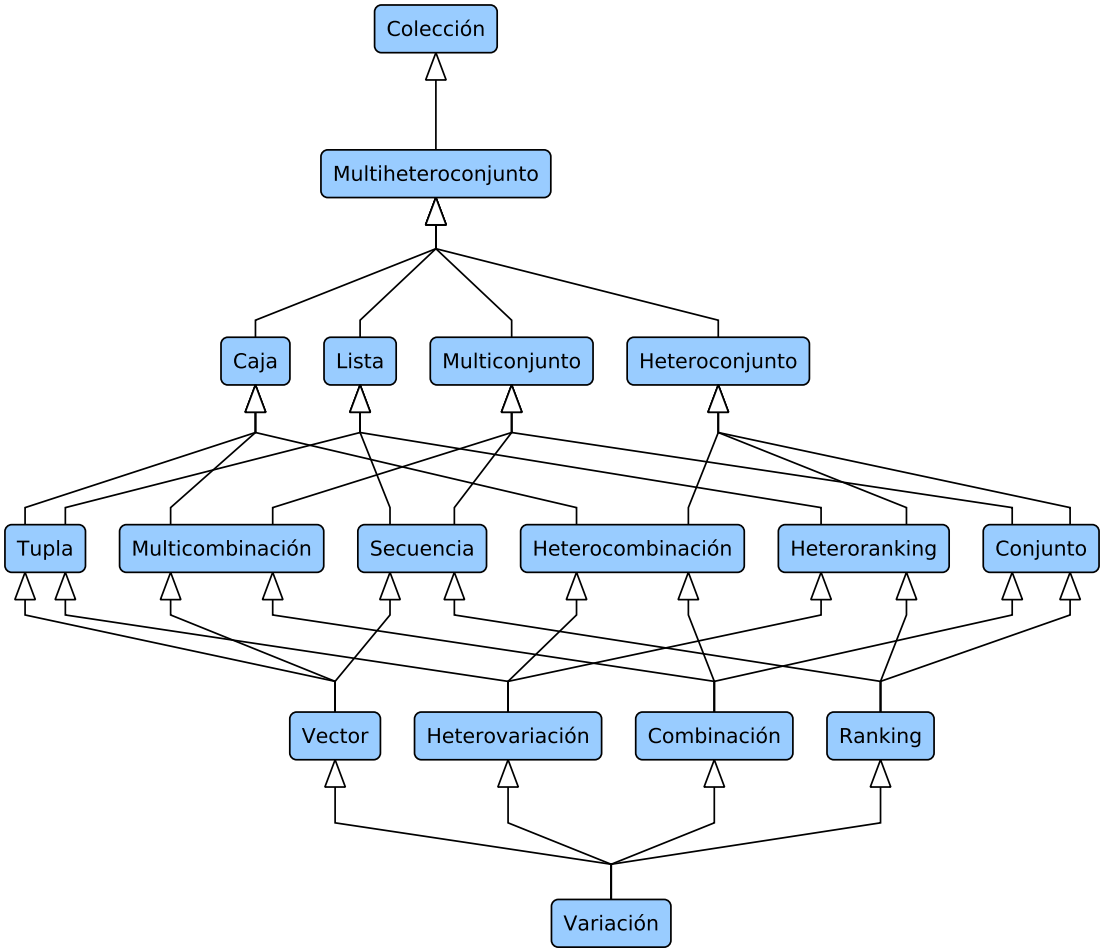


Figura 1. Taxonomía de colecciones.

Entre estos dos tipos de de colecciones existen otros 14 tipos de colecciones en distintos niveles de la taxonomía según se apliquen más o menos restricciones sobre su estructura. Algunas de estas colecciones bien conocidas por ser clásicos en distintas ramas de la Matemática, como los *conjuntos*, las *combinaciones*, los *ranking*, las *tuplas* o los *vectores*, aunque estos últimos suelen limitarse a elementos numéricos.

Dados n universos de elementos $E_1, \ldots, E_n$, sea $\cup E_i = \{E_1 \cup \cdots \cup E_n\}$ la unión de los universos. Dependiendo de las características estructurales de las colecciones se pueden definir las siguientes operaciones lógicas:

Multiplicidad La *multiplicidad* o *frecuencia* de un elemento e en una colección A es el número de veces que el elemento x está contenido en A, es decir, el número de veces que se repite, y se denota $f_A(x)$.

En general $f_A(x) \in \mathbb{Z}_{\geq 0}$ aunque en el caso de colecciones con unicidad $f_A(x) \in \{0,1\}$.

Cardinalidad El *cardinal* de una colección A, que se denota $|A|$, es el número de elementos que contiene, es decir, $|A| = \sum_{x \in \cup E_i} f_A(x)$.

En el caso de colecciones con cardinalidad fija, todas las colecciones de ese tipo tendrán el mismo número de elementos.

Pertenencia Un elemento x *pertenece* a una colección A, y se denota $x \in A$, si y solo si $f_A(x) > 0$.

Inclusión Una colección A está *incluida* en otra B, y se denota $A \subseteq B$, si $f_A(x) \leq f_B(x)\ \forall x \in \cup E_i$.

Suma La *suma* de dos colecciones sin unicidad A y B, que se denota $A \uplus B$, es la colección formada por los elementos de A y B sumando sus multiplicidades, es decir, $f_{A \uplus B}(x) = f_A(x) + f_B(x)\ \forall x \in \cup E_i$.

Esta operación no tiene sentido para colecciones con unicidad.

Unión La *unión* de dos colecciones A y B, que se denota $A \cup B$, es la colección formada por los elementos de A o B con su máxima multiplicidad, es decir, $f_{A \cup B}(x) = \text{máx}\{f_A(x), f_B(x)\}$ $\forall x \in \cup E_i$.

Intersección La *intersección* de dos colecciones A y B, que se denota $A \cap B$, es el la colección formada por los elementos comunes de A y B con su mínima multiplicidad, es decir, $f_{A \cap B}(x) = \text{mín}\{f_A(x), f_B(x)\}\ \forall x \in \cup E_i$.

Diferencia La *diferencia* de dos colecciones A y B, que se denota $A - B$, es la colección que resulta de eliminar de A los elementos de B, teniendo en cuentas sus multiplicidades, es decir, $f_{A-B}(x) = \text{máx}\{f_A(x) - f_B(x), 0\}\ \forall x \in \cup E_i$.

Complementario La colección *complementaria* de una colección con unicidad A, que se denota $\overline{A}$, es la colección formada por los elementos de $\cup E_i$ que no pertenecen a A, es decir, $f_{\overline{A}}(x) = 1 - f_A(x)\ \forall x \in \cup E_i$.

Esta operación no tiene sentido para colecciones sin unicidad.

Elemento i-ésimo El *elemento i-ésimo* de una colección con orden A, $i \leq |A|$, es el elemento que ocupa la posición i, y se denota $A[i]$.

Esta operación no tiene sentido para colecciones sin orden.

Porción La *porción* de una colección con orden A desde la posición i hasta la posición j, $i \leq j \leq |A|$, que se denota $A[i,j]$, es la colección formada por los elementos de A que ocupan las posiciones desde la i hasta la j, es decir, $A[i,j] = [A[i], A[i+1], \ldots, A[j]]$.

Esta operación no tiene sentido para colecciones sin orden.

Concatenación La *concatenación* de las colecciones con orden A y B, que se denota $A + B$, es la lista formada por los elementos de A seguidos de los elementos de B, es decir, $A + B = [A[1], \ldots, a[|A|], B[1], \ldots, B[|B|]$.

Esta operación no tiene sentido para colecciones sin orden.

Subcolección Una colección con orden A es una *subcolección* de otra B, que se denota $A \sqsubseteq B$, si existe una secuencia creciente de enteros i_k, $k = 1, \ldots, |A|$, tal que $A[k] = B[i_k]$.

Esta operación no tiene sentido para colecciones sin orden.

Subcolección consecutiva Una colección con orden A es una *subcolección consecutiva* de otra B, que se denota $A \preceq B$ si existen colecciones con orden tales que $B = C + A + D$. Se cumple además que existen $i \leq j \leq |B|$ tales que $A = B[i, j]$.

Esta operación no tiene sentido para colecciones sin orden.

3. Comparación de colecciones

Al igual que cualquier otro proceso de comparación, la comparación de colecciones se realiza a través del estudio de las similitudes o diferencias entre los elementos que las componen por medio de funciones de similitud o disimilitud.

3.1. Funciones de similitud

Definición 1 (Función de similitud)**.** Dado un conjunto E de elementos, una *función de similitud* es una función $\sigma : E \times E \longrightarrow \mathbb{R}^+$ que asocia a cada par de elementos de E un número real que expresa el grado de parecido entre dichas elementos y que cumple los siguientes axiomas:

$$\text{No negatividad: } \forall x, y \in E,\ \sigma(x, y) \geq 0,$$
$$\text{Maximalidad: } \forall x, y \in E,\ \sigma(x, x) \geq \sigma(x, y).$$

Algunos autores añaden a estas propiedades la propiedad de *simetría*:

$$\forall x, y \in E, \sigma(x, y) = \sigma(y, x).$$

No obstante, tal y como refleja [32], existen situaciones en determinados contextos en las que esta propiedad no se cumple y por tanto no se ha incluido en la definición anterior.

El ejemplo más trivial de función de similitud es la función de similitud identidad, que sirve para cualquier tipo de dato.

Definición 2 (Función de similitud identidad)**.** Dado un universo E de elementos de cualquier tipo de datos, la *función de similitud identidad* es una función de similitud $\sigma_{id} : E \times E \longrightarrow [0, 1]$ tal que

$$\sigma_{id}(s, t) = \begin{cases} 1, & \text{si } s = t; \\ 0, & \text{si } s \neq t. \end{cases}$$

De igual modo que se define una función de similitud se puede definir una función de disimilitud:

Definición 3 (Función de disimilitud)**.** Dado un conjunto E de elementos, una *función de disimilitud* es una función $\delta : E \times E \longrightarrow \mathbb{R}^+$ que asocia a cada par de elementos de E un número real que expresa el grado de diferencia entre dichas elementos y que cumple los siguientes axiomas:

$$\text{No negatividad: } \forall x, y \in E,\ \delta(x, y) \geq 0,$$
$$\text{Minimalidad: } \forall x \in E,\ \delta(x, x) = 0.$$

En ocasiones se suelen utilizar definiciones más restringidas de funciones de disimilitud como son las distancias o métricas.

Definición 4 (Distancia)**.** Dado un conjunto E de elementos, una *distancia* es una función de disimilitud que además cumple los siguientes axiomas:

$$\text{Coincidencia: } \forall x, y \in E,\ \delta(x,y) = 0 \text{ si y solo si } x = y,$$
$$\text{Simetría: } \forall x, y \in E,\ \delta(x,y) = \delta(y,x),$$
$$\text{Desigualdad triangular: } \forall x, y, z \in E,\ \delta(x,y) + \delta(y,z) \geq \delta(x,z).$$

Habitualmente las funciones de similitud o disimilitud suelen normalizarse, restringiendo su rango al intervalo real $[0,1]$ para facilitar la comparación entre distintas medidas.

Definición 5 (Función de similitud normalizada)**.** Se dice que una función de (di)similitud está *normalizada* si su rango es el intervalo real $[0,1]$.

La mayoría de las funciones de similitud que se presentan en este artículo son funciones normalizadas, como por ejemplo la función de similitud identidad, aunque cuando pueda haber dudas, se utilizará la notación $\overline{\sigma}$ y $\overline{\delta}$ para distinguir las versiones normalizadas de las no normalizadas σ y δ respectivamente.

Muy a menudo, las funciones de similitud se construyen a partir de funciones de disimilitud o distancias.

Definición 6 (Correspondencia entre funciones de similitud y disimilitud)**.** Dada una función de similitud normalizada $\overline{\sigma} : E \times E \longrightarrow [0,1]$, y una función de disimilitud $\delta : E \times E \longrightarrow \mathbb{R}$, diremos que $\overline{\sigma}$ y δ se *corresponden*, si

$$\overline{\sigma}(x,y) = \pi(\delta(x,y)),$$

para algún isomorfismo $\pi : R \longrightarrow [0,1]$ que invierta el orden (monótono decreciente) y tal que $\pi(0) = 1$.

Algunos isomorfismos habituales para π son:

- $\pi(x) = 1 - x$, cuando δ está normalizada.
- $\pi(x) = 1 - \dfrac{x}{\max \delta}$, cuando δ tiene un valor máximo.
- $\pi(x) = 1 - \dfrac{x}{1+x}$, cuando δ no está acotada.

Uno de los principales problemas en un proceso de comparación es seleccionar las funciones de similitud apropiadas al contexto en el que se realizan las comparaciones. En el caso de la comparación de colecciones hay que tener en cuenta las propiedades estructurales del tipo de colecciones que se quieren comparar.

En las próximas secciones se presenta un catálogo de funciones de similitud o disimilitud para los tipos colecciones identificados en la sección 2. Las funciones de este catálogo se han organizado siguiendo la misma clasificación taxonómica de colecciones allí definida. De esta manera, es trivial determinar las funciones de similitud que pueden utilizarse para comparar dos colecciones sin más que saber el tipo al que pertenecen. Otra ventaja de la organización taxonómica es que se pueden comparar incluso colecciones de distintos tipos, utilizando las funciones de similitud de los tipos de colecciones que son ancestros comunes en la taxonomía de colecciones.

Sin duda, esta catalogación puede ayudar al usuario a seleccionar las funciones de similitud más apropiadas y también puede facilitar la definición de reglas de aplicación que automaticen el proceso de comparación, como por ejemplo la aplicación de funciones de similitud en cascada.

3.2. Catálogo de funciones de similitud

3.2.1. Funciones de similitud para multiheteroconjuntos

Los multiheteroconjuntos son las colecciones menos estructuradas ya que no tienen restricciones sobre la homogeneidad, la cardinalidad, el orden o la unicidad. Esto dificulta su comparación, especialmente cuando los elementos que los componen son incomparables, y por tanto, las funciones de similitud para compararlos son poco precisas. La medida de similitud más simple para multiheteroconjuntos surge de la comparación de sus cardinalidad.

Definición 7 (Función de similitud de cardinalidad)**.** Dados dos multiheteroconjuntos A y B, la *función de similitud de cardinalidad* es una función de similitud se define como

$$\sigma(A,B) = 1 - \frac{||A| - |B||}{\text{máx}\{|A|,|B|\}}.$$

Ejemplo 1 (Función de similitud de cardinalidad)**.** Dados los multiheteroconjuntos $A = \{1,1,2,2,3,4,a,a,b\}$ y $B = \{1,3,3,4,4,4,a,a,a,a,c\}$,

$$\sigma(A,B) = 1 - \frac{||A| - |B||}{\text{máx}\{|A|,|B|\}} = 1 - \frac{|9-11|}{\text{máx}\{9,11\}} = 1 - \frac{2}{11} = 0{,}82.$$

Esta función de similitud solo tiene en cuenta el número de elementos en los multiheteroconjuntos, pero no su contenido. Es posible precisar más la similitud teniendo en cuenta los elementos comunes y sus repeticiones en los multiheteroconjuntos. Con este enfoque, la función de similitud más intuitiva que surge es la propuesta en [17] y que resulta de comparar la intersección y la unión.

Definición 8 (Función de similitud de Jaccard)**.** Dados dos multiheteroconjuntos A y B, la *función de similitud de Jaccard* se define como

$$\sigma(A,B) = \frac{|A \cap B|}{|A \cup B|}.$$

Ejemplo 2 (Función de similitud de Jaccard)**.** Dados los multiheteroconjuntos $A = \{1,1,2,2,3,4,a,a,b\}$ y $B = \{1,3,3,4,4,4,a,a,a,a,c\}$,

$$\sigma(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|\{1,3,4,a,a\}|}{|\{1,1,2,2,3,3,4,4,4,a,a,a,a,b,c\}|}\frac{5}{15} = 0{,}33.$$

Otra función de similitud más optimista que utiliza este mismo enfoque se propone en [4], y compara la intersección con el mayor de los dos conjuntos, en lugar de con la unión.

Definición 9 (Función de similitud de Braun-Blanquet)**.** Dados dos multiheteroconjuntos A y B, la *función de similitud de Braun-Blanquet* se define como

$$\sigma(A,B) = \frac{|A \cap B|}{\text{máx}\{|A|,|B|\}}.$$

Ejemplo 3 (Función de similitud de Braun-Blanquet)**.** Dados los multiheteroconjuntos $A = \{1,1,2,2,3,4,a,a,b\}$ y $B = \{1,3,3,4,4,4,a,a,a,a,c\}$,

$$\sigma(A,B) = \frac{|A \cap B|}{\text{máx}\{|A|,|B|\}} = \frac{5}{11} = 0{,}45.$$

Más optimista aún es la función de similitud de [31], que compara la intersección con el menor de los dos conjuntos.

Definición 10 (Función de similitud de Simpson)**.** Dados dos multiheteroconjuntos A y B, la *función de similitud de Simpson* se define como

$$\sigma(A,B) = \frac{|A \cap B|}{\text{mín}\{|A|,|B|\}}.$$

Ejemplo 4 (Función de similitud de Simpson)**.** Dados los multiheteroconjuntos $A = \{1,1,2,2,3,4,a,a,b\}$ y $B = \{1,3,3,4,4,4,a,a,a,a,c\}$,

$$\sigma(A,B) = \frac{|A \cap B|}{\text{mín}\{|A|,|B|\}}. = \frac{5}{9} = 0{,}55.$$

Entre la medida de similitud de Braun-Blanquet y la de Simpson existe todo un repertorio de funciones de similitud que comparan el número de elementos de la intersección con un número entre la cardinalidad de A y la de B, como por ejemplo las funciones de similitud de [8], [27] y [23] que toman la media aritmética, geométrica y armónica del número de elementos de A y B, respectivamente.

Definición 11 (Función de similitud de Dice)**.** Dados dos multiheteroconjuntos A y B, la *función de similitud de Dice* se define como

$$\sigma(A,B) = \frac{|A \cap B|}{\frac{|A|+|B|}{2}} = \frac{2|A \cap B|}{|A|+|B|}.$$

Ejemplo 5 (Función de similitud de Dice)**.** Dados los multiheteroconjuntos $A = \{1,1,2,2,3,4,a,a,b\}$ y $B = \{1,3,3,4,4,4,a,a,a,a,c\}$,

$$\sigma(A,B) = \frac{2|A \cap B|}{|A|+|B|} = \frac{2 \cdot 5}{9+11} = 0{,}5.$$

Definición 12 (Función de similitud de Ochiai)**.** Dados dos multiheteroconjuntos A y B, la *función de similitud de Ochiai* se define como

$$\sigma(A,B) = \frac{|A \cap B|}{\sqrt{|A||B|}}.$$

Ejemplo 6 (Función de similitud de Ochiai)**.** Dados los multiheteroconjuntos $A = \{1,1,2,2,3,4,a,a,b\}$ y $B = \{1,3,3,4,4,4,a,a,a,a,c\}$,

$$\sigma(A,B) = \frac{|A \cap B|}{\sqrt{|A||B|}} = \frac{5}{\sqrt{9 \cdot 11}} = 0{,}5.$$

Definición 13 (Función de similitud de Kulczynski)**.** Dados dos multiheteroconjuntos A y B, la *función de similitud de Kulczynski* se define como

$$\sigma(A,B) = \frac{|A \cap B|}{\frac{2}{\frac{1}{|A|}+\frac{1}{|B|}}} = \frac{1}{2}\left(\frac{|A \cap B|}{|A|} + \frac{|A \cap B|}{|B|}\right).$$

Ejemplo 7 (Función de similitud de Kulczynski)**.** Dados los multiheteroconjuntos $A = \{1,1,2,2,3,4,a,a,b\}$ y $B = \{1,3,3,4,4,4,a,a,a,a,c\}$,

$$\sigma(A,B) = \frac{1}{2}\left(\frac{|A \cap B|}{|A|} + \frac{|A \cap B|}{|B|}\right) = \frac{1}{2}\left(\frac{5}{9} + \frac{5}{11}\right) = 0{,}5.$$

Otra función de similitud más general, utilizada habitualmente en el campo de la Recuperación de Información, que también puede extenderse fácilmente a la comparación de multiconjuntos es la medida F [34]. En el ámbito de la Recuperación de la Información, cuando se quiere comparar un conjunto de resultados obtenidos A con el conjunto de los resultados que debería obtenerse B, se utilizan habitualmente dos medidas conocidas como *precisión* y *exhaustividad*, que se definen de la siguiente manera

$$\text{Precisión } P = \frac{|A \cap B|}{|A|},$$
$$\text{Exhaustividad } E = \frac{|A \cap B|}{|B|},$$

y la medida F se define como la media armónica ponderada entre ellas

$$F = \frac{1}{\alpha \frac{1}{P} + (1-\alpha)\frac{1}{E}} = \frac{(1+\beta^2)P\,E}{\beta^2 P + E}, \text{con } \beta^2 = \frac{1-\alpha}{\alpha}$$

donde $\alpha \in [0,1]$ y $\beta^2 \in [0,\infty]$.

Su extensión a multiheteroconjuntos da lugar a la siguiente medida de similitud.

Definición 14 (Función de similitud F)**.** Dados dos multiheteroconjuntos A y B, la *función de similitud F* se define como

$$\sigma_\beta(A,B) = \frac{(1+\beta^2)|A \cap B|}{|A| + \beta^2|B|}.$$

donde $\beta \in \mathbb{R}^+$.

Esta función no es simétrica ya que para $\beta < 1$ se da más importancia a A y para $\beta > 1$ se da más importancia B. El único caso en que es simétrica es para $\beta = 1$ y en ese caso se obtiene la función de similitud de Dice.

Ejemplo 8 (Función de similitud F)**.** Dados los multiheteroconjuntos $A = \{1,1,2,2,3,4,a,a,b\}$ y $B = \{1,3,3,4,4,4,a,a,a,a,c\}$, y tomando $\beta = 0{,}5$, se tiene

$$\sigma_{0,5}(A,B) = \frac{(1+0{,}5^2)|A \cap B|}{|A| + 0{,}5^2|B|} = \frac{1{,}25 \cdot 5}{9 + 0{,}25 \cdot 11} = 0{,}53.$$

Mientras que tomando $\beta = 2$ se tiene

$$\sigma_2(A,B) = \frac{(1+2^2)|A \cap B|}{|A| + 2^2|B|} = \frac{5 \cdot 5}{9 + 4 \cdot 11} = 0{,}47.$$

En general, casi todas estas funciones de similitud se derivan de dos familias. La primera se debe a [5].

Definición 15 (Función de similitud de Caillez-Kuntz)**.** Dados dos multiheteroconjuntos A y B, la *función de similitud de Caillez-Kuntz* se define como

$$\sigma_\alpha(A,B) = \frac{|A \cap B|}{\sqrt[\alpha]{\frac{|A|^\alpha + |B|^\alpha}{2}}}.$$

donde $\alpha \in \mathbb{R}$.

De esta familia se obtienen algunas de las funciones anteriores para distintos valores de α. Por ejemplo, para $\alpha = -\infty$ se obtiene la función de similitud de Simpson, para $\alpha = -1$ la función de similitud de Kulczynski, para $\alpha = 0$ la función de similitud de Ochiai, para $\alpha = 1$ la función de similitud de Dice y para $\alpha = \infty$ la función de similitud de Braun-Blanquet.

La segunda familia de funciones se debe a [12].

Definición 16 (Función de similitud de Gower-Legendre)**.** Dados dos multiheteroconjuntos A y B, la *función de similitud de Gower-Legendre* se define como

$$\sigma_\eta(A,B) = \frac{\eta|A \cap B|}{|A| + |B| + (\eta - 2)|A \cap B|}.$$

donde $\eta \in \mathbb{R}$.

De esta familia también se obtienen algunas de las funciones anteriores para distintos valores de η, como por ejemplo la función de similitud de Jaccard para $\eta = 1$ o la función de similitud de Dice para $\eta = 2$.

Otra familia de funciones de similitud aún más general surge del modelo de ratio de Tversky [32], donde las similitudes entre dos objetos A y B dependen de las características comunes a A y B, las características de A que no tiene B y las características de B que no tiene A. Esto puede traducirse a multiheteroconjuntos como los elementos comunes a A y B, es decir $A \cap B$; los elementos en A que no están en B, es decir $A - B$; y los elementos en B que no están en A, $B - A$. Estos multiheteroconjuntos aparecen representados gráficamente en la figura 2.

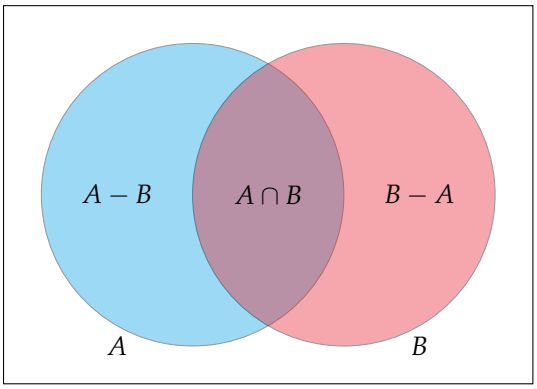


Figura 2. Multiheteroconjuntos involucrados en la comparación de dos multiheteroconjuntos A y B según el modelo de contraste de Tversky.

Definición 17 (Función de similitud de Tversky)**.** Dados dos multiheteroconjuntos A y B, la *función de similitud de Tversky* se define como

$$\sigma_{\beta,\gamma}(A,B) = \frac{|A \cap B|}{|A \cap B| + \beta|A - B| + \gamma|B - A|},$$

donde $\beta, \gamma \in \mathbb{R}^+$ son pesos que le dan mayor o menor importancia a los elementos no comunes de ambos multiheteroconjuntos.

Esta función de similitud no es simétrica si $\beta \neq \gamma$. Esto tiene sentido si, al comparar los multiheteroconjuntos uno de ellos es considerado como referente [32]. Cuando, por ejemplo, el segundo multiheteroconjunto se toma como referente, el peso de los elementos de B que no están en A, debe ser más grande que el peso de los elementos en A que no están B, de manera que $\beta < \gamma$.

Ejemplo 9 (Función de similitud de Tversky)**.** Dados los multiheteroconjuntos $A = \{1, 1, 2, 2, 3, 4, a, a, b\}$ y $B = \{1, 3, 3, 4, 4, 4, a, a, a, a, c\}$, y tomando $\beta = 1$ y $\gamma = 1$, se tiene

$$\sigma(A,B) = \frac{|A \cap B|}{|A \cap B| + |A - B| + |B - A|} = \frac{5}{5 + 4 + 6} = \frac{5}{15} = 0{,}33.$$

Si se toma como referente el segundo multiheteroconjunto, tomando por ejemplo $\beta = 0{,}25$ y $\gamma = 0{,}75$, entonces se tiene

$$\sigma(A,B) = \frac{|A \cap B|}{|A \cap B| + 0{,}25|A - B| + 0{,}75|B - A|} = \frac{5}{5 + 0{,}25 \cdot 4 + 0{,}75 \cdot 6} = 0{,}48.$$

que es diferente de la medida de similitud que obtiene al comparar B con A

$$\sigma(B,A) = \frac{|B \cap A|}{|B \cap A| + 0{,}25|B - A| + 0{,}75|A - B|} = \frac{5}{5 + 0{,}25 \cdot 6 + 0{,}75 \cdot 4} = 0{,}53.$$

Dependiendo de los valores de β y γ se obtienen diferentes funciones de similitud. Por ejemplo, para $\beta = \gamma = 1$ se obtienen la función de similitud de Jaccard, y para $\beta = \gamma = 1/2$, se obtiene la función de similitud de Dice. También se pueden deducir familias enteras de funciones de similitud como la función de similitud F, tomando $\beta = 1/(1 + \beta'^2)$ y $\gamma = \beta'^2/(1 + \beta'^2)$, donde β' es el parámetro asociado a esta Función, o en general la familia de funciones de similitud en las que se cumple $\beta + \gamma = 1$ [28].

Teniendo en cuenta las definiciones de estas familias de funciones de similitud, es posible definir otra familia más general aún que las engloba. Esta nueva función se ha llamado *función de similitud de Tversky generalizada*.

Definición 18 (Función de similitud de Tversky generalizada)**.** Dados dos multiheteroconjuntos A y B, la *función de similitud de Tversky generalizada* se define como

$$\sigma_{\alpha,\beta,\gamma}(A,B) = \frac{|A \cap B|}{\sqrt[\alpha]{\beta|A|^\alpha + \gamma|B|^\alpha + (1 - \beta - \gamma)|A \cap B|^\alpha}},$$

donde $\alpha \in \mathbb{R}$ y $\beta, \gamma \in \mathbb{R}^+$.

La demostración de que esta función de similitud engloba a las familias de funciones de similitud de Caillez-Kuntz, Gower-Legendre y Tverksy puede verse en [30].

En la tabla 2 se presenta un resumen de las funciones de similitud más comunes que resultan de instanciar la función de similitud de Tversky generalizada para diferentes valores de α, β y γ.

Disponer de una familia parametrizada de funciones de similitud no es sólo interesante desde el punto de vista de simplificar la implementación de todas ellas, sino también porque permite entender mejor la relación entre estas funciones de similitud, pero sobre todo porque permite generar nuevas funciones de similitud para distintas configuraciones de los parámetros. De hecho, esta nueva función de similitud abre la puerta a la experimentación para la determinación, mediante técnicas de aprendizaje automático, de la combinación de valores para los parámetros que dan mejores resultados en cada contexto de aplicación, disponiendo así de funciones de similitud a la carta.

3.2.2. Funciones de similitud para multiconjuntos

En primer lugar, puesto que los multiconjuntos son un caso particular de los multiheteroconjuntos, es posible utilizar cualquiera de las funciones de similitud vistas en la sección 3.2.1 para comparar multiconjuntos.

Por otro lado, puesto que los multiconjuntos pueden interpretarse como muestras con una distribución de frecuencias dada por las multiplicidades de sus elementos, es posible utilizar funciones de similitud para comparar distribuciones de frecuencias, como por ejemplo el test de la Chi-cuadrado [25].

Tabla 2. Funciones de similitud para multiheteroconjuntos derivadas de la de Tversky.

Similarity function	α	β	γ	**Formula**
Jaccard	1	1	1	$\frac{\lvert A \cap B \rvert}{\lvert A \cup B \rvert}$
Braun-Blanquet	$+\infty$	0,5	0,5	$\frac{\lvert A \cap B \rvert}{\max\{\lvert A \rvert, \lvert B \rvert\}}$
Simpson	$-\infty$	0,5	0,5	$\frac{\lvert A \cap B \rvert}{\min\{\lvert A \rvert, \lvert B \rvert\}}$
Dice	1	0,5	0,5	$\frac{2\lvert A \cap B \rvert}{\lvert A \rvert + \lvert B \rvert}$
Ochiai	0	0,5	0,5	$\frac{\lvert A \cap B \rvert}{\sqrt{\lvert A \rvert \lvert B \rvert}}$
Kulczynski	-1	0,5	0,5	$\frac{1}{2}\left(\frac{\lvert A \cap B \rvert}{\lvert A \rvert} + \frac{\lvert A \cap B \rvert}{\lvert B \rvert}\right)$
Sokal-Sneath	1	2	2	$\frac{\lvert A \cap B \rvert}{2\lvert A \rvert + \lvert 2\lvert B - 3\lvert A \cap B \rvert}$
F	1	$1/(1+\beta'^2)$	$\beta'^2/(1+\beta'^2)$	$\frac{(1+\beta'^2)\lvert A \cap B \rvert}{\lvert A \rvert + \beta'^2 \lvert B \rvert}$
Rodríguez-Egenhofer	1	$1-\gamma$	γ	$\frac{\lvert A \cap B \rvert}{\gamma\lvert A \rvert + (1-\gamma)\lvert B \rvert}$

Definición 19 (Función de similitud Chi-cuadrado)**.** Dados dos multiconjuntos A y B, la *función de similitud Chi-cuadrado* se define como

$$\sigma(A,B) = P(\chi^2(n) \geq \chi^2(A,B)),$$

donde

$$\chi^2(A,B) = \sum_{x \in A} \frac{(f_B(x) - f_A(x))^2}{f_A(x)},$$

es el estadístico Chi-cuadrado para los multiconjuntos A y B, y $\chi^2(n)$ es el valor de la función de distribución de una Chi-cuadrado con n grados de libertad, siendo n el número de elementos distintos de A menos 1.

Ejemplo 10 (Función de similitud Chi-cuadrado)**.** Dados los multiconjuntos $A = \{a,a,b,c,d,d,e\}$ y $B = \{a,a,a,a,b,b,c,d\}$, los cálculos del estadístico Chi-cuadrado aparecen en la siguiente tabla

x	$f_A(x)$	$f_B(x)$	$\frac{(f_B(x)-f_A(x))^2}{f_A(x)}$
a	2	4	2
b	1	2	1
c	1	1	0
d	2	1	0,5
e	1	0	1
$\sum$			4,5

con lo que se tiene

$$\sigma(A,B) = P(\chi^2(4) \geq 4{,}5) = 0{,}343,$$

Obsérvese que esta medida de similitud no es simétrica y además no puede calcularse cuando alguno de los elementos del segundo multiconjunto tiene multiplicidad 0. En tal caso podría utilizarse test exacto de Fisher.

Un enfoque diferente consiste en asignar un peso a cada elemento que depende no solo de la multiplicidad en el multiconjunto, sino también de su multiplicidad en otros multiconjuntos del mismo dominio de aplicación. Una función bien conocida que utiliza este enfoque es la Frecuencia de Términos × Frecuencia Inversa de Documentos, más conocida como TF×IDF (del inglés Term Frequency × Inverse Document Frequency). Esta función se aplica habitualmente en el modelo de espacio de vectores para comparar documentos, vistos como multiconjuntos de palabras [29]. La idea consiste en cuantificar la importancia de un término en un documento teniendo en cuenta su frecuencia dentro del documento y fuera de él en un corpus de documentos de referencia. La importancia del término aumenta a medida que su frecuencia aumenta en el documento, y disminuye a medida que su frecuencia aumenta en el corpus.

Definición 20 (TF×IDF)**.** Dados un multiconjunto A y un corpus de multiconjuntos $\mathcal{C}$ en el mismo universo, se define la función

$$\text{TF} \times \text{IDF}_{A,\mathcal{C}}(x) = f_A(x) \log(|\mathcal{C}|/|\mathcal{C}(x)|),$$

donde $\mathcal{C}(x)$ es el subconjunto de multiconjuntos de $\mathcal{C}$ que contienen el término x.

La base del logaritmo no es importante aunque habitualmente suele tomarse 2.

Para definir la función de similitud para multiconjuntos basada en la función TF×IDF, un multiconjunto A es considerado como un vector donde cada dimensión corresponde a un elemento del universo de los multiconjuntos $E = \{x_1, \ldots, x_n\}$, y la componente i es el peso del elemento x_i en el multiconjunto, es decir, $\text{TF}\times\text{IDF}_{A,\mathcal{C}}(x_i)$. Con estos vectores, la similitud entre multiconjuntos puede calcularse con cualquiera de las distancias en el espacio Euclideo n dimensional, como por ejemplo, la distancia Euclidea, el producto escalar o el coseno (ver sección 3.2.14).

Definición 21 (Función de similitud TF×IDF)**.** Dados dos multiconjuntos A y B, y un corpus de multiconjuntos $\mathcal{C}$ en el mismo universo, la *función de similitud TF×IDF* se define como

$$\sigma_{\mathcal{C}}(A,B) = \frac{\sum_{i=1}^{n} \text{TF} \times \text{IDF}_{A,\mathcal{C}}(x_i) \cdot \text{TF} \times \text{IDF}_{B,\mathcal{C}}(x_i)}{\sqrt{\sum_{i=1}^{n} (\text{TF} \times \text{IDF}_{A,\mathcal{C}}(x_i))^2 \cdot \sum_{i=1}^{n} (\text{TF} \times \text{IDF}_{B,\mathcal{C}}(x_i))^2}}$$

3.2.3. Funciones de similitud para heteroconjuntos

Los heteroconjuntos son un caso particular de multiheteroconjuntos, de manera que pueden utilizarse todas las funciones de similitud de la sección 3.2.1 para comparar heteroconjuntos. Más allá de estas funciones, en la literatura científica no se han encontrado otras funciones que utilicen además la condición de unicidad específica de los heteroconjuntos.

3.2.4. Funciones de similitud para listas

Las listas son un caso particular de multiheteroconjuntos, de manera que pueden utilizarse todas las funciones de similitud de la sección 3.2.1 para comparar listas.

Por otro lado, la restricción de orden que incorporan las listas le da mayor semántica a la colección y puede aprovecharse en distintas estrategias de comparación de listas. La estrategia más sencilla consiste en contar el número de elementos que coinciden en la misma posición de dos listas y fue propuesta en [14] para la detección de errores en la transmisión de cadenas de bits.

Definición 22 (Función de similitud de Hamming para listas)**.** Dadas dos listas A y B, la *función de similitud de Hamming* se define como

$$\sigma(A,B) = \frac{\sum_{i=1}^{\text{mín}\{|A|,|B|\}} \sigma_{id}(A[i],B[i])}{\text{máx}\{|A|,|B|\}},$$

donde σ_{id} es la función de similitud identidad.

Ejemplo 11 (Función de similitud de Hamming)**.** Dadas las las listas $A = [a,b,c,a,d,c,a,b]$ y $B = [a,d,e,a,f,c,a,d,c,a,b]$, los elementos que coinciden en la misma posición aparecen sombreados en la siguiente tabla:

Posición	1	2	3	4	5	6	7	8	9	10	11
A	a	b	c	a	d	c	a	b			
B	a	d	e	a	f	c	a	d	c	a	b

Por tanto, la similitud de Hamming de A y B es

$$\sigma(A,B) = \frac{\sum_{i=1}^{\text{mín}\{|A|,|B|\}} \sigma_{id}(A[i],B[i])}{\text{máx}\{|A|,|B|\}} = \frac{1+0+0+1+0+1+1+0}{\text{máx}\{8,11\}} = \frac{4}{11} = 0{,}37.$$

Esta función solo tiene en cuenta las coincidencias de elementos en la misma posición. Sin embargo, es posible considerar las coincidencias de elementos dentro de una ventana de un determinado tamaño.

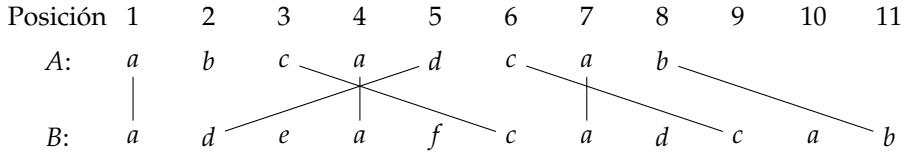
Definición 23 (Ventana de coincidencia)**.** Dadas dos listas A y B, la *ventana de coincidencia* de tamaño k, que se denota $\text{match}(A,B,k)$, es el conjunto de pares de posiciones (i,j), $1 \leq i \leq |A|$ y $1 \leq j \leq |B|$, tales que

- $A[i] = B[j]$,
- $i - k \leq j \leq i + k$,
- $\nexists l < j$ con $(i,l) \in \text{match}(A,B,k)$,
- $\nexists m < i$ con $(m,j) \in \text{match}(A,B,k)$.

Los elementos correspondientes a las posiciones de $\text{match}(A,B,k)$ son una sublista de A que se llama *sublista de la ventana de coincidencia de tamaño k* y es

$$\text{lmatch}(A,B,k) = \{A[i] \mid \exists j \text{ tal que } (i,j) \in \text{match}(A,B,k)\}$$

Ejemplo 12 (Ventana de coincidencia)**.** Dadas las listas $A = [a,b,c,a,d,c,a,b]$ y $B = [a,d,e,a,f,c,a,d,c,a,b]$, para una ventana de coincidencia de tamaño 4, se tienen las siguientes coincidencias



de manera que

$$\text{match}(A,B,4) = \{(1,1),(3,6),(4,4),(5,2),(6,9),(7,7),(8,11)\}$$

Del mismo modo se puede comprobar que

$$\text{match}(B,A,4) = \{(1,1),(2,5),(4,4),(6,3),(7,7),(9,6),(11,8)\}$$

Obsérvese $B[8] = d$ no coincide con $A[5]$ ya que esta posición fue previamente emparejada con $B[2]$.

Utilizando la ventana de coincidencia se puede definir la siguiente función de similitud.

Definición 24 (Función de similitud de la ventana de coincidencia)**.** Dadas dos listas A y B, la *función de similitud de ventana de coincidencia* de tamaño k se define como

$$\sigma_k(A,B) = \frac{|\operatorname{match}(A,B,k)|}{\max\{|A|,|B|\}}.$$

Ejemplo 13 (Función de similitud de la ventana de coincidencia)**.** Según las ventanas de coincidencia calculadas en el ejemplo anterior, la similitud de ventana de coincidencia de tamaño 4 para las listas $A = [a,b,c,a,d,c,a,b]$ y $B = [a,d,e,a,f,c,a,d,c,a,b]$ es

$$\sigma_4(A,B) = \frac{|\operatorname{match}(A,B,4)|}{\max\{|A|,|B|\}} = \frac{7}{11} = 0{,}64.$$

Una variación de esta función de similitud que también tiene en cuenta las transposiciones de elementos en la ventana de coincidencia se debe a [18].

Definición 25 (Función de similitud de Jaro)**.** Dadas dos listas A y B, la *función de similitud de Jaro* se define como

$$\sigma(A,B) = \frac{1}{3}\left(\frac{|\operatorname{match}(A,B)|}{|A|} + \frac{|\operatorname{match}(A,B)|}{|B|} + \frac{|\operatorname{match}(A,B)| - |\operatorname{lmatch}(A,B)|}{|\operatorname{match}(A,B)|}\right)$$

donde $\operatorname{match}(A,B) = \operatorname{match}(A,B,\frac{\max\{|A|,|B|\}}{2} - 1)$ (es decir, el tamaño de la ventana de coincidencia es la mitad del tamaño de la mayor lista menos uno) y $\operatorname{lmatch}(A,B)$ es el número de elementos en $\operatorname{lmatch}(A,B,\frac{\max\{|A|,|B|\}}{2} - 1)$ que no coinciden con los elementos de $\operatorname{lmatch}(B,A,\frac{\max\{|A|,|B|\}}{2} - 1)$ en la misma posición, dividido por 2.

Ejemplo 14 (Función de similitud de Jaro)**.** Considerando de nuevo las listas $A = [a,b,c,a,d,c,a,b]$ y $B = [a,d,e,a,f,c,a,d,c,a,b]$, el tamaño de la ventana de coincidencia es $\frac{\max\{|A|,|B|\}}{2} - 1 = 4$, que es el mismo tamaño de la ventana de coincidencia del ejemplo anterior. Teniendo en cuenta además los elementos que no aparecen en la misma posición de las sublistas correspondientes a las ventanas de coincidencia, que aparecen sombreadas en la siguiente tabla

$\operatorname{lmatch}(A,B,4)$	a	c	a	d	c	a	b
$\operatorname{lmatch}(B,A,4)$	a	d	a	c	a	c	b

se concluye que la similitud de Jaro es

$$\begin{aligned}\sigma(A,B) &= \frac{1}{3}\left(\frac{|\operatorname{match}(A,B)|}{|P|} + \frac{|\operatorname{match}(A,B)|}{|B|} + \frac{|\operatorname{match}(A,B)| - |\operatorname{lmatch}(A,B)|}{|\operatorname{match}(A,B)|}\right) = \\ &= \frac{1}{3}\left(\frac{7}{8} + \frac{7}{11} + \frac{7-2}{7}\right) = 0{,}74.\end{aligned}$$

Cuando se aplica a palabras de un lenguaje, se ha comprobado que esta función de similitud funciona muy bien para identificar errores ortográficos.

Otra variación de esta función de similitud que da mayor peso a los prefijos comunes se debe a [37].

Definición 26 (Función de similitud de Jaro-Winkler)**.** Dadas dos listas A y B, la *función de similitud de Jaro-Winkler* se define como

$$\sigma(A,B) = \sigma_J(A,B) + \alpha|\operatorname{prefix}(A,B)|(1-\sigma_J(A,B)),$$

donde $\sigma_J(s,t)$ es la función de similitud de Jaro, $\operatorname{prefix}(A,B)$ es la sublista consecutiva de mayor tamaño que es prefijo común a A y B, hasta un máximo de 4, y α es un factor de escalado (normalmente 0,1).

Ejemplo 15 (Función de similitud de Jaro-Winkler)**.** Considerando de nuevo las listas $A = [a,b,c,a,d,c,a,b]$ y $B = [a,d,e,a,f,c,a,d,c,a,b]$, su similitud de Jaro-Winkler es

$$\begin{aligned}\sigma(A,B) &= \sigma_J(A,B) + 0{,}1|\operatorname{prefix}(A,B)|(1-\sigma_J(A,B)) = \\ &= 0{,}74 + 0{,}1 \cdot 1 \cdot (1-0{,}74) = 0{,}77.\end{aligned}$$

Aunque las funciones de similitud de Jaro y de Jaro-Winkler son válidas para listas, en realidad provienen del área de la vinculación de registros y se utilizan fundamentalmente con cadenas de caracteres, que son un caso particular de secuencias (ver sección 3.2.7).

Otra forma natural de comparar listas es por medio de sus sublistas, especialmente cuando el orden es tan importante como para no permitir transposiciones. La comparación se hace comparando el tamaño de la mayor sublista común.

Definición 27 (Función de similitud de la mayor sublista común)**.** Dadas dos listas A y B, la *función de similitud de la mayor sublista común* se define como

$$\sigma(A,B) = \frac{2\,\text{máx}\{|C| : C \sqsubseteq A, C \sqsubseteq B\}}{|A|+|B|}$$

Ejemplo 16 (Función de similitud de la mayor sublista común)**.** La mayor sublista común de las listas $A = [a,b,c,a,d,c,a,b]$ y $B = [a,d,e,a,f,c,a,d,c,a,b]$ es $[a,c,a,d,c,a,b]$, y por tanto, la similitud de la mayor sublista común es

$$\sigma(A,B) = \frac{2|[a,c,a,d,c,a,b]|}{|A|+|B|} = \frac{2\cdot 7}{8+11} = 0{,}74.$$

También se pueden considerar sublistas consecutivas en lugar de sublistas.

Definición 28 (Función de similitud de la mayor sublista consecutiva común)**.** Dadas dos listas A y B, la *función de similitud de la mayor sublista consecutiva común* se define como

$$\sigma(A,B) = \frac{2\,\text{máx}\{|C| : C \preceq A, C \preceq B\}}{|A|+|B|}$$

Ejemplo 17 (Función de similitud de la mayor sublista consecutiva común)**.** La mayor sublista consecutiva común de las listas $A = [a,b,c,a,d,c,a,b]$ y $B = [a,d,e,a,f,c,a,d,c,a,b]$ es $[c,a,d,c,a,b]$, y por tanto, la similitud de la mayor sublista consecutiva común es

$$\sigma(A,B) = \frac{2|[c,a,d,c,a,b]|}{|A|+|B|} = \frac{2\cdot 6}{8+11} = 0{,}63.$$

Se pueden considerar también variaciones de esta función de similitud que solamente tengan en cuenta los prefijos y los sufijos de las listas.

Definición 29. Función de similitud del mayor prefijo común Dadas dos listas A y B, la *función de similitud del mayor prefijo común* se define como

$$\sigma(A,B) = \frac{2\,\text{máx}\{|C| : A = C + A', B = C + B'\}}{|A|+|B|}$$

Ejemplo 18 (Función de similitud del mayor prefijo común)**.** El mayor prefijo común de las listas $A = [a, b, c, a, d, c, a, b]$ y $B = [a, d, e, a, f, c, a, d, c, a, b]$ es $[a]$, y por tanto, la similitud del mayor prefijo común es

$$\sigma(A, B) = \frac{2 \cdot |[a]|}{|A| + |B|} = \frac{2 \cdot 1}{8 + 11} = 0{,}11.$$

Definición 30 (Función de similitud del mayor sufijo común)**.** Dadas dos listas A y B, la *función de similitud del mayor sufijo común* se define como

$$\sigma(A, B) = \frac{2 \max\{|C| : A = A' + C, B = B' + C\}}{|A| + |B|}$$

Ejemplo 19 (Función de similitud del mayor sufijo común)**.** El mayor sufijo común de las listas $A = [a, b, c, a, d, c, a, b]$ y $B = [a, d, e, a, f, c, a, d, c, a, b]$ es $[c, a, d, c, a, b]$, y por tanto, la similitud del mayor sufijo común es

$$\sigma(A, B) = \frac{2|[c, a, d, c, a, b]|}{|A| + |B|} = \frac{2 \cdot 6}{8 + 11} = 0{,}63.$$

Por supuesto también se pueden idear funciones de similitud que combinen prefijos y sufijos.

3.2.5. Funciones de similitud para cajas

Las cajas son un caso particular de multiheteroconjuntos, de manera que pueden utilizarse todas las funciones de similitud de la sección 3.2.1 para comparar cajas. Más allá de estas funciones, en la literatura científica no se han encontrado otras funciones que utilicen además la condición de cardinalidad fija específica de las cajas.

3.2.6. Funciones de similitud para conjuntos

Los conjuntos son un caso particular de multiconjuntos y de heteroconjuntos, de manera que pueden utilizarse todas las funciones de similitud de las secciones 3.2.2 y 3.2.3 para comparar conjuntos. Las funciones de similitud definidas en esas secciones solamente tienen en cuenta los elementos comunes y no comunes en los conjuntos; sin embargo, cuando los elementos son diferentes no se tienen en cuenta sus diferencias. Si existe una función de similitud o disimilitud para los elementos del universo, es posible calcular la similitud entre dos conjuntos como una función de la similitud entre sus elementos. Existen varias alternativas; la más optimista es tomar los dos elementos más similares.

Definición 31 (Función de similitud de elementos máxima)**.** Dados dos conjuntos A y B definidos sobre un mismo universo E y una función de similitud normalizada σ_E para los elementos de E, la *función de similitud de elementos máxima* se define como

$$\sigma(A, B) = \max\{\sigma_E(x, y) : x \in A, y \in B\}.$$

La opción más pesimista consiste en tomar los dos elementos menos similares

Definición 32 (Función de similitud de elementos mínima)**.** Dados dos conjuntos A y B definidos sobre un mismo universo E y una función de similitud normalizada σ_E para los elementos de E, la *función de similitud de elementos mínima* se define como

$$\sigma(A, B) = \min\{\sigma_E(x, y) : x \in A, y \in B\}.$$

El problema de las opciones más optimista y más pesimista es que solamente tienen en cuenta un par de elementos, los más similares y los menos, respectivamente. En muchos casos es más razonable considerar la similitud entre todos los posibles pares.

Definición 33 (Función de similitud de elementos media)**.** Dados dos conjuntos A y B definidos sobre un mismo universo E y una función de similitud normalizada σ_E para los elementos de E, la *función de similitud de elementos media* se define como

$$\sigma(A,B) = \frac{\sum_{(x,y)\in A\times B}\sigma_E(x,y)}{|A||B|}.$$

Ejemplo 20 (Función de similitud de elementos máxima, mínima y media)**.** Dados los conjuntos A={1,2,5,6,8} y B={1,3,4,7,8,9} sobre el universo de los números enteros del 0 al 10 sobre los que se define la función de similitud normalizada $\sigma_E(x,y) = 1 - \frac{|x-y|}{10}$. La similitud de todos los posibles pares de elementos de A y B de acuerdo a esta medida de similitud aparece en la tabla **??**. Según estas similitudes la similitud de elementos máxima de A y B es

$$\sigma(A,B) = \text{máx}\{\sigma_E(x,y) : x \in A, y \in B\} = 1,$$

ya que ambos comparten elementos comunes y la similitud entre dos elementos iguales es 1.

La similitud de elementos mínima es

$$\sigma(A,B) = \text{máx}\{\sigma_E(x,y) : x \in A, y \in B\} = 0{,}87,$$

que corresponde a la similitud entre 1 y 9 que son los más distintos.

Y la similitud de elementos media entre las cartas de vinos de los restaurantes A y B es

$$\sigma(A,B) = \frac{\sum_{(x,y)\in A\times B}\sigma_E(x,y)}{|A||B|} = \frac{61{,}44}{8\cdot 8} = 0{,}67.$$

Tabla 3. Medidas de similitud entre los elementos de los conjuntos A={1,2,5,6,8} y B={1,3,4,7,8,9} mediante la función de similitud $\sigma_E(x,y) = 1 - \frac{|x-y|}{10}$.

$A\backslash B$	1	3	4	7	8	9	$\sum$
1	1	0,8	0,7	0,4	0,3	0,2	3,4
2	0,9	0,9	0,8	0,5	0,4	0,3	3,8
5	0,6	0,8	0,9	0,8	0,7	0,6	4,4
6	0,5	0,7	0,8	0,9	0,8	0,7	4,4
8	0,3	0,5	0,6	0,9	1	0,9	4,2
						$\sum$	20,2

Sin embargo, la función de similitud de elementos media, que parece la más razonable, no es en realidad una función de similitud ya que no cumple la condición de maximalidad. Según [33], para satisfacer la condición de maximalidad es necesario emparejar los elementos y medir la similitud de los elementos emparejados. Para que el emparejamiento sea óptimo debe cumplirse:

a) Cada elemento de A debe emparejarse a lo sumo con uno de B y viceversa.
b) El número de elementos emparejados debe ser máximo.
c) El emparejamiento debe hacer la medida de similitud máxima.

Definición 34 (Función de similitud de emparejamiento óptimo)**.** Dados dos conjuntos A y B definidos sobre un mismo universo E y una función de similitud normalizada σ_E para los elementos de E, la *función de similitud de elementos media* se define como

$$\sigma(A,B) = \frac{\sum_{(x,y)\in \text{match}(A,B)} \sigma_E(x,y)}{\text{máx}\{|A|,|B|\}},$$

donde $\text{match}(A,B)$ es un emparejamiento óptimo de los elementos de A y B.

Ejemplo 21 (Función de similitud de emparejamiento óptimo)**.** Siguiendo con el ejemplo anterior, es fácil ver que el emparejamiento óptimo entre elementos de A y B es el que aparece en la tabla 4. Por tanto la similitud de elementos media de A y B es

$$\sigma(A,B) = \frac{\sum_{(x,y)\in \text{match}(A,B)} \sigma_E(x,y)}{\text{máx}\{|A|,|B|\}} = \frac{4{,}7}{6} = 0{,}78.$$

Tabla 4. Emparejamiento óptimo de los elementos de los conjuntos A={1,2,5,6,8} y B={1,3,4,7,8,9} de acuerdo a la función de similitud $\sigma_E(x,y) = 1 - \frac{|x-y|}{10}$.

$x \in A$	$y \in B$	$\sigma_E(x,y)$
1	1	1
2	3	0,9
5	4	0,9
6	7	0,9
8	8	1
	$\sum$	4,7

3.2.7. Funciones de similitud para secuencias

Las secuencias son un caso particular de multiconjuntos y de listas, de manera que pueden utilizarse todas las funciones de similitud de las secciones 3.2.2 y 3.2.4 para comparar secuencias.

Como en el caso de las listas, la estrategia más trivial para comparar secuencias es comparar los elementos que ocupan la misma posición. Si solo se tiene en cuenta las coincidencias, entonces se tiene la función de similitud de Hamming. Sin embargo, cuando los elementos no coinciden, esta función no mide las discrepancias. Puesto que los elementos de las secuencias son todos del mismo tipo, y por tanto son comparables, es posible extender la función de similitud de Hamming para que tenga en cuenta la similitud de los elementos en la misma posición.

Definición 35 (Función de similitud de Hamming extendida)**.** Dadas dos secuencias A y B definidas sobre un mismo universo E y una función de similitud normalizada σ_E para los elementos de E, la *función de similitud de Hamming extendida* se define como

$$\sigma(A,B) = \frac{\sum_{i=1}^{\text{mín}\{|A|,|B|\}} \sigma_E(A[i],B[i])}{\text{máx}\{|A|,|B|\}}.$$

Otra forma de comparar secuencias es comparando subsecuencias suyas. Es habitual considerar subsecuencias consecutivas de un tamaño fijo que se conocen como n-gramas [22].

Definición 36 (n-grama)**.** Un *n-grama* de una secuencia A es cualquier subsecuencia consecutiva de A de cardinalidad n.

Cualquier secuencia A tiene exactamente $|A| - n + 1$ n-gramas, que son

$$n\text{-grams}(A) = \{A[1, n], A[2, n + 1], \cdots, A[|A| - n + 1, |A|]\}$$

Ejemplo 22 (n-gramas)**.** La secuencia correspondiente a la palabra, `similitud` tiene 6 4-gramas, que son

$$\text{4-gramas}(\mathsf{similitud}) = \{\mathsf{simi}, \mathsf{imil}, \mathsf{mili}, \mathsf{ilit}, \mathsf{litu}, \mathsf{itud}\}$$

La colección de todos los n-gramas de una secuencia es en realidad un multiconjunto, y por tanto, es posible utilizar cualquiera de las funciones de similitud presentadas en la sección 3.2.2 para compararlos, como por ejemplo, la función de similitud de Jaccard (ver definición **??**).

Definición 37 (Función de similitud de n-gramas)**.** Dadas dos secuencias A y B, la *función de similitud de n-gramas* se define como

$$\sigma(A, B) = \frac{|n\text{-gramas}(A) \cap n\text{-gramas}(B)|}{|n\text{-gramas}(A) \cup n\text{-gramas}(B)|}$$

Ejemplo 23 (Función de similitud de n-gramas)**.** Considerando las secuencias de nucleótidos $A = [g, a, c, c, a, a, c, a, t, t]$ y $B = [a, t, g, a, c, c, a, a, c, a]$ del gen del citocromo b en el perro y el gato respectivamente, y teniendo en cuenta que los nucleótidos suelen agruparse en bloques de 3, llamados *codones* que dan lugar a los aminoácidos que forman las proteínas, tiene sentido comparar las secuencias utilizando 3-gramas. Los $3-$*gramas* correspondientes a las secuencias anteriores son

$$\begin{aligned}
\text{3-gramas}(P)= &\quad \{\texttt{gac}, \texttt{acc}, \texttt{cca}, \texttt{caa}, \texttt{aac}, \texttt{aca}, \texttt{cat}, \texttt{att}\}\\
\text{3-gramas}(G)= &\quad \{\texttt{atg}, \texttt{tga}, \texttt{gac}, \texttt{acc}, \texttt{cca}, \texttt{caa}, \texttt{aac}, \texttt{aca}\}\\
\text{3-gramas}(P) \cap \text{3-gramas}(G)= &\quad \{\texttt{gac}, \texttt{acc}, \texttt{cca}, \texttt{caa}, \texttt{aac}, \texttt{aca}\}\\
\text{3-gramas}(P) \cup \text{3-gramas}(G)= &\quad \{\texttt{gac}, \texttt{acc}, \texttt{cca}, \texttt{caa}, \texttt{aac}, \texttt{aca}, \texttt{cat}, \texttt{att}, \texttt{atg}, \texttt{tga}\}
\end{aligned}$$

y la similitud de 3-gramas de las secuencias genéticas del citocromo b del perro y el gato es

$$\sigma(P, G) = \frac{|\text{3-gramas}(P) \cap \text{3-gramas}(G)|}{|\text{3-gramas}(P) \cup \text{3-gramas}(G)|} = \frac{6}{10} = 0{,}6.$$

Otra forma común de comparar secuencias es medir el número de operaciones elementales que se requieren para transformar una secuencia en la otra [13]. Una famosa distancia que utiliza este enfoque es la distancia de edición de [24], que solamente considera tres posibles operaciones: inserción, eliminación y sustitución de un elemento.

Definición 38 (Función de similitud de Levenshtein)**.** Dadas dos secuencias A y B, la *distancia de edición de Levenshtein* se define como

$$\delta(A, B) = \frac{\text{mín}\{n : B = op_1 \circ \cdots \circ op_n(A)\}}{\text{máx}\{|A|, |B|\}},$$

donde op_i es una de las siguientes operaciones elementales :

- $\text{ins}(A, i, e)$ insertar el elemento e en la posición i- de la secuencia A y desplazar el resto de elementos.
- $\text{del}(A, i)$ eliminar el elemento de las posición i-th de la secuencia A.
- $\text{sub}(A, i, e)$ sustituir el elemento de la posición i de la secuencia A por el elemento e.

La correspondiente función de similitud $\sigma(A,B) = 1 - \delta(A,B)$ se conoce como *función de similitud de Levenshtein*.

El algoritmo para calcular el mínimo número de operaciones elementales necesarias para convertir una secuencia en la otra es un algoritmo de programación dinámica bastante conocido [6].

Ejemplo 24 (Función de similitud de Levenshtein)**.** Considerando de nuevo las secuencias de nucleótidos A y B del ejemplo anterior correspondientes al gen del citocromo b en el perro y el gato, se necesitan un mínimo de 4 operaciones elementales para transformar la cadena `gaccaacatt` en `atgaccaaca`, que son

$$\begin{aligned} \text{ins}(\texttt{gaccaacatt}, 1, \texttt{a}) &= \texttt{agaccaacatt} \\ \text{ins}(\texttt{agaccaacatt}, 2, \texttt{t}) &= \texttt{atgaccaacatt} \\ \text{del}(\texttt{atgaccaacatt}, 12) &= \texttt{atgaccaacat} \\ \text{del}(\texttt{atgaccaacat}, 11) &= \texttt{atgaccaaca} \end{aligned}$$

Por lo tanto, la distancia de edición de Levenshtein entre las secuencias genéticas del citocromo b del perro y el gato es

$$\delta(A,B) = \frac{4}{10} = 0{,}4,$$

y la función de similitud asociada es

$$\sigma(A,B) = 1 - \frac{4}{10} = 0{,}6.$$

Esta función de similitud asocia el mismo peso a las tres operaciones elementales aunque es posible asignar distintos pesos a cada operación como ocurre con la distancia de Needleman-Wunch, que asigna más peso a las inserciones y las eliminaciones que a las sustituciones [26].

3.2.8. Funciones de similitud para multicombinaciones

Las multicombinaciones son casos particulares de multiconjuntos en los que se ha fijado la cardinalidad, y también de cajas en las que todos los elementos pertenecen al mismo universo, de manera que pueden utilizarse todas las funciones de similitud de las secciones 3.2.2 y 3.2.5 para comparar multicombinaciones. Más allá de estas funciones, en la literatura científica no se han encontrado otras funciones que utilicen la condición de cardinalidad fija o de homogeneidad.

3.2.9. Funciones de similitud para heterorankings

Los heterorankings son casos particulares de heteroconjuntos en los que se ha establecido un orden, y también de listas en las que no se pueden repetir elementos, de manera que pueden utilizarse todas las funciones de similitud de las secciones 3.2.3 y 3.2.4 para comparar heterorankings.

Como en los heterorankings no se pueden repetir los elementos, para compararlos se pueden utilizar técnicas para comparar permutaciones, adaptándolas para cuando las colecciones no son del mismo tamaño y no contienen exactamente los mismos elementos. Las tres medidas más conocidas para comparar permutaciones son la distancia de Spearman, la distancia Rho de Spearman y la distancia Tau de Kendall [19].

La distancia de Spearman entre dos permutaciones ρ_A y ρ_B de n elementos es la distancia de Manhattan (ver definición 31) entre los rangos o posiciones de los elementos en las dos permutaciones, es decir,

$$\delta(\rho_A, \rho_B) = \sum_{i=1}^{n} |\rho_A(i) - \rho_B(i)|.$$

Esta distancia está acotada ya que su valor máximo es $n^2/2$ si n es par, y $(n+1)(n-1)/2$ si n es impar, de manera que puede normalizarse fácilmente dividiéndola por estos valores.

La distancia Rho de Spearman entre dos permutaciones ρ_A y ρ_B de n elementos es la distancia euclídea (ver definición 32) entre los rangos o posiciones de los elementos en las dos permutaciones, es decir,

$$\delta(\rho_A, \rho_B) = \left(\sum_{i=1}^{n} (\rho_A(i) - \rho_B(i))^2\right)^{1/2}.$$

Esta distancia está acotada ya que su valor máximo es $(n(n+1)(2n+1)/3)^{1/2}$, de manera que puede normalizarse fácilmente dividiéndola por este valor.

La Tau de Kendall mide el número de pares de elementos que aparecen en distinto orden en las dos permutaciones, es decir,

$$\begin{aligned}\delta(\rho_A, \rho_B) = |\{(i,j) : i < j, (\rho_A(i) < \rho_A(j) \wedge \rho_B(i) > \rho_B(j)) \vee \\ \vee (\rho_A(i) > \rho_A(j) \wedge \rho_B(i) < \rho_B(j))\}|.\end{aligned}$$

Esta distancia también está acotada y su valor máximo es $n(n-1)/2$, por lo que también se puede normalizar fácilmente.

En [10] se proponen varias extensiones de ambas distancias para la comparación de rankings, cuando los rankings pueden contener distintos elementos. No obstante, estas extensiones siguen considerando rankings del mismo tamaño, es decir, heterovariaciones (ver sección 3.2.15).

A continuación se presentan varias extensiones de estas distancias para heterorankings, que son novedosas. Si se tienen dos heterorankings A y B, y $k = |A \cap B|$ es el número de elementos comunes, en el caso de la distancia de Spearman una extensión natural consiste en comparar las dos permutaciones que contienen los elementos comunes, añadiendo los elementos de $A - B$ en la posición $|B| + 1$ de B, y los elementos de $B - A$ en la posición $|A| + 1$ de A.

Definición 39 (Distancia de Spearman para heterorankings)**.** Dados dos heterorankings A y B, la *distancia de Spearman para heterorankings* se define como

$$\begin{aligned}\delta(A,B) &= \sum_{i\in A\cap B} |\rho_A(i) - \rho_B(i)| + \sum_{i\in A-B} ((|B|+1) - \rho_A(i)) + \sum_{i\in B-A} ((|A|+1) - \rho_B(i)) = \\ &= \sum_{i\in A\cap B} |\rho_A(i) - \rho_B(i)| + |A-B|(|B|+1) - \sum_{i\in A-B} \rho_A(i) + |B-A|(|A|+1) - \sum_{i\in B-A} \rho_B(i),\end{aligned}$$

donde $\rho_X(i)$ es la posición que ocupa el elemento i en el heteroranking X.

El máximo valor que puede tomar esta distancia es $\frac{|A|(|B|+1)+|B|(|A|+1)}{2}$, por lo que es fácil obtener una función de similitud normalizada a partir de ella.

Definición 40 (Función de similitud de Spearman para heterorankings)**.** Dados dos heterorankings A y B, la *función de similitud de Spearman para heterorankings* se defie como

$$\sigma(A,B) = 1 - 2\frac{\delta(A,B)}{|A|(|B|+1) + |B|(|A|+1)},$$

donde $\delta(A,B)$ es la distancia de Spearman para los heterorankings A y B.

Ejemplo 25 (Función de similitud de Spearman para heterorankings)**.** Considerando los heterorankings $A = [a, \alpha, b, \beta, c, \gamma]$ y $B = [c, \alpha, b, \delta, \epsilon]$, el rango o posición de cada elemento en el heteroranking es

Rango $\rho(i)$	1	2	3	4	5	6
A	a	α	b	β	c	γ
B	c	α	b	δ	ϵ	

En este caso $A \cap B = \{\alpha, b, c\}$, $A - B = \{a, \beta, \gamma\}$ y $B - A = \{\delta, \epsilon\}$, de manera que la similitud de Spearman entre A y B es

$$\sigma(A, B) = 1 - 2\frac{3 + 3 \cdot 7 - 11 + 2 \cdot 6 - 9}{6 \cdot 7 + 5 \cdot 6} = 1 - 2\frac{16}{72} = 0{,}55.$$

Para la Tau de Kendall, una posible extensión es la siguiente.

Definición 41 (Distancia Tau de Kendall para heterorankings)**.** Dados dos heterorankings A y B, la *distancia Tau de Kendall para heterorankings* se define como

$$\delta_p(A, B) = \sum_{(i,j) \in C(A \cup B)} \tau(i, j)$$

donde $C(A \cup B)$ son las combinaciones de dos elementos tomados de $A \cup B$, es decir, el conjunto de pares de elementos distintos en $A \cup B$ sin tener en cuenta el orden, y $\tau(i, j)$ es una función de penalización que se define de la siguiente manera:

- Si $i, j \in A \cap B$ entonces, si i y j aparecen en el mismo orden en A y B, $\tau(i, j) = 0$, y si aparece en distinto orden $\tau(i, j) = 1$.
- Si i, j aparecen en un heteroranking, por ejemplo A, pero sólo uno de ellos aparece en B, por ejemplo i, entonces si i aparece delante de j en A, $\tau(i, j) = 0$, y en caso contrario, $\tau(i, j) = 1$.
- Si i aparece en un heteroranking, por ejemplo A, pero no j, y j aparece en el otro, pero no i, entonces $\tau(i, j) = 1$.
- Si i, j aparece en un heteroranking, por ejemplo A, pero no aparecen en el otro, entonces $\tau(i, j) = p$, con $p \in [0, 1]$.

Obsérvese que esta distancia es en realidad una familia de distancias que depende del valor de la penalización p. La elección más optimista corresponde a $p = 0$ y la más pesimista a $p = 1$. En el peor caso, el valor máximo que podría tomar esta distancia es $|A \cup B|(|A \cup B| - 1)/2$, de manera que es fácil obtener una función de similitud normalizada a partir de esta distancia.

Definición 42 (Función de similitud Tau de Kendall para heterorankings)**.** Dados dos heterorankings A y B, la *función de similitud Tau de Kendall para heterorankings* se define como

$$\sigma_p(A, B) = 1 - 2\frac{\delta_p(A, B)}{|A \cup B|(|A \cup B| - 1)}$$

donde $\delta(A, B)$ es la distancia Tau de Kendall para los heterorankings A y B.

Ejemplo 26 (Función de similitud Tau de Kendall para heterorankings)**.** Tomando los mismo heterorankings del ejemplo anterior $A = [a, \alpha, b, \beta, c, \gamma]$ y $B = [c, \alpha, b, \delta, \epsilon]$, la similitud Tau de Kendall entre A y B es

$$\sigma(A, B) = 1 - 2\frac{3 + 3 \cdot 7 - 11 + 2 \cdot 6 - 9}{6 \cdot 7 + 5 \cdot 6} = 1 - 2\frac{16}{72} = 0{,}55.$$

Estas medidas de similitud otorgan el mismo peso a las discrepancias independientemente de la posición que ocupen en los rankings. En muchas ocasiones, como por ejemplo en los rankings de páginas web devueltas por los buscadores, las discrepancias en las primeras posiciones de los rankings tienen más importancia que en las últimas posiciones, por lo que se suelen utilizar funciones de (di)similitud para la comparación de rankings que otorgan diferentes pesos a las discrepancias dependiendo de la posición en la que ocurren.

Una de estas funciones, ampliamente utilizada en el ámbito de la Recuperación de Información, es el solapamiento parcial de rangos [35]. Esta función compara dos rankings por medio de la comparación de sus prefijos. Dados dos rankings A y B, para cada prefijo de longitud k, se calcula la proporción de solapamiento de los prefijos $\frac{|A[1,k] \cap B[1,k]|}{k}$. Finalmente estas proporciones de solapamiento se agregan mediante una media ponderada, otorgando a la proporción de solapamiento de los prefijos de longitud k un peso p^{k-1},

$$\sigma_p(A,B) = (1-p)\sum_{k=1}^{\infty} p^{k-1}\frac{|A[1,k] \cap B[1,k]|}{k},$$

donde $p \in (0,1)$ es un parámetro que determina la velocidad de decrecimiento de los pesos con la longitud de los prefijos, es decir, cuanto más pequeño sea p, más peso tendrán los elementos del comienzo de los ranking. Esta función de similitud está normalizada puesto que $\sum_{k=1}^{\infty} p^{k-1} = \frac{1}{1-p}$.

Aunque esta función está pensada para rankings potencialmente infinitos, puede adaptarse fácilmente para heterorankings finitos de distinto tamaño.

Definición 43 (Función de similitud del solapamiento parcial de heterorankings)**.** Dados dos heterorankings A y B, la *función de similitud del solapamiento parcial de heterorankings* se define como

$$\sigma_p(A,B) = \frac{\sum_{k=1}^{\text{máx}\{|A|,|B|\}} p^{k-1}\frac{|A[1,k] \cap B[1,k]|}{k}}{\sum_{k=1}^{\text{máx}\{|A|,|B|\}} p^{k-1}}.$$

donde $p \in (0,1]$.

Ejemplo 27 (Función de similitud del solapamiento parcial de heterorankings)**.** Tomando de nuevo los heterorankings $A = [a, \alpha, b, \beta, c, \gamma]$ y $B = [c, \alpha, b, \delta, \epsilon]$, en la tabla 5 se muestra para varios valores de p el solapamiento para cada prefijo de longitud $k = 1, \ldots, \text{máx}\{|A|, |B|\}$ y la media ponderada para cada k.

Tomando $p = 1$ se otorga el mismo peso a todos prefijos. En este caso la similitud del solapamiento parcial entre A y B es

$$\sigma_1(A,B) = \frac{\sum_{k=1}^{6}\frac{|A[1,k] \cap B[1,k]|}{k}}{6} = \frac{2{,}77}{6} = 0{,}46.$$

Tomando $p = 1/2$ se otorga el doble de peso a cada elemento que a su sucesor. En este caso la similitud del solapamiento parcial entre A y B es

$$\sigma_1(A,B) = \frac{\sum_{k=1}^{6} 0{,}5^{k-1}\frac{|A[1,k] \cap B[1,k]|}{k}}{\sum_{k=1}^{6} 0{,}5^{k-1}} = \frac{0{,}53}{1{,}97} = 0{,}27.$$

En este caso la medida de similitud es claramente inferior ya que el primer elemento de los rankings no coincide.

Tabla 5. Solapamiento de los heterorankings $A = [a, \alpha, b, \beta, c, \gamma]$ y $B = [c, \alpha, b, \delta, \epsilon]$.

k	$A[1,k]$	$B[1,k]$	$\lvert A[1,k] \cap B[1,k] \rvert$	$\lvert A[1,k] \cap B[1,k] \rvert / k$
1	$[a]$	$[c]$	0	0
2	$[a, \alpha]$	$[c, \alpha]$	1	0,5
3	$[a, \alpha, b]$	$[c, \alpha, b]$	2	0,67
4	$[a, \alpha, b, \beta]$	$[c, \alpha, b, \delta]$	2	0,5
5	$[a, \alpha, b, \beta, c]$	$[c, \alpha, b, \delta, \epsilon]$	3	0,6
6	$[a, \alpha, b, \beta, c, \gamma]$	$[c, \alpha, b, \delta, \epsilon]$	3	0,5

3.2.10. Funciones de similitud para heterocombinaciones

Las heterocombinaciones son casos particulares de heteroconjuntos en los que se ha fijado la cardinalidad, y también de cajas en las que los elementos no pueden repetirse, de manera que pueden utilizarse todas las funciones de similitud de las secciones 3.2.3 y 3.2.5 para comparar heterocombinaciones. Más allá de estas funciones, en la literatura científica no se han encontrado otras funciones que utilicen la condición de cardinalidad fija o de unicidad.

3.2.11. Funciones de similitud para tuplas

Las tuplas son un caso particular de listas y de cajas, de manera que pueden utilizarse todas las funciones de similitud de las secciones 3.2.4 y 3.2.5 para comparar tuplas.

Dado que las tuplas tienen una cardinalidad fija, en el caso de que los elementos en la misma posición pertenecen al mismo universo, es posible comparar los elementos que ocupan la misma posición utilizando funciones de similitud definidas sobre los elementos de cada universo y después combinar las medidas de similitud mediante una función de agregación.

Definición 44 (Función de agregación)**.** Una *función de agregación* es una función $f : [0,1]^n \longrightarrow [0,1]$ que cumple las siguientes propiedades:

a) $f(0, \ldots, 0) = 0$.
b) $f(1, \ldots, 1) = 1$.
c) f es monótona en cada argumento, es decir,

$$f(x_1, \ldots, x_i, \ldots, x_n) \leq f(y_1, \ldots, y_i, \ldots, y_n) \text{ si } x_j = y_j \; \forall j \neq i \text{ y } x_i \leq y_i.$$

Esta definición de función de agregación está pensada para agregar medidas realizadas con funciones de similitud normalizadas.

La función de agregación más pesimista corresponde al mínimo, y la más optimista al máximo.

Definición 45 (Función de similitud mínima)**.** Dadas dos tuplas A y B con elementos en k universos de elementos $E_1, \ldots, E_k$, y $\sigma_i : E_i \times E_i \longrightarrow [0,1]$ $(i = 1, \ldots, k)$ funciones de similitud para los elementos de los universos E_i, la *función de similitud mínima* se define como

$$\sigma(A, B) = \min_{i=1,\ldots,k} \sigma_i(A[i], B[i]).$$

Definición 46 (Función de similitud máxima)**.** Dadas dos tuplas A y B con elementos en k universos de elementos $E_1, \ldots, E_k$, y $\sigma_i : E_i \times E_i \longrightarrow [0,1]$ $(i = 1, \ldots, k)$ funciones de similitud para los elementos de los universos E_i,la *función de similitud máxima* se define como

$$\sigma(A,B) = \max_{i=1,\ldots,k} \sigma_i(A[i], B[i]).$$

La familias de funciones de similitud para tuplas más habituales utilizan medias como funciones de agregación.

Definición 47 (Función de similitud media)**.** Dadas dos tuplas A y B con elementos en k universos de elementos $E_1, \ldots, E_k$, y $\sigma_i : E_i \times E_i \longrightarrow [0,1]$ $(i = 1, \ldots, k)$ funciones de similitud para los elementos de los universos E_i,la *función de similitud mínima* se define como

$$\sigma(A,B) = \frac{\sum_{i=1}^{k} \sigma_i(A[i], B[i])}{k}.$$

Dependiendo de las funciones de similitud utilizadas para cada universo E_i se obtienen diferentes funciones de similitud. Por ejemplo, tomando σ_i como la función de similitud identidad, la función de similitud media resultante es la función de similitud de Hamming (ver definición **??**).

Ejemplo 28 (Función de similitud media)**.** Considérense las tuplas correspondientes a la descripción de dos vinos $A = [14{,}0; 17; \text{Excelente}; 80]$ y $B = [13{,}5; 4; \text{Muy bueno}; 10]$ donde la primera componente es el contenido alcohólico (de 0 a 15 grados), la segunda el tiempo de envejecimiento en barrica (de 0 a 24 meses), la tercera la valoración de cata (malo, medio, bueno, muy bueno, excelente) y la última el precio (de 0 a 100 €). Puesto que el contenido alcohólico, el tiempo de envejecimiento y el precio son valores numéricos se puede utilizar la función de similitud de la diferencia para compararlos. Para la valoración de cata, puesto que se trata de una escala ordinal, también se puede utilizar la función de similitud de la diferencia sobre el número de orden de cada categoría en la escala (1:malo, 2:medio, 3:bueno, 4:muy bueno, 5:excelente). Así, la similitud de cada uno de los componentes de las tuplas son

Vino	Alcohol	Tiempo barrica	Valoración cata	Precio
A	14.0	17	Excelente	80
B	13.5	4	Muy bueno	10
Diferencia δ	$\frac{14-13{,}5}{15-0} = 0{,}033$	$\frac{17-4}{24-0} = 0{,}542$	$\frac{5-4}{5-1} = 0{,}25$	$\frac{80-10}{100-0} = 0{,}7$
Similitud $\sigma = 1 - \delta$	0,967	0,458	0,75	0,3

de manera que la similitud media de las tuplas correspondientes a los los vinos A y B es

$$\sigma(A,B) = \frac{0{,}967 + 0{,}458 + 0{,}75 + 0{,}3}{4} = 0{,}619.$$

Cuando los elementos de las tuplas tienen distinta importancia es mejor utilizar como función de agregación la media ponderada.

Definición 48 (Función de similitud media ponderada)**.** Dadas dos tuplas A y B con elementos en k universos de elementos $E_1, \ldots, E_k$, y $\sigma_i : E_i \times E_i \longrightarrow [0,1]$ $(i = 1, \ldots, k)$ funciones de similitud para los elementos de los universos E_i, la *función de similitud media ponderada* se define como

$$\sigma(A,B) = \frac{\sum_{i=1}^{k} w_i \sigma_i(A[i], B[i])}{\sum_{i=1}^{k} w_i},$$

donde $w_i \in \mathbb{R}^+$ $(i = 1, \ldots, k)$ es el peso de la componente i-ésima de la tupla.

Ejemplo 29 (Función de similitud media ponderada)**.** Considerando el ejemplo anterior de las tuplas correspondientes a los vinos A y B, si se le da la misma importancia al tiempo de envejecimiento que al contenido alcohólico, el doble de importancia a la valoración de cata que al contenido alcohólico, y el triple de importancia al precio que al contenido alcohólico, la similitud media ponderada de las tuplas correspondientes a los los vinos A y B es

$$\sigma(A,B) = \frac{1 \cdot 0{,}967 + 1 \cdot 0{,}458 + 2 \cdot 0{,}75 + 3 \cdot 0{,}3}{1+1+2+3} = 0{,}546.$$

También es posible utilizar funciones de agregación no lineales.

Definición 49 (Función de similitud media generalizada)**.** Dadas dos tuplas A y B con elementos en k universos de elementos $E_1, \ldots, E_k$, y $\sigma_i : E_i \times E_i \longrightarrow [0,1]$ $(i = 1, \ldots, k)$ funciones de similitud para los elementos de los universos E_i,la *función de similitud media generalizada* se define como

$$\sigma(A,B) = \sqrt[\alpha]{\sum_{i=1}^{k} w_i \sigma_i(A[i], B[i])^\alpha},$$

donde $w_i \in \mathbb{R}^+$ $(i = 1, \ldots, k)$ es el peso de la componente i-ésima de la tupla y $\alpha \in \mathbb{R}^+$.

En ocasiones conviene aplicar una función de tipo sigmoidal a las medidas de similitud antes de agregarlas para sobrevalorar las similitudes altas e infravalorar las similitudes bajas. Esta decisión puede justificarse en algunos casos como por ejemplo en la comparación de cadenas. Si la diferencia entre dos cadenas es solo uno o dos caracteres, entonces es muy probable que ambas cadenas representen el mismo concepto, mientras que si las cadenas solo tienen en común dos o tres caracteres, su similitud es prácticamente nula.

3.2.12. Funciones de similitud para rankings

Los rankings son un caso particular de conjuntos, secuencias y heterorankings, de manera que pueden utilizarse todas las funciones de similitud de las secciones 3.2.6, 3.2.7 y 3.2.9 para comparar rankings.

3.2.13. Funciones de similitud para combinaciones

Las combinaciones son un caso particular de conjuntos, multicombinaciones y heterocombinaciones, de manera que pueden utilizarse todas las funciones de similitud de las secciones 3.2.6, 3.2.8 y 3.2.10 para comparar combinaciones.

3.2.14. Funciones de similitud para vectores

Los vectores son un caso particular de secuencias, multicombinaciones y tuplas, por lo que pueden utilizarse todas las funciones de similitud de las secciones 3.2.7, 3.2.8 y 3.2.11 para comparar vectores.

Como todos los elementos de un vector pertenecen al mismo universo, es posible comparar los elementos que ocupan la misma posición, al igual que para las tuplas, utilizando una misma función de similitud, y luego combinar esas medidas de similitud. Existen diferentes formas de combinar las medidas de similitud de los elementos, siendo las más comunes las derivadas de la función de similitud media generalizada. Una de las más conocidas se debe a Minkowski [38].

Definición 50 (Distancia de Minkowski)**.** Dados dos vectores A y B definidos sobre un mismo universo E y una distancia normalizada δ_E para los elementos de E, la *distancia de Minkowski* se define como

$$\delta(A,B) = \left(\sum_{i=1}^{n} \delta_E(A[i],B[i])^p\right)^{1/p},$$

donde $p \in \mathbb{Z}_{\geq 0}$.

La función de similitud asociada se conoce como *función de similitud de Minkowski*

$$\sigma(A,B) = 1 - \delta(A,B).$$

Obsérvese que si δ_E no está normalizada, esta distancia tampoco lo está.

Ejemplo 30 (Distancia de Minkowski)**.** Considérense los vectores $A = [9,10,9]$ y $B = [8,7,7]$ correspondientes a las puntuaciones de dos alumnos en tres pruebas. Puesto que el universo de los elementos de los vectores son números, es posible compararlos mediante la distancia de la diferencia normalizada, teniendo en cuenta que las puntuaciones de 0 a 10. De este modo la similitud de Minkowski de A y B tomando $p = 3$ es

$$\delta(A,B) = \left(\left(\frac{|9-8|}{10}\right)^3 + \left(\frac{|10-7|}{10}\right)^3 + \left(\frac{|9-7|}{10}\right)^3\right)^{1/3} = 0{,}33,$$

y la medida de similitud de Minkowski es

$$\sigma(A,B) = 1 - 0{,}33 = 0{,}67.$$

Esta familia de distancias da lugar a varias distancias bien conocidas para distintos valores de p. Por ejemplo, cuando el universo E es numérico, δ_E es la distancia de la diferencia no normalizada y $p = 1$, se obtiene la distancia de Manhattan [7].

Definición 51 (Distancia de Manhattan)**.** Dados dos vectores A y B definidos sobre un mismo universo numérico E, la *distancia de Manhattan* se define como

$$\delta(A,B) = \sum_{i=1}^{n} |A[i] - B[i]|.$$

Ejemplo 31 (Distancia de Manhattan)**.** Siguiendo con el ejemplo anterior, la distancia de Manhattan de los vectores A y B correspondientes a las puntuaciones de los alumnos es

$$\delta(A,B) = |9-8| + |10-7| + |9-7| = 6.$$

Teniendo en cuenta que las puntuaciones van de 0 a 10, se puede normalizar esta distancia dividiendo en entre la mayor distancia que es $10 + 10 + 10 = 30$,

$$\overline{\delta}(A,B) = \frac{6}{30} = 0{,}2,$$

y la medida de similitud asociada es

$$\sigma(A,B) = 1 - 0{,}2 = 0{,}8.$$

Para $p = 2$ resulta la distancia Euclídea [7].

Definición 52 (Distancia euclídea)**.** Dados dos vectores A y B definidos sobre un mismo universo numérico E, la *distancia euclídea* se define como

$$\delta(A,B) = \sqrt{\sum_{i=1}^{n} (A[i] - B[i])^2}.$$

Ejemplo 32 (Distancia euclídea)**.** Siguiendo con el ejemplo anterior, la distancia euclídea de los vectores A y B correspondientes a las puntuaciones de los alumnos es

$$\delta(A,B) = \sqrt{(9-8)^2 + (10-7)^2 + (9-7)^2)} = 3{,}742.$$

Teniendo en cuenta que las puntuaciones van de 0 a 10, se puede normalizar esta distancia dividiendo en entre la mayor distancia que es $\sqrt{10^2 + 10^2 + 10^2} = 17{,}321$,

$$\overline{\delta}(A,B) = \frac{3{,}742}{17{,}321} = 0{,}216,$$

y la medida de similitud asociada es

$$\sigma(A,B) = 1 - 0{,}216 = 0{,}784.$$

Y para $p = +\infty$ se tiene la distancia de Tchebychev [7].

Definición 53 (Distancia de Tchebychev)**.** Dados dos vectores A y B definidos sobre un mismo universo numérico E, la *distancia de Tchebychev* se define como

$$\delta(A,B) = \lim_{p\to+\infty} \left(\sum_{i=1}^{n} (A[i] - A[i])^p \right)^{1/p} = \max_{i=1}^{n} |A[i] - B[i]|.$$

Ejemplo 33 (Distancia de Tchebychev)**.** Tomando de nuevo el ejemplo anterior, la distancia de Tchebychev de los vectores A y B es

$$\delta(A,B) = \max\{|9-8|, |10-7|, |9-7|\} = 3.$$

Teniendo en cuenta que las puntuaciones de la cata van de 0 a 10, se puede normalizar esta distancia dividiendo en entre la mayor distancia que es 10,

$$\overline{\delta}(A,B) = \frac{3}{10} = 0{,}3,$$

y la medida de similitud asociada es

$$\sigma(A,B) = 1 - 0{,}3 = 0{,}7.$$

Cuando los universos que dan soporte a los vectores son numéricos los vectores pueden representarse en un espacio real n-dimensional y es posible aprovechar aspectos geométricos, como el ángulo que forman, para compararlos. Las siguientes funciones de similitud utilizan este enfoque [7].

Definición 54 (Función de similitud coseno)**.** Dados dos vectores A y B definidos sobre un mismo universo numérico E, la *función de similitud coseno* se define como

$$\sigma(A,B) = \frac{A \cdot B}{|A||B|},$$

donde $A \cdot B$ es el producto escalar de los vectores A y B.

Ejemplo 34 (Función de similitud coseno)**.** Siguiendo con el ejemplo anterior, la función de similitud coseno los vectores A y B es

$$\sigma(A,B) = \frac{9\cdot 8 + 10\cdot 7 + 9\cdot 7}{\sqrt{9^2 + 10^2 + 9^2}\sqrt{8^2 + 7^2 + 7^2}} = 0{,}995.$$

Definición 55 (Función de similitud de Tanimoto)**.** Dados dos vectores A y B definidos sobre un mismo universo numérico E, la *función de similitud de Tanimoto* se define como

$$\sigma(A,B) = \frac{A \cdot B}{|A|^2 + |B|^2 - A \cdot B},$$

donde $A \cdot B$ es el producto escalar de los vectores A y B.

Ejemplo 35 (Función de similitud de Tanimoto)**.** Utilizando una vez más el ejemplo anterior, la función de similitud de Tanimoto de los vectores A y B es

$$\sigma(A,B) = \frac{9 \cdot 8 + 10 \cdot 7 + 9 \cdot 7}{(9^2 + 10^2 + 9^2) + (8^2 + 7^2 + 7^2) - (9 \cdot 8 + 10 \cdot 7 + 9 \cdot 7)} = 0{,}936.$$

3.2.15. Funciones de similitud para heterovariaciones

Las heterovariaciones son casos particulares de heterorankings, heterocombinaciones y tuplas, de manera que pueden utilizarse todas las funciones de similitud de las secciones 3.2.9, 3.2.10 y 3.2.11 para comparar heterovariaciones.

Como las heterovariaciones tienen cardinalidad fija, para compararlas es posible utilizar las extensiones de la distancia de Spearman, el coeficiente de correlación de Spearman y la distancia Tau de Kendall propuestas por [10] para permutaciones.

Definición 56 (Distancia de Spearman para heterovariaciones)**.** Dadas dos heterovariaciones A y B, la *distancia de Spearman para heterovariaciones* se define como

$$\begin{aligned}\delta(A,B) &= \sum_{i \in A \cap B} |\rho_A(i) - \rho_B(i)| + \sum_{i \in A - B} ((k+1) - \rho_A(i)) + \sum_{i \in B - A} ((k+1) - \rho_B(i)) = \\ &= 2|A - B|(k+1) + \sum_{i \in A \cap B} |\rho_A(i) - \rho_B(i)| - \sum_{i \in A - B} \rho_A(i) - \sum_{i \in B - A} \rho_B(i),\end{aligned}$$

donde $\rho_X(i)$ es la posición que ocupa el elemento i en la heterovariación X.

El máximo valor que puede tomar esta distancia es $k(k+1)$ por lo que es fácil obtener la función de similitud normalizada asociada.

Definición 57 (Función de similitud de Spearman para heterovariaciones)**.** Dadas dos heterovariaciones A y B, la *función de similitud de Spearman para heterovariaciones* se define como

$$\sigma(A,B) = 1 - \frac{\delta(A,B)}{k(k+1)},$$

donde $\delta(A,B)$ es la distancia de Spearman para heterovariaciones.

Definición 58 (Distancia Tau de Kendall para heterovariaciones)**.** Dadas dos heterovariaciones A y B, la *distancia Tau de Kendall para heterovariaciones* se define como

$$\delta_p(A,B) = k - |A \cap B|((2+p)k - p|A \cap B| + 1 - p) + \sum_{i,j \in A \cap B} \tau(i,j) - \sum_{j \in A - B} \rho_A(j) - \sum_{j \in B - A} \rho_B(j).$$

donde $\rho_X(i)$ es la posición que ocupa el elemento i en la heterovariación X, y $\tau(i,j)$ es una función tal que $\tau(i,j) = 0$ si i y j aparecen en el mismo orden en A y B, y $\tau_p(i,j) = 1$ si aparecen en distinto orden.

El máximo valor que puede tomar esta distancia es $2k^2 - k$ por lo que es fácil obtener la función de similitud normalizada asociada.

Definición 59 (Función de similitud Tau de Kendall para heterovariaciones)**.** Dadas dos heterovariaciones A y B, la *función de similitud Tau de Kendall para heterovariaciones* se define como

$$\sigma_p(A,B) = 1 - \frac{\delta_p(A,B)}{2k^2 - k},$$

donde $\delta_p(A,B)$ es la distancia Tau de Kendall para heterovariaciones.

3.2.16. Funciones de similitud para variaciones

Las variaciones son casos particulares de rankings, combinaciones, vectores y heterovariaciones, de manera que pueden utilizarse todas las funciones de similitud de las secciones 3.2.12, 3.2.13, 3.2.14 y 3.2.15 para comparar variaciones. En definitiva, al ser subclases de todas las demás clases de colecciones, pueden utilizarse todas las funciones de similitud vistas en esta sección.

3.3. Conclusiones

En este trabajo se ha presentado una nueva catalogación de funciones de similitud para la comparación de coleccione o agrupaciones de objetos, estructurada de acuerdo una clasificación taxonómica de los tipos de colecciones según sus características estructurales (homogeneidad, unicidad, orden y cardinalidad).

Esta taxonomía resulta útil para identificar las funciones de similitud más apropiadas para comparar dos colecciones dadas, incluso sin ser del mismo tipo, utilizando las funciones de similitud de los tipos de colecciones que son ancestros comunes en la taxonomía de colecciones.

Sin duda, esta catalogación puede ayudar al usuario a seleccionar las funciones de similitud más apropiadas y también puede facilitar la definición de reglas de aplicación que automaticen el proceso de comparación, como por ejemplo la aplicación de funciones de similitud en cascada.

La utilidad de este catálogo queda demostrada en [30] donde se ha aplicado con éxito en el dominio de la enología para la comparación de vinos.

En un próximo futuro se pretende implementar todas las funciones de este catálogo en una librería para los principales lenguajes de programación.

Referencias

[1] AIGNER, A. *Combinatorial Theory*. Springer Verlag, New York/Berlin, 1979.

[2] BIZER, C., HEATH, T. and BERNERS-LEE, T. *Linked Data – The Story So Far*. International Journal on Semantic Web and Information Systems, 5(3):1–22, 2009.

[3] BLIZARD, W. *Multiset Theory*. Notre Dame Journal of Formal Logic, 30(1):36–66, 1989.

[4] BRAUN-BLANQUET, J. *Plant sociology: the study of plant communities*, page 439. McGraw-Hill, 1932.

[5] CAILLEZ, F. and KUNTZ, P. *A contribution to the study of the metric and euclidiean structures of dissimilarities*. Psychometrika, 61(2):241–253, 1996.

[6] CORMEN, T. ET AL. *Introduction to Algorithms*. MIT Press, Cambridge, Massachusetts, 3rd edition, 2009.

[7] DEZA, M. and DEZA, E. *Encyclopedia of distances*. Springer, Berlin, 2009.

[8] DICE, L. *Measures of the Amount of Ecologic Association Between Species*. Ecology, 26(3):297–302, July 1945.

[9] EUZENAT, J. and SHVAIKO, P. *Ontology Matching*. Springer, Berlin, Heidelberg, 2007.

[10] FAGIN, R., KUMAR, R. and SIVAKUMAR, D. *Comparing Top k Lists*. SIAM Journal on Discrete Mathematics, 17(1):134–160, 2003.

[11] FOREMAN, M. and KANAMORI, A. *Handbook Set Theory*. Springer, Berlin, 2006.

[12] GOWER, J. and LEGENDRE, P. *Metric and euclidean properties of dissimilarity coefficients*. Journal of classification, 3(1):5–48, 1986.

[13] HAHN, U. and CHATER, N. *Concepts and similarity*. In L. Lamberts and D. Shanks, editors, Knowledge, Concepts and Categories, Cambridge, Massachusetts, 1997. Physhology Press/MIT Press.

[14] HAMMING, R. *Error-detecting and error-correcting codes*. 29(2):147–160, 1950.

[15] HARTMANN, S. LINK, S. *Collection Type Constructors in Entity-Relationship Modeling*. In Proceedings of the 26th International Conference on Conceptual Modeling, Lecture Notes in Computer Science, pages 307–322. Springer, 2007.

[16] HENDLER, J. *Web 3.0: The Dawn of Semantic Search*. Computer, 43(1):77–80, 2010.

[17] JACCARD, P. *Étude comparative de la distribution florale dans une portion des Alpes et des Jura*. Bulletin del la Société Vaudoise des Sciences Naturelles, 37:547–579, 1901.

[18] JARO, M. *Advances in Record-Linkage Methodology as Applied to Matching the 1985 Census of Tampa, Florida*. Journal of the American Statistical Association, 84(406):414–420, 1989.

[19] KENDALL, M. and GIBBONS, J. *Rank Correlation Methods*. A Charles Griffin Title, 5 edition, September 1990.

[20] KENT, W. *The Semantics of Aggregate Objects*. Technical report, Hewlett-Packard Laboratories, Database Technology Department, Palo Alto, California, October 1988.

[21] KENT, W. *Carriers And Cargos: A General Paradigm For Modeling Collections, Tables, Multimeda, Spatial Constructs, And Other Intensional/Extionsional Objects*. Technical report, Hewlett-Packard Laboratories, Database Technology Department, Palo Alto, California, November 1993.

[22] KONDRAK, G. *N-Gram Similarity and Distance.* In Mariano P. Consens and Gonzalo Navarro, editors, 12th International Conference String Processing and Information Retrieval (SPIRE), volume 3772 of Lecture Notes in Computer Science, Berlin, Germany, pages 115–126, Buenos Aires, Argentina, 2005. Springer.

[23] KULCZYNSKI, S. *Die Pflanzenassoziationen der Pieninen*. Bulletin International de l'Academie Polonaise des Sciences et des Lettres, Classe des Sciences Mathematiques et Naturelles, B (Sciences Naturelles), II:57–203, 1927.

[24] LEVENSHTEIN, V. *Binary codes capable of correcting deletions, insertions, and reversals*. Cybernetics and Control Theory, 10(8):707–710, 1966. Original in Doklady Akademii Nauk SSSR 163(4): 845–848 (1965).

[25] MANNING, C. and SCHÜTZE, H. *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, Massachusetts, 1999.

[26] NEEDLEMAN, S. and WUNSCH, C. *A general method applicable to the search for similarities in the amino acid sequence of two proteins*. Journal of Molecular Biology, 48(3):443–453, 1970.

[27] OCHIAI, A. *Zoogeographical studies on the soleoid fishes found Japan and its neighboring regions*. Bulletin of the Japanese Society of Scientific Fischeries, 22(9):526–530, 1957.

[28] RODRÍGUEZ, A. and EGENHOFER, M. *Determining Semantic Similarity Among Entity Classes from Different Ontologies*. IEEE Transactions on Knowledge and Data Engineering, 15(2):442–456, 2003.

[29] SALTON, G., WONG, A. and YANG, C. *A vector space model for automatic indexing*. Communications of the ACM, 18(11):613–620, 1975.

[30] SÁNCHEZ-ALBERCA, A. *Modelado y comparación de colecciones en la Web Semántica*. PhD thesis, ETS de Ingenieros Informáticos (UPM Madrid), 2015.

[31] SIMPSON, G. *Notes on the measurement of faunal resemblance*. American Journal of Science, 258-A:300–311, 1960.

[32] TVERSKY, A. *Features of Similarity*. Psychological Review, 84(4):327–352, July 1977.

[33] VALTCHEV, P. and EUZENAT, J. *Dissimilarity Measure for Collections of Objects and Values.* In Xiaohui Liu, Paul R. Cohen, and Michael R. Berthold, editors, Advances in Intelligent Data Analysis, Reasoning about Data, Second International Symposium, IDA-97, volume 1280 of Lecture Notes in Computer Science, Berlin, Germany, pages 259–272, London, United Kingdom, 1997. Springer.

[34] VAN RIJSBERGEN, C. *Information retrieval*. Butterworths, London (United Kingdom), 2 edition, 1979.

[35] WEBBER, W., MOFFAT, A. and ZOBEL, J. *A similarity measure for indefinite rankings*. ACM Transactions on Information Systems, 28(4):1–38, November 2010.

[36] WEI, W. *Semantic Search: Bringing Semantic Web Technologies to Information Retrieval*. PhD thesis, School of Computer Science, The University of Nottingham, 2009.

[37] WINKLER, W. *The State of Record Linkage and Current Research Problems*. Technical Report Statistical Research Report Series RR99/04, U.S. Bureau of the Census, Washington, D.C., 1999.

[38] XU, R. and WUNSCH, D. *Clustering*. Wiley-IEEE Press, 2008.

Sobre el autor:
Nombre: Alfredo Sánchez Alberca
Correo electrónico: asalber@ceu.es
Web: `http://aprendeconalf.es`
Institución: Universidad CEU San Pablo.
Departamento: Matemática Aplicada y Estadística.