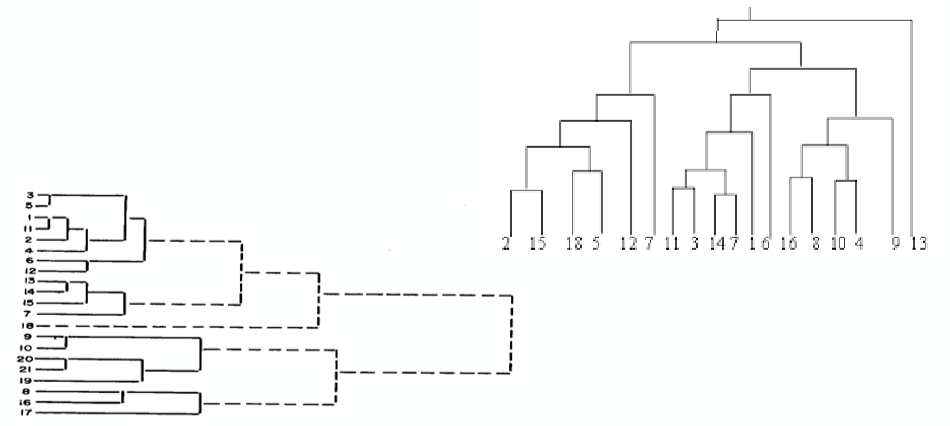
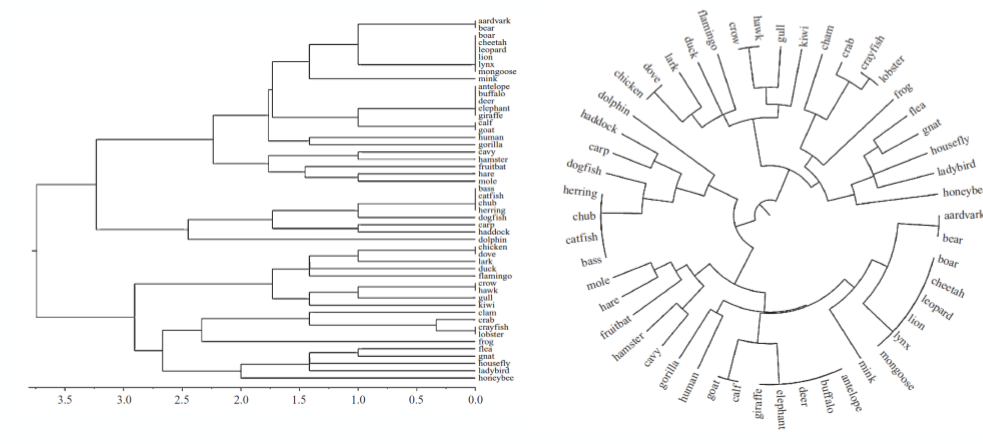
Clustering Jerárquico

* La formación de un par inicial de grupos y luego considerar de forma recursiva si vale la pena dividirlos produce una jerarquía que se puede representar como un árbol binario llamado dendrograma
* La misma información podría ser representada como un diagrama de Venn de conjuntos y subconjuntos
  + La restricción que la estructura es jerárquica corresponde al hecho de que, a pesar de esto, los subconjuntos pueden incluir uno del otro, no pueden entrelazarse.
  + En algunos casos existe una medida del grado de disimilitud entre los grupos en cada conjunto; a continuación, la altura de cada nodo en el dendrograma se puede hacer proporcional a la disimilitud entre sus hijos. Esto proporciona un diagrama de fácil interpretación de un agrupamiento jerarquico.

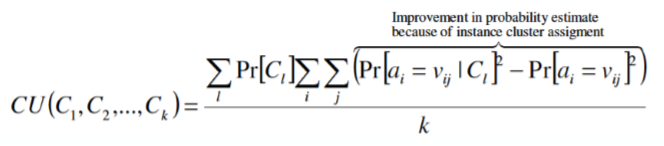




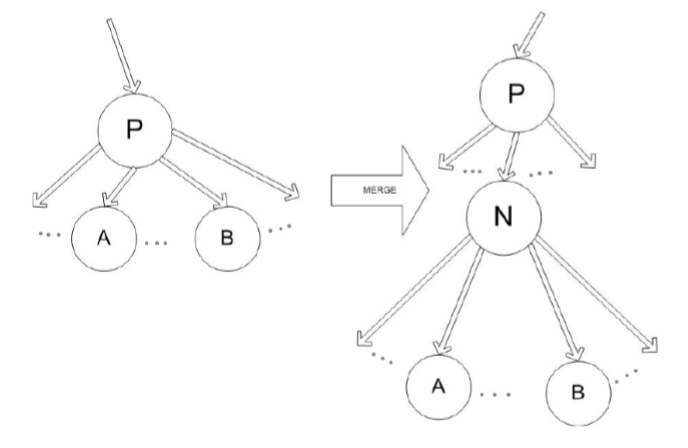
* Una alternativa al método top-down para formar una estructura jerárquica de grupos es utilizar un **enfoque bottom-up**, que se llama la agrupación de aglomeración.
  + Esta idea fue propuesta hace muchos años y ha disfrutado recientemente un resurgimiento en popularidad.
  + El algoritmo básico es simple. Todo lo que se necesita es **una medida de distancia** (o una medida de similitud) entre los dos grupos.
  + Se empieza asignando a cada caso como un grupo en sí mismo; luego se encuentran los dos grupos más cercanos, se **combinan**, y sigue haciendo esto hasta que sólo un grupo quede.
  + El registro de mergings forma una estructura en forma de una **agrupación** **dendrográmica binaria** jerárquica.
* Hay numerosas posibilidades para la medida de distancia.
* Una de ellas es la distancia mínima entre los clusters.
  + La distancia entre sus dos miembros más cercanos.
  + Esto produce lo que se llama el algoritmo de clustering un solo vínculo.
  + Dado que esta medida tiene en cuenta sólo los dos miembros más cercanos de un par de grupos, el procedimiento es sensible a los valores atípicos: La adición de una sola nueva instancia puede alterar radicalmente toda la estructura de la agrupación.
  + Además, si se deﬁne el diámetro de un grupo como la mayor distancia entre sus miembros, un solo vínculo de agrupación puede producir clusters con diámetros muy grandes.
* Otra medida es la distancia máxima entre los grupos, en lugar del mínimo.
  + Se consideran dos grupos cercanos sólo si todas las instancias en su unión son relativamente similares, a veces llamado el método de vinculación exhaustividad.
  + Esta medida, que también es sensible a los valores atípicos, busca racimos compactos con diámetros pequeños
  + Sin embargo, algunos casos pueden terminar mucho más cerca de otros clusters de lo que son para el resto de su propio cluster.

CobWeb

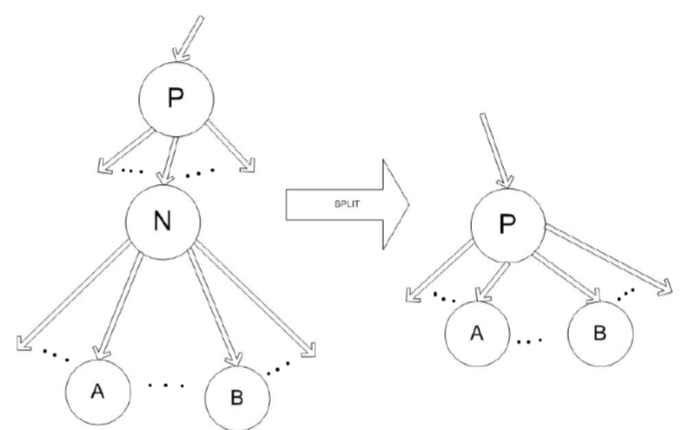
* Desarrollado en 1980 por investigadores de ML, fue creado para agrupar objetos en un conjunto de datos objeto-atributo.
* El algoritmo Cobweb genera un dendograma de agrupamiento llamado árbol de clasificación que caracteriza a cada grupo.
* Este algoritmo construye un árbol de clasificación de forma incremental mediante la inserción de objetos en el árbol de clasificación uno por uno.
* Al insertar un objeto en el árbol de clasificación, el algoritmo cobweb atraviesa el árbol de arriba hacia abajo a partir del nodo raíz
* En cada nodo, el algoritmo considera 4 operaciones posibles y selecciona la que obtiene el valor más alto de la función utilidad de la categoría (CU):
  + Insert
  + Create
  + Merge
  + Split
* Este algoritmo funciona basándose en la llamada función utilidad de la categoría (CU) que mide la calidad de la agrupación
* Si dividimos un conjunto de objetos en m grupos, entonces el CU de esta partición en particular es:



* Inserción significa que un nuevo objeto se insertará en uno de los nodos secundarios existentes.
  + El algoritmo evalúa el valor de la función CU de insertar el nuevo objeto en cada uno de los nodos secundarios existentes y selecciona el que tiene la puntuación más alta.
* El algoritmo también considera la creación de un nuevo nodo hijo especíﬁcamente para el nuevo objeto.
* El algoritmo Cobweb considera también la fusión (merging) de los dos nodos secundarios existentes con el más alto puntaje y con el segundo puntaje más alto



* El algoritmo Cobweb considera dividir (split) un nodo hijo existente con la puntuación más alta



* Entrada
  + El nodo actual N en la jerarquía de conceptos.
  + Una instancia I sin clasiﬁcar (atributo-valor).
* Resultados
  + Una jerarquía de conceptos que clasiﬁca a la instancia.
* Invocación
  + Cobweb (nodo-raíz, I).
* Variables
  + C, P, Q, y R son nodos en la jerarquía.
  + U, V, W y X son los grupos (partición).

