Métodos de teoría de Grafos en aprendizaje no supervisado y clustering

Jorge Luis Guevara Díaz

15 de enero de 2011

1. Teoría de Grafos
2. Métodos de la Teoría de Grafos
3. Clasificación de Documentos Web
4. Clustering Incremental de Documentos
5. Clustering de genes y analisis de expresion de metagenes basados en grafos

Teoría de Grafos I

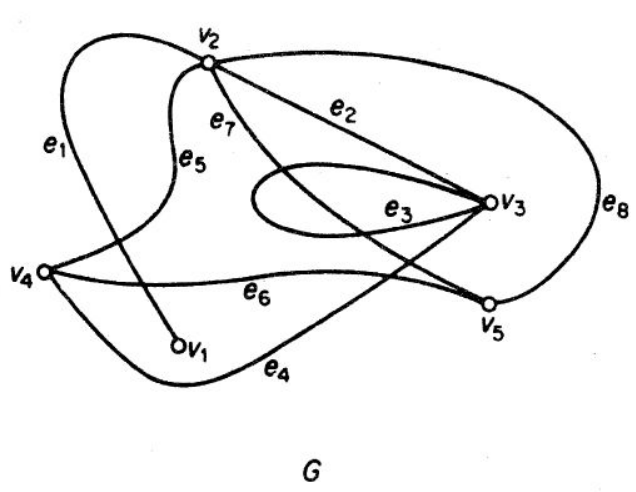
1. Un Grafo es una terna ordenada $(V(G), E(G), \psi_G)$, donde $V(G) \neq \phi$ es un conjunto de *vértices*, $E(G)$ es el conjunto de *aristas*, tal que $E(G) \cap V(G) = \phi$, ψ_G es la *función de incidencia* que asocia cada arista de G un par cualquiera de vértices no necesariamente distintos de G, tal que si e es una arista, y u, v son dos vértices, entonces $\psi(e) = uv$. [1]
2. Ejemplo

$$\begin{aligned} G &= (V(G), E(G), \psi_G) \\ V(G) &= \{v_1, v_2, v_3, v_4, v_5\} \\ E(G) &= \{e_1, e_2, e_3, e_4, e_5, e_6, e_7, e_8\} \end{aligned}$$

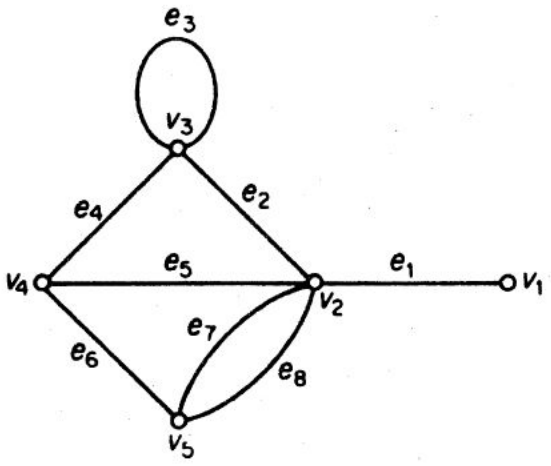
y la función de incidencia ψ_G definida como:

Teoría de Grafos II

$$\psi_G(e_1) = v_1 v_2,\ \psi_G(e_2) = v_2 v_3,\ \psi_G(e_3) = v_3 v_3,$$
$$\psi_G(e_4) = v_3 v_4,\ \psi_G(e_5) = v_2 v_4,\ \psi_G(e_6) = v_4 v_5,$$
$$\psi_G(e_7) = v_2 v_5,\ \psi_G(e_8) = v_2 v_5$$



(a) Diagrama del grafo G



(b) Otro diagrama del grafo G

Figura: Diagramas del grafo G

Aplicaciones de la teoría de grafos I

1. Problema del camino mas corto, Algoritmo de Dikstra.

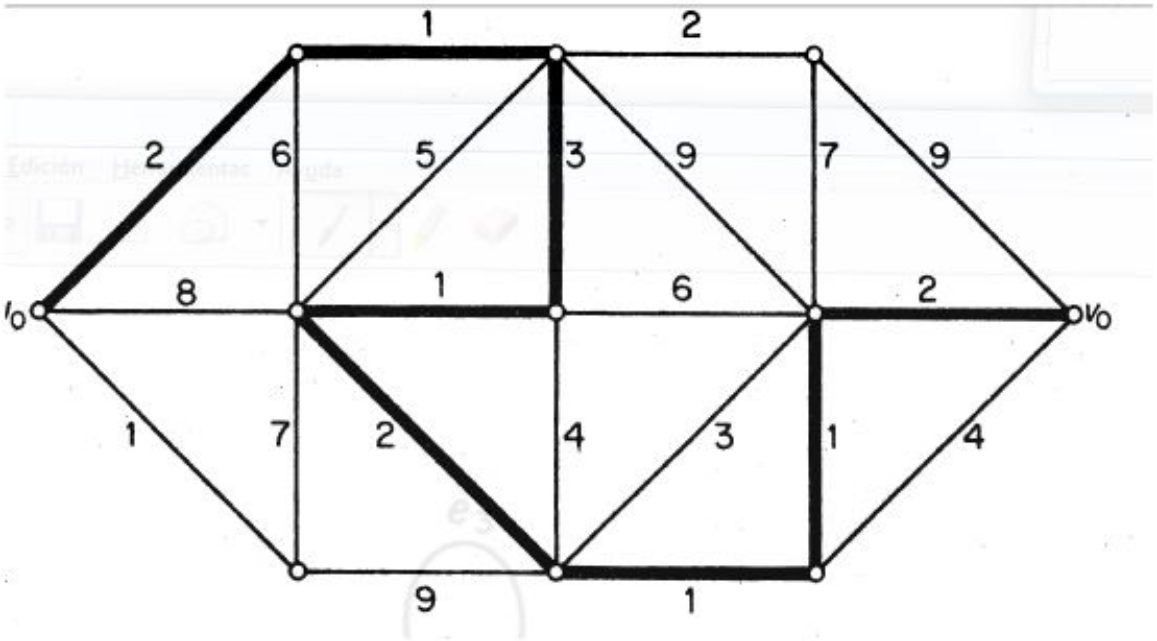


Figura: Camino mas corto

2. Problema de Conexión, Algoritmo de Kruskal.

Aplicaciones de la teoría de grafos II

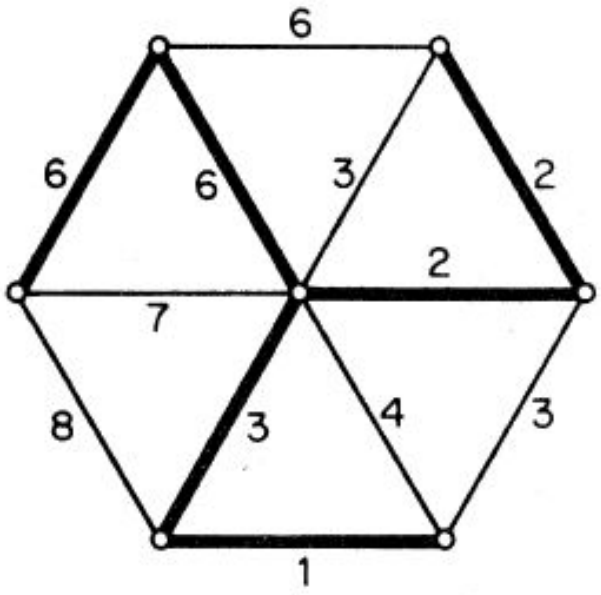


Figura: Árbol óptimo en un grafo ponderado

3. Construcción de redes de comunicación confiables. Teoria de conectividad.

Aplicaciones de la teoría de grafos III

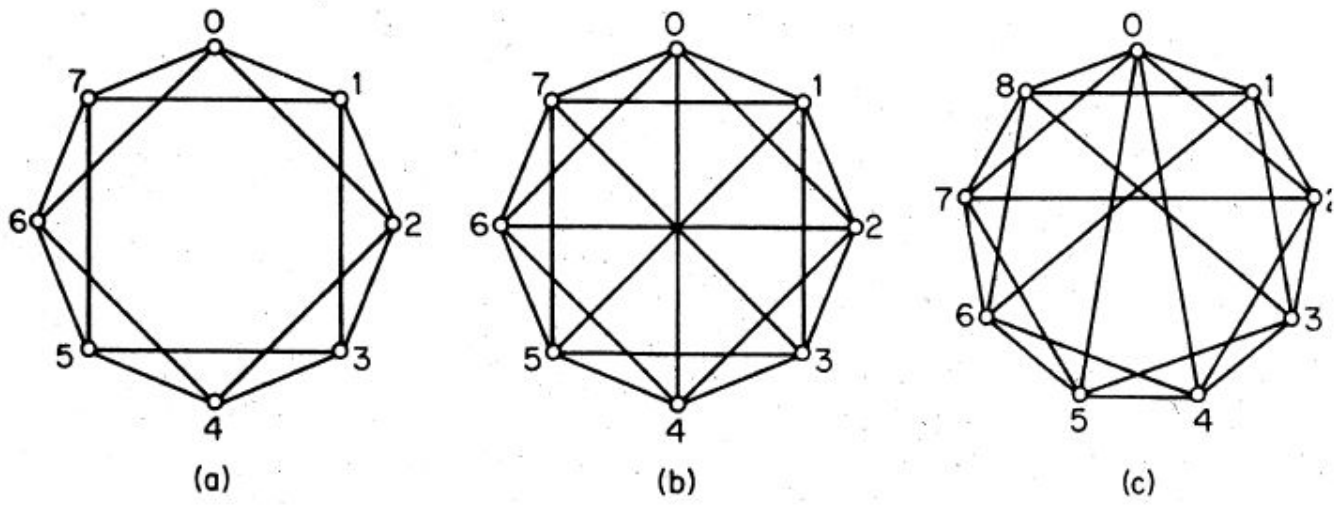


Figura: Tres casos de conectividad

4. El problema del cartero chino. Algoritmo de Fleury. Tour de Euler

Aplicaciones de la teoría de grafos IV

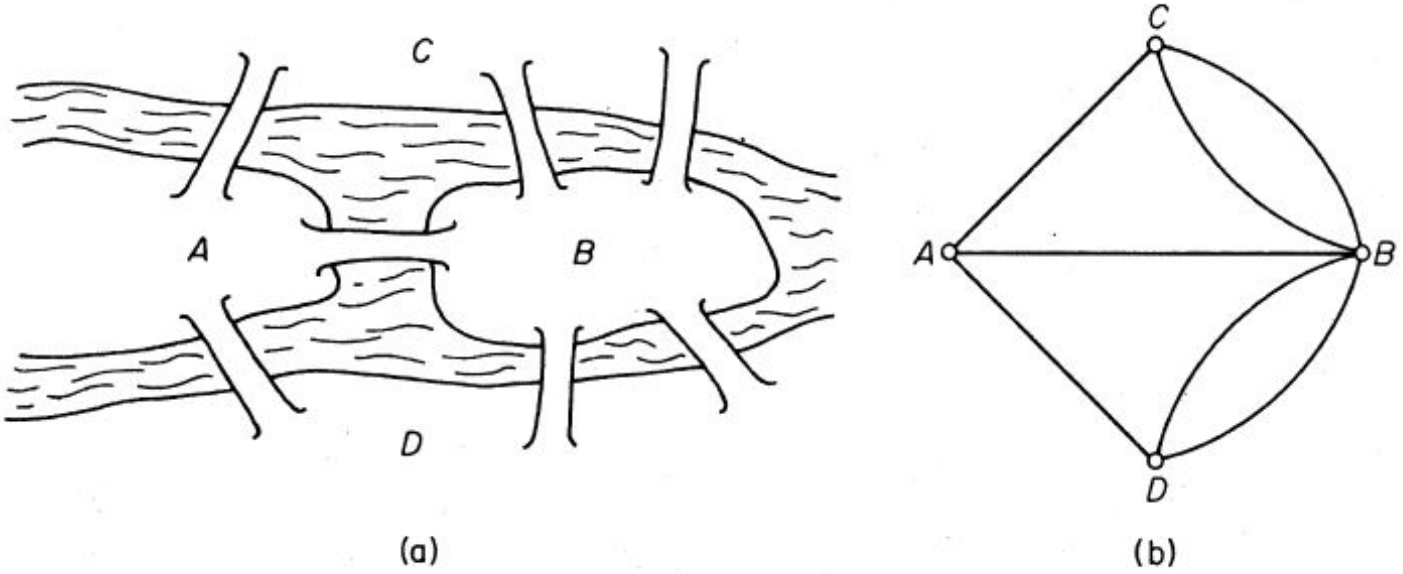


Figura: Puentes de Konigsberg y su grafo respectivo

5. El problema del Agente viajero. Circuito Hamiltoniano

Aplicaciones de la teoría de grafos V

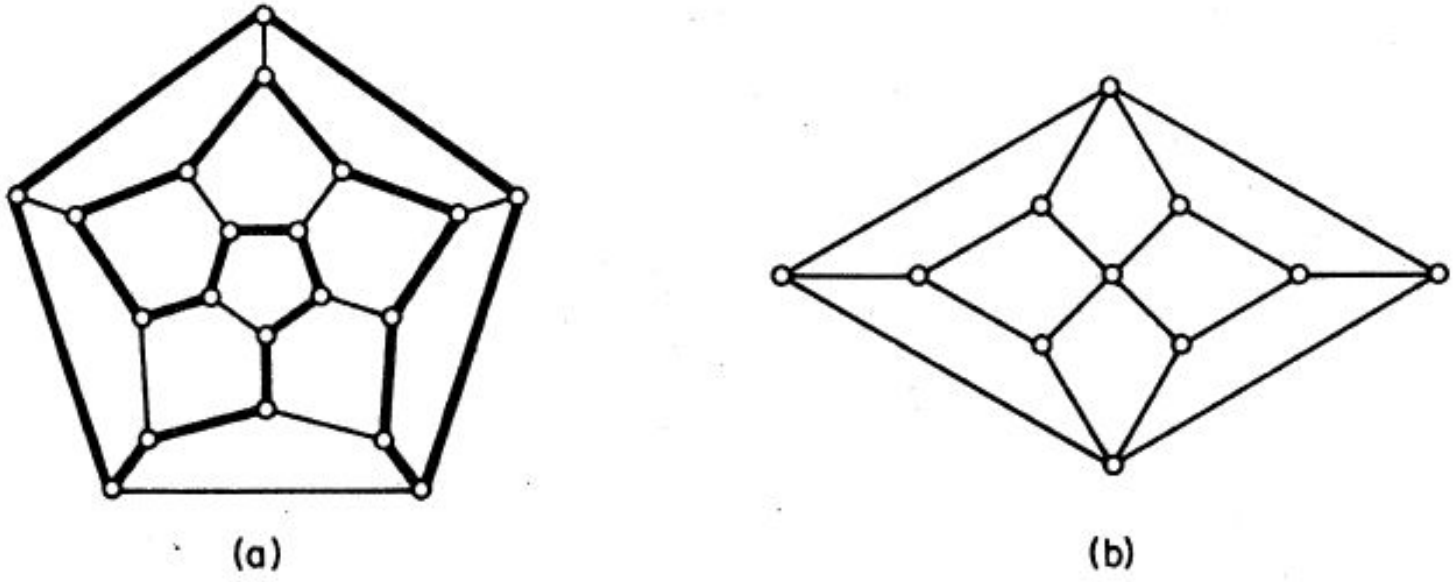


Figura: Dodecaedro y grafo de Herschel

6. El problema de asignación de personal. Algoritmo Hungariano. Matching de grafos

Aplicaciones de la teoría de grafos VI

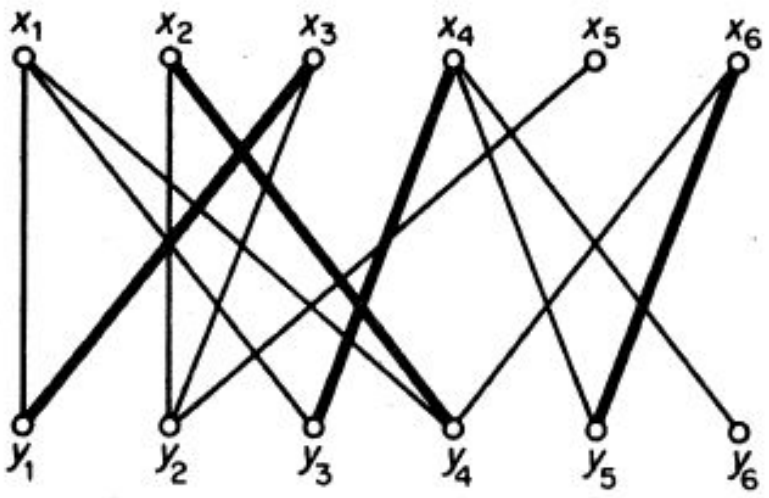


Figura: Matching de un grafo

7. El problema de los horarios. Coloración de aristas
8. El problema del almacenamiento. Coloración de vertices

Métodos de la Teoría de Grafos I

1. Aprendizaje no supervisado: Usa datos no etiquetados, es decir no se conoce la categoría a la que pertenecen. [2].

Métodos de la Teoría de Grafos

1. Clustering: Realiza una descripción de los datos (puntos d-dimensionales) en términos de clusters o grupos, usando algún criterio de similaridad (dist euclidiana, manhatan, canberra, etc) y una función criterio a optimizar (suma de los cuadrados del error, criterio de varianza minima, etc). [2]

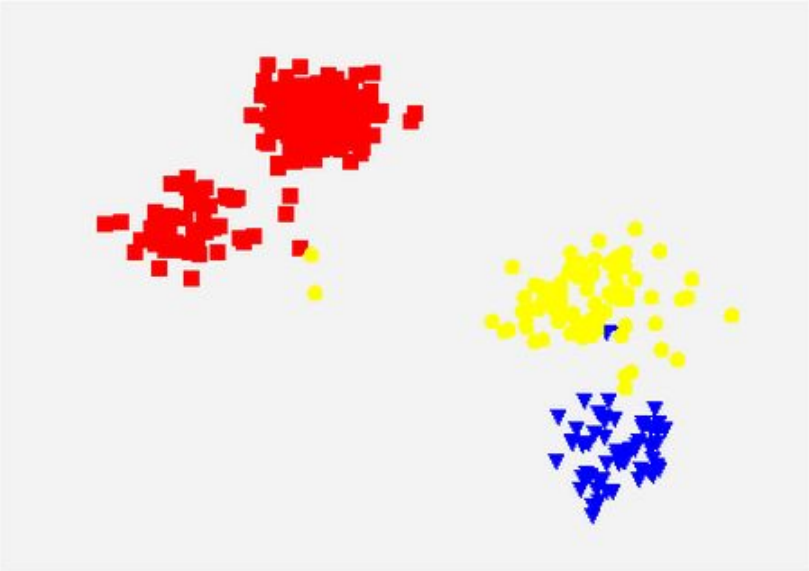


Figura: Clustering de datos bidimensionales

Matriz de similaridad

1. Matriz de Similaridad: Sea $S = [s_{ij}]$ la matriz de similaridad $n \times n$ definida por:

$$s_{ij} = \begin{cases} 1 & \text{si } s(x_i, x_j) > d_0 \\ 0 & \text{caso contrario} \end{cases} \tag{1}$$

donde d_0 es un valor umbral y $s(x_i, x_j)$ es una medida de similaridad para los puntos x_i e x_j

Grafo de Similaridad

1. Grafo de Similaridad: Sea el grafo $G = (V(E), E(G), \psi)$ inducido por la matriz de similaridad, donde los vértices corresponden a los puntos y las aristas unen los vértices i e j si y solo si $s_{ij} = 1$

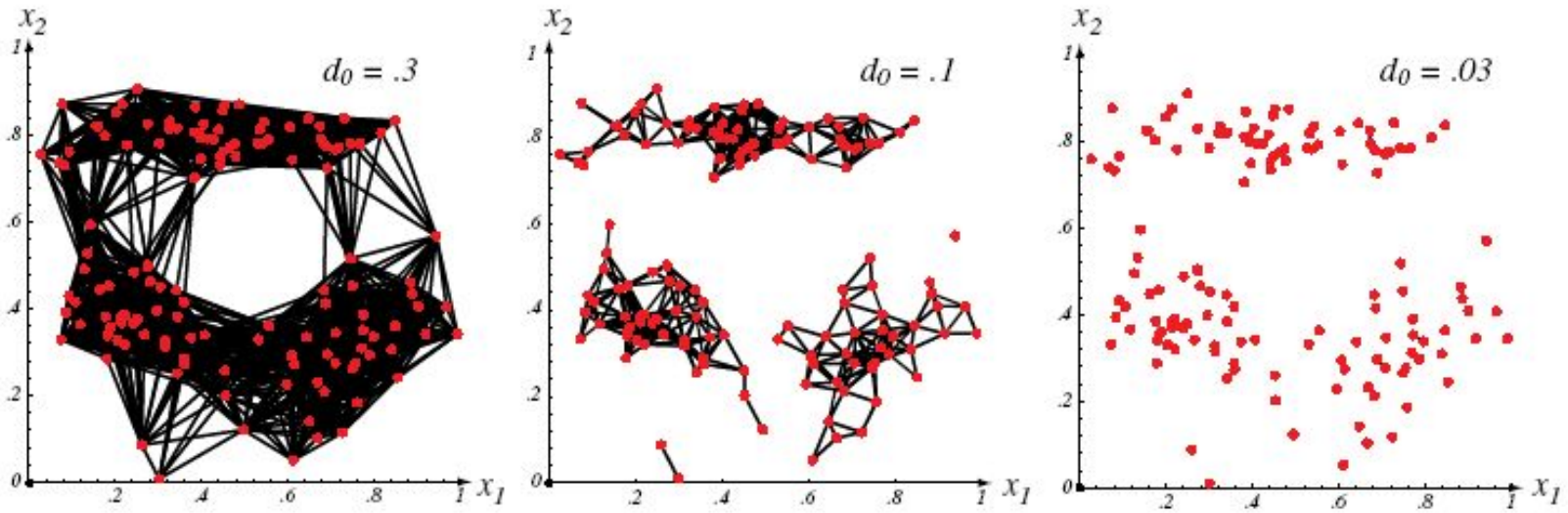
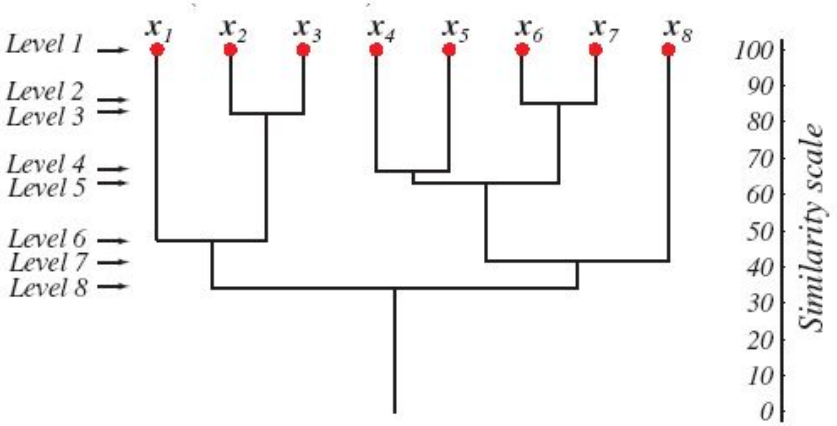


Figura: El valor umbral afecta al tamaño y número de clusters

Clustering Jerárquico Aglomerativo

Algoritmo-Clustering-Jerárquico

Begin Initialize c, $\hat{c} \leftarrow n$, $D_i \leftarrow X_i, i = 1, ..., n$
  do $\hat{c} \leftarrow \hat{c} - 1$
    Find nearest clusters, say D_i and D_j
    Merge D_i, and D_j
    Until $c = \hat{c}$
return c clusters



Clustering Jerárquico Aglomerativo

1. Single linkage algorithm usa la medida de distancia entre clusters:

$$d_{min}(D_i, D_j) = min_{x \in D_i, x' \in D_j} \|x - x'\| \tag{2}$$

Arbol de cobertura mínima en cada cluster. Dos puntos (vértices) digamos x y x' están en el mismo cluster si existe un camino $c = x, x_1, ..., x'$.

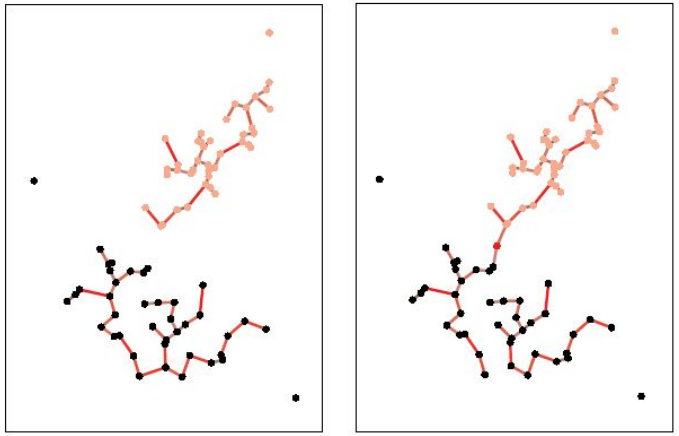


Figura: El algoritmo es sensitivo a los detalles

Clustering Jerárquico Aglomerativo

1. Complete linkage algorithm usa la medida de distancia entre clusters:

$$d_{max}(D_i, D_j) = max_{x \in D_i, x' \in D_j} \|x - x'\| \tag{3}$$

Subgrafos completos maximales del grafo de similaridad

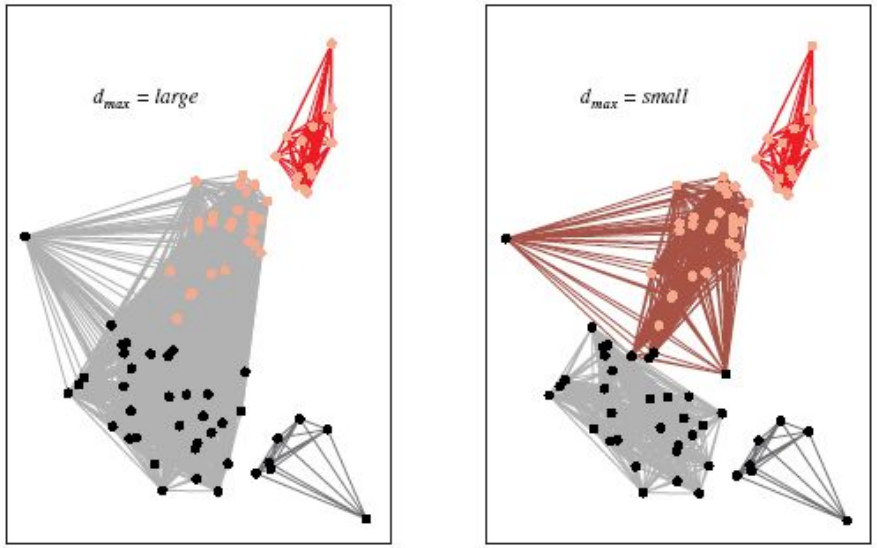


Figura: El número de clusters depende del valor umbral

Arista Inconsistente

1. Dado un árbol de cobertura mínima, remover la arista de mayor longitud, obteniendo dos clusters, luego la siguiente arista de mayor longitud y así de manera sucesiva. Otro enfoque es remover una arista inconsistente.

Arista Inconsistente

1. Arista Inconsistente: Sea l la longitud de una arista digamos e. Sea $\bar{l}$ la longitud promedio de todas las demás aristas incidentes a los vértices de la arista e. La arista e es inconsistente si l es significativamente mayor que $\bar{l}$, por ejemplo $l > 2\bar{l}$

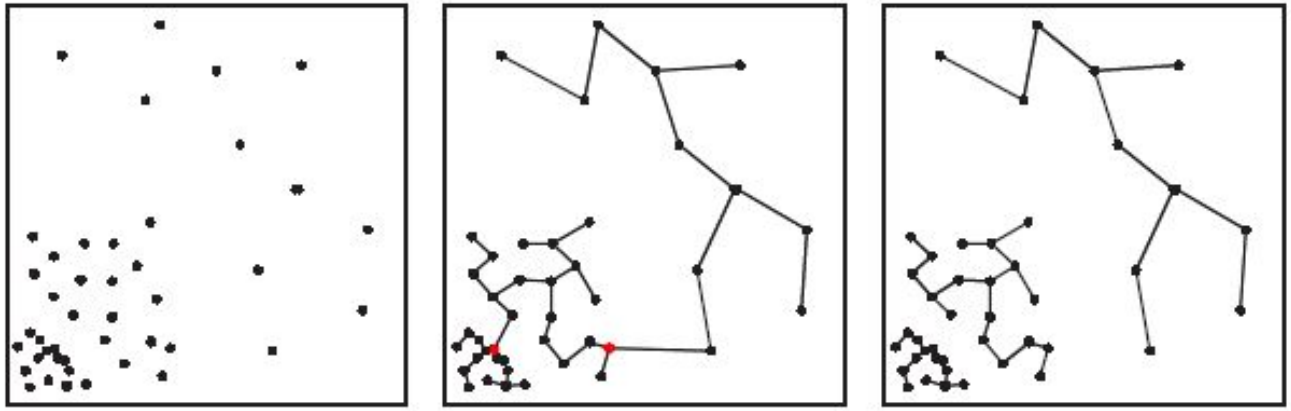


Figura: Datos originales, árbol de cobertura mínima, y clusters obtenidos eliminando aristas inconsistentes

Clasificación de Documentos Web usando un Modelo de Grafo [5] I

1. Medida de similaridad

$$d_{MCS}(G_1, G_2) = 1 - \frac{|msc(G_1, G_2)|}{max(|G_1|, |G_2|)} \tag{4}$$

G_1 y G_2 son grafos, $msc(G_1, G_2)$ es el grafo común máximo.

Clasificación de Documentos Web usando un Modelo de Grafo [5] II

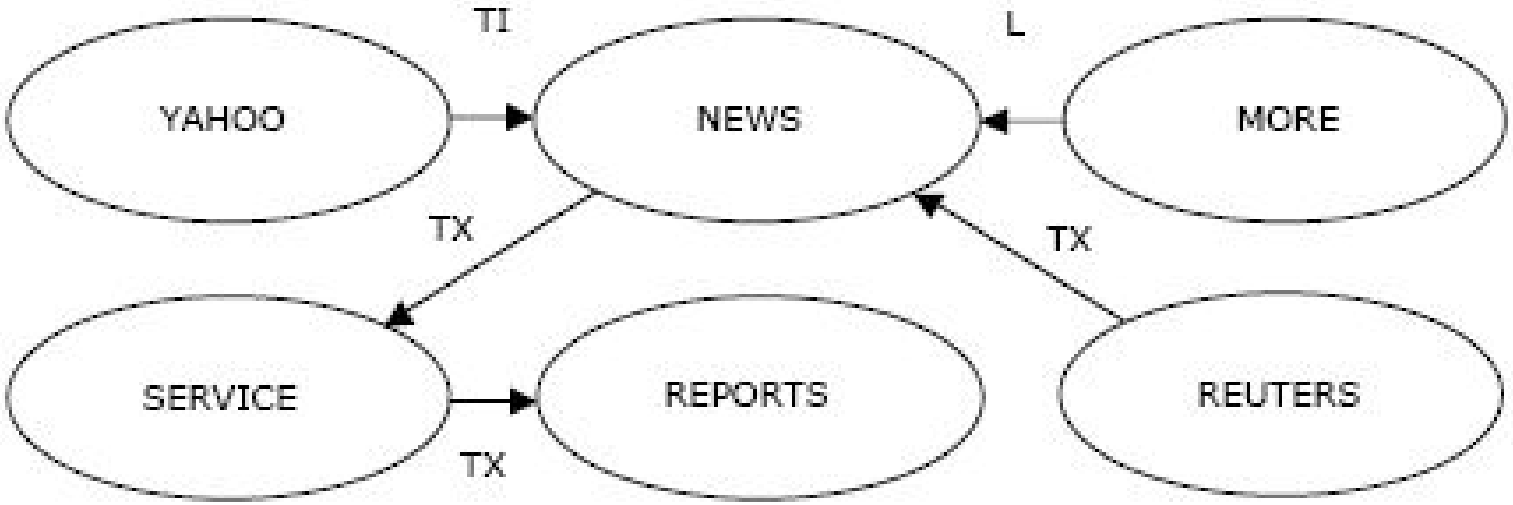


Figura: Representación de un documento mediante un grafo

Clasificación de Documentos Web usando un Modelo de Grafo [5] III

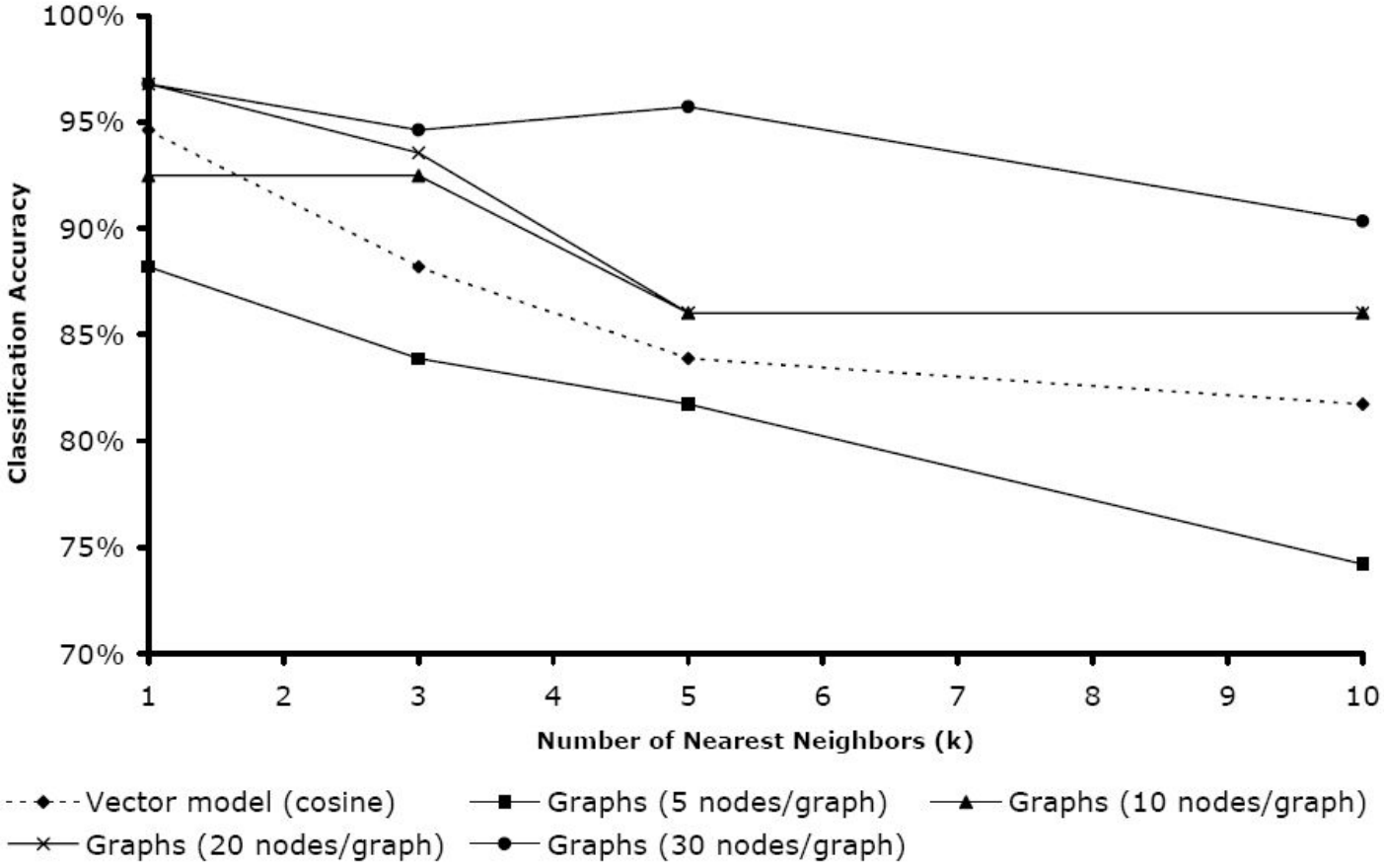


Figura: Resultados

Clustering Incremental de Documentos Basados en Grafos [4] I

1. Modelo de representación de documentos basado en Grafos: Grafo dirigido $G = (V(E), E(G), \psi)$, $V(G)$ representa a las palabras del documento, $E(G)$, representa el orden de las palabras en el documento, por ejemplo la arista $e = uv$, representa una conexión directa de la palabra u hacia la palabra v.

Grafo de índice de Documentos [3] I

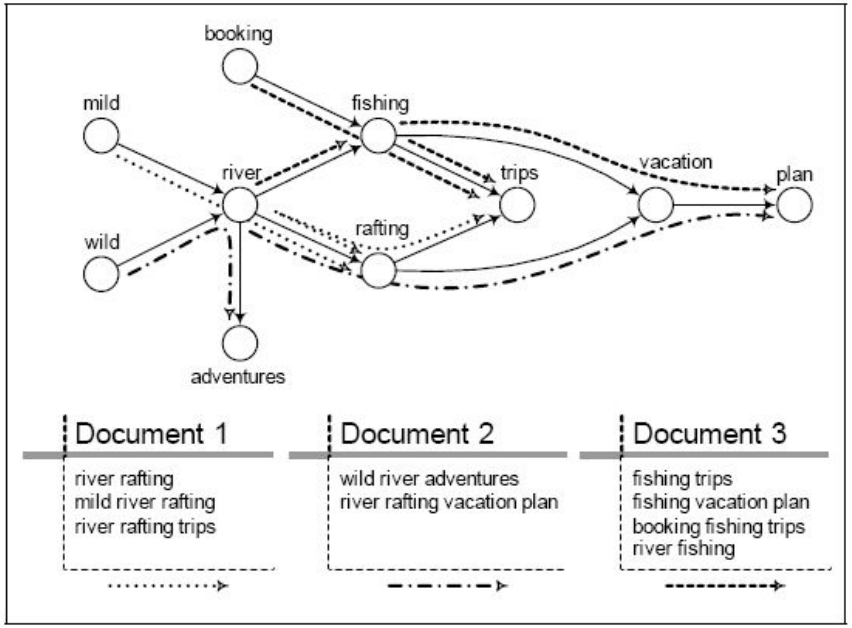


Figura: Grafo de índice de documentos

Grafo de índice de Documentos [3] I

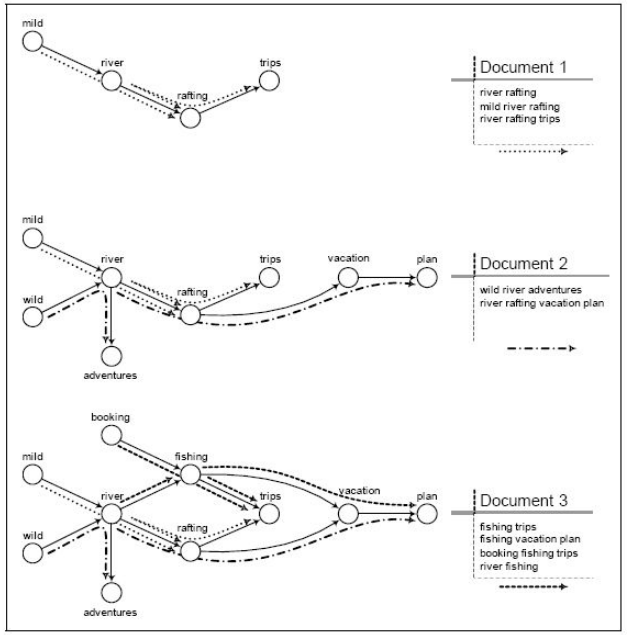


Figura: Construcción incremental del grafo de índice de documentos

Medida de similaridad I

$$sim(d_1, d_2) = \lambda sim_{df}(d_1, d_2) + (1 - \lambda) sim_{sp}(d_1, d_2) \tag{5}$$

$\lambda \in [0, 1]$

$$sim_{sp}(d_1, d_2) = \frac{\sqrt{\sum_{i=1}^{p} (\frac{l_i}{avg(|s_i|)})(f_{1i} + f_{2i})^2}}{\sum_j |s_{1j}| + \sum_k |s_{2k}|} \tag{6}$$

P=número de frases compartidas, f_{1i}, f_{2i} son las frecuencias de las frases compartidas i en los documentos d_1, d_2, l_i longitud de la frase, $|s_{ij}|$ longitud de la sentencia j en el documento d_i, $avg(|s_i|)$ longitud promedio de las sentencias conteniendo las frases compartidas i.

El proceso de clustering fué realizado con una modificación del algoritmo incremental DBSCAN [4]

Clustering de genes y analisis de expresion de metagenes basados en grafos I

1. Gen: secuencia organizada (código genético) de nucleótidos en la molécula de ADN o ARN en el caso de algunos virus. Unidad de herencia de organismos vivos. y reside en una extensión de DNA.

Clustering de genes y analisis de expresion de metagenes basados en grafos II

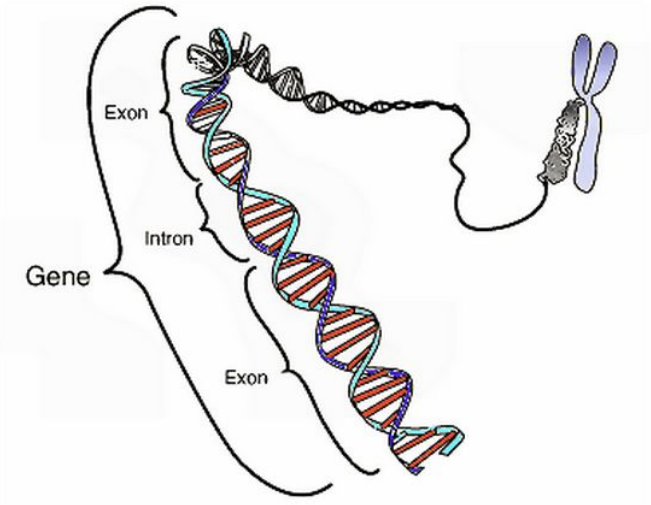


Figura: gen

Expresión génica I

1. Expresión génica: Proceso en el cual la información de un gene es usada para sintetizar proteinas, ribosomal RNA (rRNA genes) ó tRNA (tRNA genes).

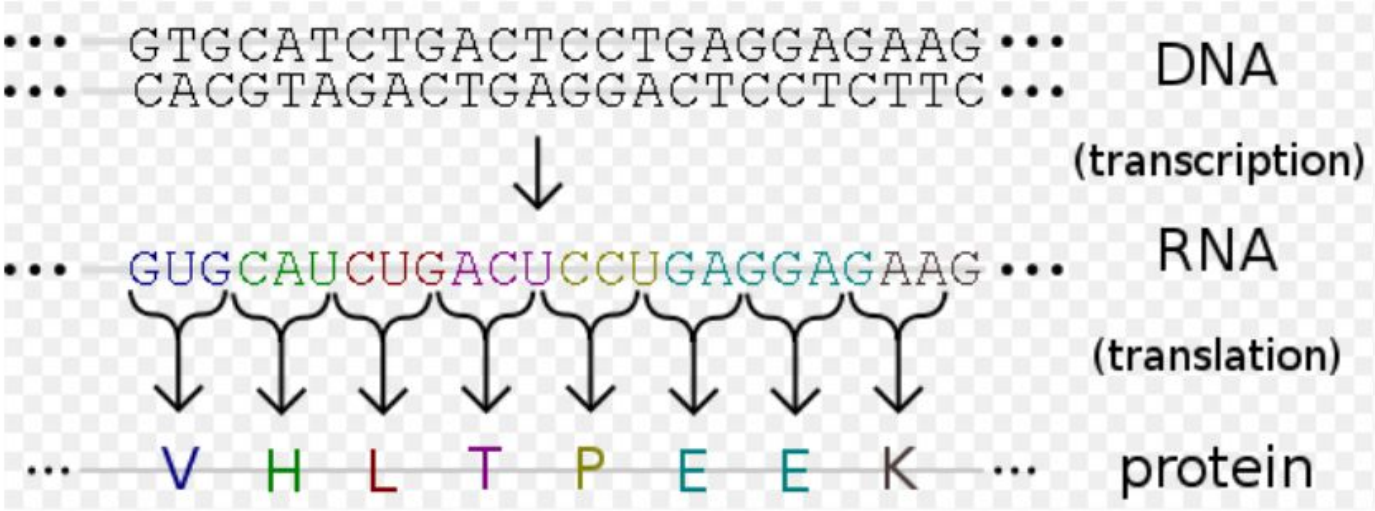


Figura: Proceso de transcripción y traslación

Microarray I

1. Microarray: Es una tecnología para medir cambios en los niveles de expresión génica

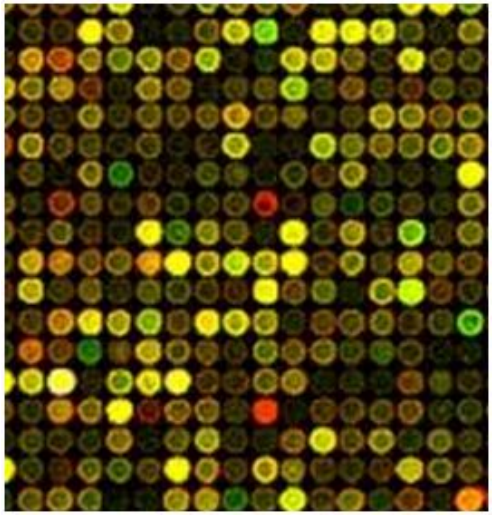


Figura: microarray

Aplicación I

1. 158 muestras de cancer de mama (Koo Foundation Sun Yat-Sen Cancer Center).
2. K-means y MetageneCreator son usados para construir los clusters y encontrar los metagenes asociados a cada cluster

Aplicación I

1. Grafo de Indenpendencia: $G = (V(E), E(G), \psi)$, donde cada gen es asociado a un vértice del grafo, y los elementos de $E(g)$ son los elementos diferentes de la diagonal y diferentes de cero de la matriz inversa de covarianza $\Omega = \Sigma^{-1}$), así dos genes están conectados si se cree que existe asociación.
2. Modelos gráficos gaussianos : $M = (\Sigma, G)$

Aplicación II

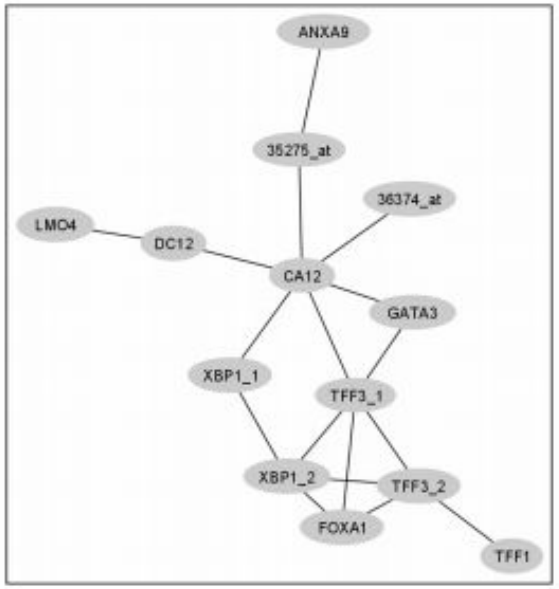
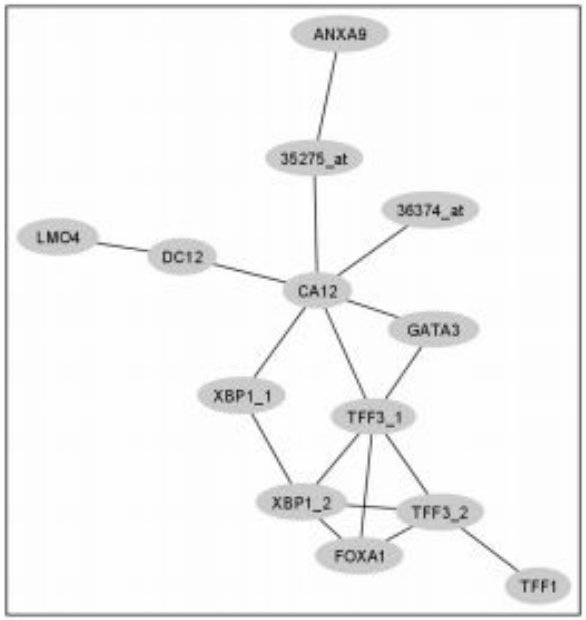
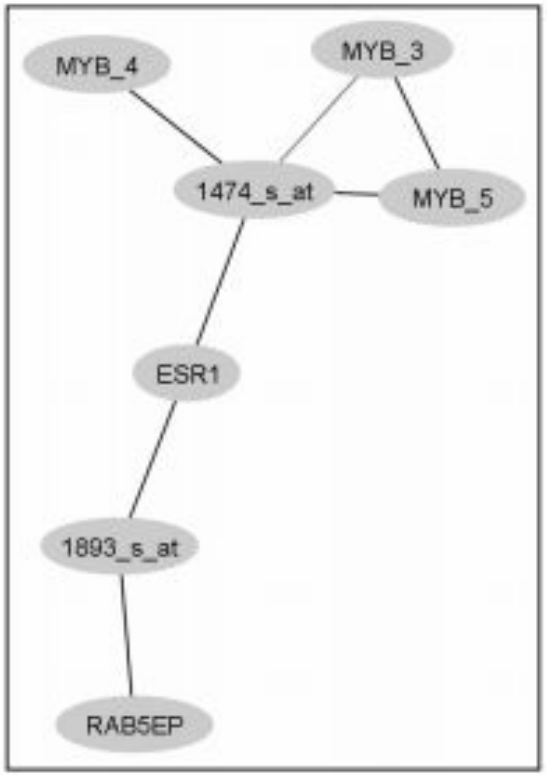


Figura: genes en el cluster 438 y la red de asociaciones obtenida usando modelos gráficos gaussianos

Aplicación III



(a) Genes en el cluster 438 y la red de asociaciones obtenida usando modelos gráficos gaussianos



(b) Genes en el cluster 398 y la red de asociaciones obtenida usando modelos gráficos gaussianos

John Adrian Bondy.
Graph Theory With Applications.
Elsevier Science Ltd, 1976.

Richard O. Duda, Peter E. Hart, and David G. Stork.
Pattern Classification (2nd Edition).
Wiley-Interscience, 2 edition, November 2001.

Khaled M. Hammouda and Mohamed S. Kamel.
Phrase-based document similarity based on an index graph model.
In *In Proceedings of the 2002 IEEE Int'l Conf. on Data Mining (ICDM'02*, pages 203–210, 2002.

Tu-Anh Nguyen-Hoang, Kiem Hoang, Danh Bui-Thi, and Anh-Thy Nguyen.
Incremental document clustering based on graph model.

In *Proceedings of the 5th International Conference on Advanced Data Mining and Applications*, ADMA '09, pages 569–576, Berlin, Heidelberg, 2009. Springer-Verlag.

Adam Schenker, Mark Last, Horst Bunke, and Abraham Kandel.
Classification of web documents using a graph model.
In *Seventh International Conference on Document Analysis and Recognition*, pages 240–244, 2003.