CAPÍTULO 4

CLASIFICACIÓN DE DOCUMENTOS CON *BLINDLIGHT*

Un primer mecanismo para enfrentarse a una colección muy grande de documentos es su clasificación, es decir, su división en grupos de documentos más pequeños y homogéneos que permitan deducir la estructura subyacente a la colección y faciliten su exploración. El problema de la clasificación no supervisada de colecciones de documentos no es nuevo y los beneficios que aporta al campo de la recuperación de información son bien conocidos. Existe una gran variedad de métodos de clasificación y se dispone de técnicas que permiten evaluar la calidad de las clasificaciones obtenidas, ya sea estudiando características internas de las mismas o comparándolas con clasificaciones "externas". En este capítulo se repasarán brevemente algunas de las principales técnicas de clasificación automática de documentos para, seguidamente, presentar la forma en que es posible aplicar la técnica propuesta por el autor a este problema. A continuación, se describirá una serie de experimentos cuyos resultados serán comparados con los obtenidos con otras técnicas "convencionales" demostrando que, efectivamente, blindLight es capaz de obtener resultados semejantes e incluso mejores que los alcanzados por métodos ad hoc.

1 El problema de la clasificación

El problema de clasificar de manera no supervisada un conjunto de patrones, *a priori* grande, en un número reducido de grupos (*clusters*) que exhiban características similares es conocido como *clustering*, clasificación no supervisada o agrupamiento. Este problema se manifiesta en multitud de campos, incluyendo la recuperación de información[1], y admite varias aproximaciones. Sin embargo, no es objetivo de este trabajo analizar de manera

[1] La aplicación de técnicas de clasificación a la recuperación de información se basa en la denominada *cluster hypothesis* expuesta originalmente por Jardine y van Rijsbergen (1971, p. 219): *"es intuitivamente plausible que la asociación entre documentos proporciona información acerca de la relevancia de los documentos respecto a las peticiones"* y posteriormente reformulada del siguiente modo: *"documentos estrechamente asociados tienden a ser relevantes para las mismas peticiones"* (van Rijsbergen 1979).

exhaustiva las distintas alternativas para llevar a cabo clasificación de documentos puesto que existe amplia bibliografía sobre el tema. Para adquirir una visión amplia del campo son muy recomendables tanto el tercer capítulo de *"Information Retrieval"* (van Rijsbergen 1979) como la revisión hecha por Jain, Murty y Flynn (1999). Los capítulos cuarto, y en menor medida el quinto y sexto, de *"Mining the Web"* (Chakrabarti 2003) están dedicados al problema de extraer información de colecciones de documentos hipertextuales.

No obstante, puesto que el autor afirma en su tesis que la técnica que propone facilita *"la clasificación […] de documentos […] con resultados similares a los de otros métodos […]"* antes de describir la forma en que es posible aplicar *blindLight* a este problema y los resultados obtenidos es necesario hacer un brevísimo repaso de las distintas técnicas disponibles para la clasificación de documentos así como de algunas colecciones habitualmente empleadas para probar tales métodos.

1.1 Clasificación de documentos

Un método de clasificación requiere, básicamente, (1) un modo de representar los documentos, (2) una medida de similitud entre dichas representaciones y (3) un algoritmo para construir los grupos de documentos basándose en la medida anterior.

La forma de representar los documentos para su clasificación y, de hecho, cualquier patrón es, habitualmente, *"un vector multidimensional donde cada dimensión corresponde a una única característica"* (Duda y Hart 1973, citado por Jain *et al.* 1999, p. 270). Como ya se dijo anteriormente, el modelo vectorial (ya sean los pesos binarios o reales) facilita un modo de representación de documentos muy conveniente para la implementación de métodos de clasificación.

En cuanto a las medidas de similitud ya se han mostrado varias (p.ej. los coeficientes de Dice o Jaccard o la función del coseno). Según Lerman (1970, citado por van Rijsbergen 1979, p. 30) muchas de estas medidas son monótonas entre sí[1]. Por tanto, aquellos métodos de clasificación que únicamente empleen el orden establecido entre los documentos, es decir, la mayor parte, obtendrán clasificaciones idénticas con independencia de la medida de similitud empleada.

Así, de los tres pasos necesarios para clasificar documentos (representación, cálculo de similitudes y construcción de los grupos) este apartado se centrará únicamente en el tercero revisando escuetamente algunas de las distintas aproximaciones posibles.

Jain *et al.* (1999) proporcionan una posible taxonomía de métodos de clasificación que, inicialmente, pueden diferenciarse en dos grandes grupos: los jerárquicos que producen un conjunto de particiones anidadas y los particionales que producen una única partición. En segundo lugar existen una serie de características "transversales" que afectan a ambos tipos. Así, por ejemplo, los algoritmos pueden ser (a) aglomerativos o divisivos[2] según comiencen con tantos grupos como documentos que van siendo "fusionados" o lo hagan con un único grupo que es dividido. (b) Exactos si un documento es asignado a un único grupo o borrosos (*fuzzy*) si existen grados de "pertenencia". (c) Deterministas si siempre se obtiene la misma clasificación para un conjunto de partida o estocásticos si se emplean métodos aleatorios para llegar al resultado final. Y (d) incrementales o no incrementales en

[1] Dos funciones son monótonas entre sí cuando al establecer una relación de orden sobre un conjunto de elementos ambas dan lugar a la misma ordenación.

[2] Ascendentes y descendentes según Chakrabarti (2003, p. 84)

función del número de documentos a clasificar[1]. Así pues, existe una flexibilidad enorme a la hora de especificar un algoritmo de clasificación y en consecuencia existen muchísimos métodos y sutiles variantes de los mismos. A continuación se presentarán algunos de los más conocidos.

Los métodos aglomerativos comienzan situando cada documento en su propio grupo y, de forma iterativa, van agrupando grupos en funcion de su similitud. Puesto que en cada iteración existe un número distinto de grupos y éstos están constituidos a su vez por grupos más pequeños este tipo de algoritmos son jerárquicos. Las dos principales variantes de este tipo de algoritmos son las denominadas *single-link* (Sneath y Sokal 1973) y *complete-link* (King 1967). En el primer caso la distancia entre dos grupos es el mínimo de las distancias entre todos los pares formados al emparejar un elemento del primer grupo con uno del segundo. En el caso del método *complete-link* esta distancia no es el mínimo sino el máximo. En ambos casos se agrupan iterativamente aquellos grupos que se encuentran a una distancia menor.

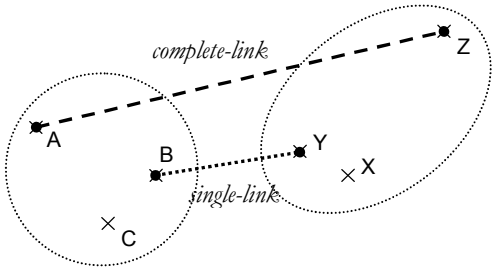


Fig. 50 Distancia entre grupos empleando *complete-link* y *single-link*.

Un segundo método muy conocido es el algoritmo de las *k*-medias que es particional, divisivo y, en su versión más simple, exacto (Duda y Hart 1973, citado por Jain *et al.* 1999). Este método clasifica una colección de patrones en *k* grupos donde el valor de *k* se establece *a priori*. Cuando se aplica a clasificación de documentos estos son representados como vectores y los grupos como el centroide de los documentos pertenecientes al mismo. En primer lugar deben establecerse *k* centroides de manera "arbitraria" asignándose cada documento de la colección al centroide más próximo. Una vez se han asignado todos los documentos se recalcula el centroide de cada grupo y se lleva a cabo una nueva fase de asignación. Este proceso se repite hasta que los cambios en los grupos obtenidos son mínimos. Una alternativa a este método es conocida como *k*-medoides donde no se utiliza el centroide del grupo sino aquel documento del mismo más próximo a éste.

Otro algoritmo interesante es el propuesto por Jarvis y Patrick (1973). Este método no sólo tiene en cuenta los documentos más próximos a uno dado sino también los vecinos que tienen en común, razón por la que también se denomina como método de "vecinos comunes" (*shared nearest neighbor clustering*). Este algoritmo requiere dos parámetros, *J* y *K*, donde *J* es el tamaño de la lista de vecinos para un punto dado y *K* es el número de vecinos comunes necesarios para formar un grupo. Según el método de Jarvis-Patrick dos documentos pertenecerán a un mismo grupo si ambos son vecinos y tienen, al menos, *K* vecinos en común.

---

[1] *"La clasificación incremental parte del supuesto de que es posible considerar los patrones de uno en uno y asignarlos a algún grupo ya disponible"* (Jain *et al.* 1999, p. 32) por lo que está especialmente indicada para colecciones muy grandes.

Además de métodos como los anteriores se han aplicado con éxito técnicas de computación flexible y probabilísticas. Por ejemplo, los Mapas Auto-Organizativos[1] (*Self-Organizing Maps* o *SOM*) se han utilizado en los proyectos *WEBSOM* (Honkela *et al.* 1996) y *SOMLib* (Rauber y Merkl 1999) y se han desarrollado métodos de clasificación jerárquicos y aglomerativos basados en la probabilidad condicionada de Bayes (Iwayama y Tokunaga 1995).

1.2 Evaluación de métodos de clasificación

Según van Rijsbergen (1979, p. 23) la clasificación en el campo de la recuperación de información se hace con un propósito y, por tanto, su bondad sólo puede medirse sobre la base del rendimiento en la fase de recuperación. De este modo, van Rijsbergen evita debatir acerca de las denominadas clasificaciones "naturales", aquellas similares a las que producirían de manera independiente distintos seres humanos.

No obstante, no es estrictamente necesario esperar a la fase de recuperación de información para evaluar un método de clasificación automática. De hecho, si se dispone de documentos que ya se han clasificado previamente (tal vez de manera manual) es posible calcular una serie de medidas para determinar la calidad de la clasificación automática: la **entropía**[2] y la **medida *F***[3] (Steinbach, Karypis y Kumar 2000) o la **pureza**[4] (Zhao y Karypis 2002). Por otro lado, en caso de no disponer de una clasificación previa puede calcularse la **similitud promedio**[5] (*overall similarity*) (Steinbach *et al.* 2000).

En cuanto a las colecciones de documentos que se utilizan con mayor frecuencia para evaluar nuevos métodos de clasificación podrían destacarse la colección de artículos de la revista *TIME*, las habituales *CACM*, *CISI*, *LISA* y *Cranfield*[6], diversas particiones

---

[1] Los Mapas Auto-Organizativos (Kohonen 1982) agrupan una serie de vectores de entrada sobre un "mapa", una red neuronal generalmente bidimensional o tridimensional, en el cual vectores "similares" aparecen en posiciones cercanas.

[2] La entropía es un criterio de evaluación externo puesto que depende de una clasificación previa con la que comparar la solución obtenida. Dado un grupo *j*, su entropía es E_j (véase ecuación al final de la nota) donde p_{ij} es la probabilidad de que un elemento de dicho grupo pertenezca a la clase *i*. Por su parte, la entropía del agrupamiento será la media ponderada de la entropía de todos los grupos (en función de la proporción entre el número de documentos del grupo y el total). $E_j = -\sum_i p_{ij} \cdot \log(p_{ij})$

[3] La medida *F* se utiliza habitualmente para evaluar sistemas de recuperación de información (véase pág. 139) y combina en un valor único las consabidas precisión y exhaustividad. En el caso de la evaluación de clasificaciones automáticas esta medida permite evaluar soluciones jerárquicas (al contrario que la entropía y la pureza que tan sólo sirven para soluciones sin grupos anidados). Para ello, se calculan los valores F_{ij} para cada grupo *j* y clase externa *i* entendiendo que la precisión es la fracción de documentos del grupo *j* que pertenecen a la clase *i* mientras la exhaustividad es la fracción de documentos de la clase *i* que aparecen en el grupo *j*. Posteriormente se determina el máximo valor *F* para cada clase y, por último, se calcula la media ponderada de estos valores.

[4] La pureza también permite evaluar una solución de agrupamiento por medio de una clasificación externa. No es más que la proporción entre el número de *ítems* pertenecientes a la clase dominante en un grupo y el tamaño de dicho grupo. Es decir, la pureza evalúa en qué medida un grupo de una clasificación automática contiene elementos de una única clase.

[5] La similitud promedio es una medida de evaluación interna y que, por tanto, no requiere ninguna clasificación con la que comparar la solución de agrupamiento. Se trata tan sólo de la similitud media entre cada par de documentos de un grupo.

[6] Se trata de colecciones para la evaluación de sistemas de recuperación de información pero que también se han usado para evaluar sistemas de clasificación automática puesto que los documentos están asignados a categorías predefinidas. La colección *CACM* consta de 3.204 artículos (título y resumen) publicados en

empleadas en las conferencias *TREC*[1], la colección de textos médicos *OHSUMED* (Hersh *et al.* 1994) y varias colecciones de artículos de la agencia *Reuters*[2].

2 Utilización de *blindLight* para la clasificación automática de documentos

La aplicación de *blindLight* a la clasificación automática de documentos es muy sencilla: los documentos se representan mediante vectores de *n*-gramas tal y como fueron descritos en el capítulo anterior mientras que la medida de similitud interdocumental es la denominada *PiRo* (véase ecuación 12 en página 66). Por lo que respecta a los métodos para obtener los grupos de documentos se han implementado dos algoritmos particionales, uno no incremental y otro incremental. En el primer caso se utiliza, además de la similitud interdocumental, información sobre los documentos "vecinos" de manera similar a la de Jarvis y Patrick (1973).

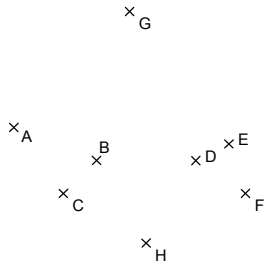


Fig. 51 Conjunto de puntos de ejemplo.

2.1 Algoritmo no incremental basado en *blindLight*

Una de las ideas más interesantes del algoritmo de clasificación de Jarvis y Patrick es la utilización de listas de vecinos "compartidos" para agrupar los patrones. En la implementación del método de clasificación *blindLight* no incremental se ha aplicado una idea similar pero sin recurrir a ningún tipo de lista de vecinos y utilizando en cambio toda la información disponible en la matriz de similitudes. Dicha matriz se construye en la fase inicial del algoritmo y almacena los valores *PiRo* para cada posible par de documentos[3]. De este modo, para cada documento D_i se dispone de un vector que contiene la similitud de D_i respecto al resto de documentos de la colección (véase Fig. 52). Estos vectores pueden asimilarse con el "comportamiento" de cada documento dentro de la colección y permitirían agrupar aquellos documentos que no sólo son similares entre sí sino que se "comportan" de manera similar respecto al resto de documentos de la colección.

$$A = \{(B:0{,}18),(C:0{,}15),(D:0{,}05),(E:0{,}04),(F:0{,}04),(G:0{,}05),(H:0{,}06)\}$$
$$B = \{(A:0{,}18),(C:0{,}31),(D:0{,}10),(E:0{,}05),(F:0{,}04),(G:0{,}03),(H:0{,}10)\}$$
$$C = \{(A:0{,}15),(B:0{,}31),(D:0{,}06),(E:0{,}04),(F:0{,}04),(G:0{,}03),(H:0{,}15)\}$$

Fig. 52 Vectores de similitudes para los puntos A, B y C de Fig. 51.

---

la revista *Communications of the ACM* entre 1.958 y 1.979. *CISI* incluye 1.460 artículos (título y resumen) compilados en el *Institute for Scientific Information*. *LISA* es una colección de 6.004 documentos (resúmenes) extraídos de la base de datos *Library and Information Science Abstracts* y, por último, *Cranfield* consta de 1400 documentos y es uno de los resultados del proyecto *Cranfield II* (véase página 140).

[1] `http://trec.nist.gov/`

[2] *Reuters Corpus* `<http://about.reuters.com/researchandstandards/corpus>` y *Reuters-21578* `<http://www.daviddlewis.com/resources/testcollections/reuters21578>`.

[3] Prescindiendo de los pares repetidos y aquellos que involucran al mismo documento.

A fin de determinar qué documentos exhiben un "comportamiento" análogo, esto es, tienen vectores de similitudes parecidos se ha empleado una medida que permite comparar las poblaciones de distintos ecosistemas, el denominado coeficiente de Bray-Curtis[1] (Bray y Curtis 1957, citado por Gauch 1982):

$$C_Z = \frac{2w}{a+b} \qquad (1)$$

En este coeficiente (véase la ecuación 1) *a* es la suma de las poblaciones de todas las especies en el primer ecosistema, *b* es la suma de las poblaciones en el segundo ecosistema y *w* es la suma de la población menor para cada especie presente en ambos ecosistemas (en Fig. 54 se muestra un ejemplo ilustrativo).

B={(*A*: 0,18), (*C*: 0,31), (*D*: 0,10), (*E*: 0,05), **(*F*: 0,04)**, **(*G*: 0,03)**, **(*H*: 0,10)**}
C={**(*A*: 0,15)**, (*B*: 0,31), **(*D*: 0,06)**, **(*E*: 0,04)**, **(*F*: 0,04)**, **(*G*: 0,03)**, (*H*: 0,15)}

a	0,18+0,31+0,10+0,05+0,04+0,03+0,10	**0,81**
b	0,15+0,31+0,06+0,04+0,04+0,03+0,15	**0,78**
w	0,15+0,06+0,04+0,04+0,03+0,10	**0,42**
C_Z	*2w/(a+b)*	**0,53**

Fig. 53 Cálculo del coeficiente de Bray-Curtis para los vectores de similitudes de los puntos B y C (véase Fig. 52).

Este coeficiente se puede aplicar al problema que nos ocupa de forma inmediata (véase Fig. 53) y permite obtener para cada pareja de documentos una nueva medida de similitud que es el producto de *PiRo* y C_Z. Posteriormente se determina un umbral adecuado para estos valores y aquellos pares de documentos cuya similitud $PiRo \cdot C_Z$ supere dicho umbral son incluidos en el mismo grupo.

Puesto que este algoritmo, véase Fig. 55, requiere $(n^2 - n)/2$ comparaciones tanto de documentos como de vectores de similitudes no es adecuado para clasificar colecciones demasiado grandes. Sin embargo, puesto que puede transformarse fácilmente en un algoritmo aglomerativo que emplea la matriz de similitudes $PiRo \cdot C_Z$ permite elaborar dendrogramas como los que se muestran en este capítulo y el anterior.

[1] Las principales características de este índice que lo hacen idóneo para este problema son dos: en primer lugar la ausencia del mismo componente en ambos vectores no se tiene en cuenta como "un punto en común" y en segundo los componentes mayores son los que dominan el coeficiente.

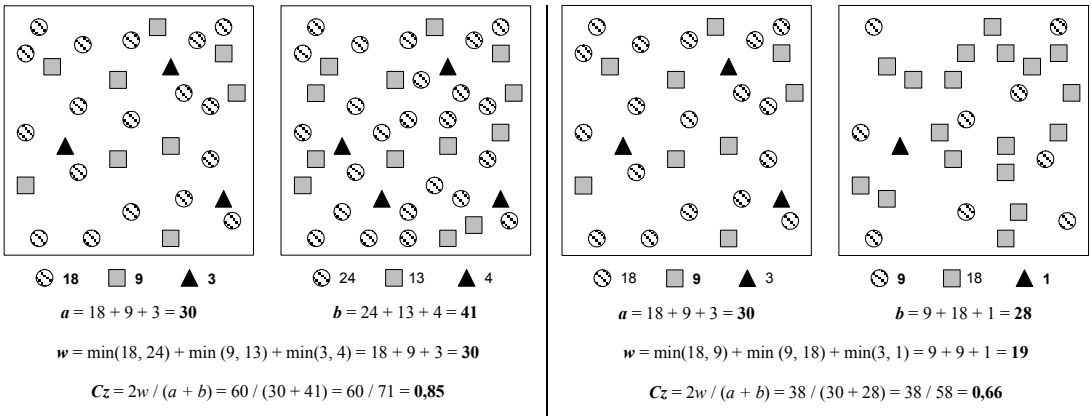


Fig. 54 Utilización del coeficiente de Bray-Curtis para la comparación de dos ecosistemas.

Se muestran aquí dos ejemplos de la evolución de un ecosistema en el que habitan 3 especies: los círculos rayados, los cuadrados y los triángulos. Los primeros son una especie vegetal que sirve de alimento a los cuadrados que son, a su vez, presa de los triángulos. A izquierda y derecha se muestra la evolución del ecosistema desde la misma situación de partida en la que hay 18 círculos, 9 cuadrados y 3 triángulos. En el caso de la izquierda la población de círculos ha pasado a contar con 24 individuos permitiendo el aumento de la población de cuadrados (13 individuos) y de triángulos (4 individuos). El coeficiente de Bray-Curtis señala que ambos ecosistemas son muy similares (0,85). En el caso de la derecha, en cambio, algo ha eliminado un número importante de triángulos lo que ha llevado a un aumento de la población de cuadrados (18) y una reducción en la población de círculos (9). El coeficiente de Bray-Curtis señala que se ha producido un cambio importante (0,66).

Algoritmo nonIncrementalBLClustering *(colección)*
Input: *colección*, una lista que contiene un vector de *n*-gramas por cada documento de la colección

1. **for each** (d_i, d_j) *en colección* **do**
2. $matriz(i)(j) \leftarrow (\mathbf{pi}(di, dj) + \mathbf{rho}(di, dj))/2$
3. $matriz(j)(i) \leftarrow matriz(i)(j)$
4. **loop**
5. **for each** (d_i, d_j) *en colección* **do**
6. $C_Z \leftarrow \mathbf{brayCurtis}(matriz(i), matriz(j))$
7. $parecidos(i)(j) \leftarrow matriz(i)(j) \cdot C_Z$
8. **loop**
9. $x \leftarrow \mathbf{media}(parecidos)$
10. $s \leftarrow \mathbf{desviacion}(parecidos)$
11. $umbral \leftarrow x + alfa \cdot s$
12. **for each** (d_i, d_j) *en colección* **do**
13. **if** $parecidos(i)(j) \geq umbral$
14. $clusters \leftarrow \mathbf{agrupar}(d_i, d_j)$
15. **end if**
16. **loop**
17. **return** *clusters*

Fig. 55 Algoritmo no incremental de clasificación automática.

2.2 Algoritmo incremental basado en *blindLight*

Cuando la colección de documentos es muy grande la utilización del algoritmo anterior puede ser enormemente costosa tanto temporal como espacialmente. Por ese motivo se ha desarrollado un algoritmo incremental que, a pesar de su sencillez, proporciona resultados adecuados al compararlo con otras técnicas de clasificación automática. En este caso, puesto que no se dispone de información completa acerca del "comportamiento" de cada documento en relación con el resto de elementos de la colección es inviable la aplicación del coeficiente de Bray-Curtis utilizado en el método anterior.

El procedimiento es relativamente sencillo (véase Fig. 56, Fig. 58 y Fig. 59). En primer lugar se obtiene el centroide de la colección a clasificar y la similitud de cada documento con el mismo. Una vez hecho esto se extraen aleatoriamente documentos individuales del conjunto original. Cada documento aleatorio es comparado con los grupos disponibles[1] obteniendo la similitud entre el documento y cada grupo. Hay que recordar que previamente se ha calculado la similitud de cada documento con el centroide de la colección y cada vez que se crea o modifica un grupo también se calcula la similitud de dicho grupo con el centroide.

Algoritmo incrementalBLclustering *(colección)*
Input: *colección*, una lista que contiene un vector de *n*-gramas por cada documento de la colección
1. *provisionales* ← ***tentativeClustering*** *(colección)*
2. *(singletons, provisionales)* ← ***separarSingletons*** *(provisionales)*
3. *(dispersos, definitivos)* ← ***testDispersion*** *(provisionales)*
4. *nuevos* ← ***tentativeClustering*** *(dispersos)*
5. *semilla* ← *definitivos + nuevos*
6. *clusters* ← ***tentativeClustering*** *(singletons, semilla)*
7. **return** *clusters*

Fig. 56 Algoritmo incremental de clasificación automática de documentos.

De este modo es posible determinar la similitud existente entre el documento y el grupo, y la media de las similitudes entre el centroide y el documento y entre el centroide y el grupo, respectivamente. Si la similitud entre el documento y el grupo es mayor que la segunda se asigna el documento al grupo (y se recalcula la similitud del mismo con el centroide) y en caso contrario se transforma el documento en un nuevo grupo (véase Fig. 57). Este proceso se repite hasta que la colección queda vacía.

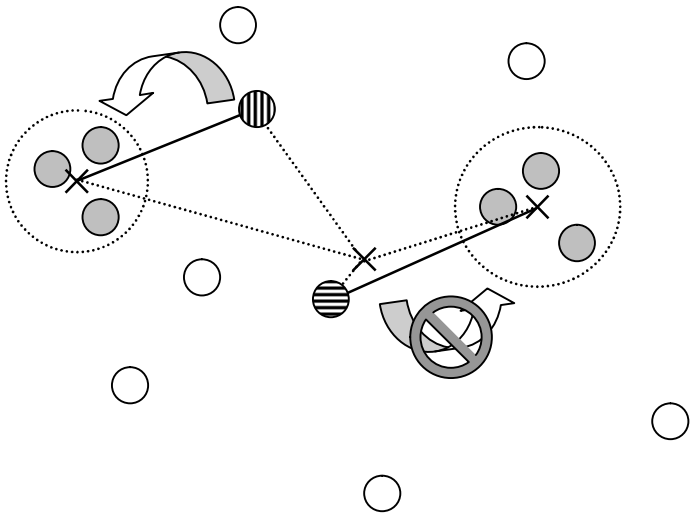


Fig. 57 Proceso de asignación de un documento a un grupo.

Un documento es asignado a un grupo si el parecido entre el grupo y el documento es superior a la media de las similitudes respectivas del grupo y el documento con el centroide de la colección. En la figura el documento rayado verticalmente puede asignarse al grupo de la izquierda mientras que el rayado horizontalmente no se asigna al de la derecha.

[1] Naturalmente, el primer documento extraído es convertido automáticamente en un grupo con un único documento.

Algoritmo tentativeClustering *(colección, semilla=λ)*
Input: *colección*, una lista que contiene un vector de *n*-gramas por cada documento de la colección

```
1.  if semilla ≠ λ
2.      clusters ← semilla
3.  end if
4.  centroide ← centroid (colección)
5.  for each documento d_i en colección do
6.      simDocs(i) ← ( pi (d_i, centroide) + rho (d_i, centroide))/2
7.  loop
8.  from i ← 0 to tamaño de colección do
9.      n ← random (tamaño de colección)
10.     if tamaño de clusters = 0 matriz(j)(i) ← matriz(i)(j)
11.         clusters(i)=d_n
12.         simClusters(i)=simDocs(n)
13.     else
14.         for each cluster k en clusters do
15.             sim ← ( pi (d_n, k) + rho (d_n, k))/2
16.             if sim > maxSim
17.                 maxSim ← sim
18.                 candidato ← k
19.             end if
20.         loop
21.         simDoc ← simDocs(n)
22.         simK ← simClusters(candidato)
23.         simAvg ← (simDoc + simK) / 2
24.         if maxSim > simAvg
25.             candidato ← agrupar (candidato, d_n)
26.             simClusters(candidato)←( pi (candidato, centroide) + rho (candidato,
    centroide))/2
27.         end if
28.     end if
29. loop
30. return clusters
```

Fig. 58 Algoritmo para obtener un conjunto de grupos "provisionales" (admite opcionalmente un conjunto de grupos "semilla").

Algoritmo testDispersion *(clusters)*
Input: *clusters*, un conjunto de grupos de documentos

```
1.  for each cluster k en clusters do
2.      variacionSims(k) ← coeficienteVariacion (k)
3.  loop
4.  x ← media (variacionSims)
5.  desv ← desviacion (variacionSims)
6.  umbral ← x + alfa ·desv
7.  for each cluster k en clusters do
8.      if variacionSims(k) > umbral
9.          coleccion ← extraerDocumentos (k)
10.     else
11.         clustersDef ← k
12.     end if
13. loop
14. return (coleccion, clustersDef)
```

Fig. 59 Algoritmo para determinar que grupos están "dispersos" y cuales son "definitivos".

En ese momento se dispone de una serie de grupos provisionales y, muy probablemente, de una serie de grupos constituidos por un único documento (*singletons*). Se aislan los primeros y se determina cuales constituyen grupos "dispersos". Para ello se calcula el coeficiente de variación de la similitud intra-grupal y, posteriormente, la media y desviación típica de dichos coeficientes para todos los grupos provisionales obteniendo un umbral. Aquellos grupos cuyo coeficiente de variación de similitud intra-grupal supere el umbral se consideran "dispersos" y sus documentos son trasladados a una nueva colección que es clasificada aplicando el procedimiento original. Los grupos obtenidos en esta clasificación son añadidos a los grupos provisionales no dispersos y todos pasan a considerarse grupos definitivos.

En este instante se dispone de un conjunto de grupos de documentos no dispersos y una serie de *singletons*. A fin de asignar, si es posible, dichos documentos aislados a un grupo se extrae el **medoide**[1] de cada grupo convirtiéndolo en un grupo por sí mismo y se aplica el algoritmo original sobre los *singletons* y estos grupos "semilla". Finalizada esta fase, algunos de los *singleton* habrán sido agrupados con algún medoide mientras que otros permanecerán aislados. Los que se han asociado a un medoide se integran en el grupo correspondiente mientras que los *singletons* se consideran grupos definitivos y el algoritmo finaliza con esta partición del conjunto original.

3 Algunos resultados de la aplicación de *blindLight* a la clasificación automática

A continuación se presentarán algunos resultados relevantes de la aplicación de *blindLight* a la clasificación automática de documentos. En primer lugar se describirá un experimento que permitió establecer una clasificación de distintos lenguajes naturales basándose tanto en datos léxicos como fonológicos. Seguidamente se mostrarán una serie de experimentos cuyos resultados dan soporte a la afirmación del autor acerca de la capacidad de *blindLight* para ofrecer resultados similares (y en algunos casos superiores) a los de métodos específicos.

3.1 Clasificación genética (y automática) de lenguajes naturales[2]

La mayor parte de lenguajes humanos están asignados a una familia que agrupa una serie de lenguas derivadas de un único idioma anterior. Un ejemplo que se cita con frecuencia son las lenguas romances que descienden del latín. Sin embargo, en muchos casos la lengua de origen es desconocida y es necesario reconstruir sus características a fin de determinar la evolución que siguió dicho idioma hasta producir lenguas conocidas. Este método, conocido como método comparativo, tiene sus orígenes en los estudios que realizó en el S. XIX August Schleicher sobre las lenguas indoeuropeas.

El método comparativo establece que dos idiomas están relacionados tan sólo si es posible reconstruir (aunque sea parcialmente) un idioma ancestro común. La razón es simple, tan sólo se consideran relaciones entre idiomas que han evolucionado por transmisión de padres a hijos durante generaciones. Así pues, la aplicación del método es lenta y compleja, y en muchos casos resulta imposible reconstruir un idioma anterior común, razón por la cual muchas lenguas humanas aparecen "aisladas" como únicos ejemplares de su propia familia (casos del vasco, el japonés o el coreano).

---

[1] El documento más parecido al centroide de un grupo.

[2] Gran parte de este apartado apareció en (Gayo Avello, Álvarez Gutiérrez y Gayo Avello 2004b).

Se han propuesto otras técnicas alternativas que también tienen como objetivo establecer vínculos de parentesco entre lenguas sin la necesidad de reconstruir ningún ancestro común y empleando tan sólo información léxica. Por ejemplo, la léxico-estadística (Swadesh 1950) o glotocronología (Lees 1953) y la comparación léxica masiva (Greenberg 1966).

La lexico-estadística trata de reconstruir árboles lingüísticos para lenguajes pertenecientes a la misma familia lingüística analizando el porcentaje de palabras afines[1] mientras que la glotocronología pretende, además, estimar la fecha en que dos lenguas divergieron a partir de un ancestro común. Por lo que se refiere a la comparación léxica masiva se basa en la comparación de palabras equivalentes (no necesariamente afines) en múltiples idiomas buscando sonidos similares en las mismas. Como principal éxito de esta última cabe destacar la clasificación por parte de Joseph H. Greenberg de los distintos idiomas africanos en cuatro grandes familias (Greenberg 1966); esta clasificación aunque inicialmente muy polémica es actualmente aceptada. Igualmente polémica es su clasificación de los idiomas nativos de América en tres grandes familias lingüísticas de las cuales la Amerindia no se admite generalmente como una familia única.

Es necesario decir que la mayor parte de los lingüistas no consideran estas técnicas en absoluto ortodoxas puesto que se basan únicamente en parecidos superficiales a nivel léxico y no fonético, por ejemplo, Poser y Campbell (1992) o Goddard y Campbell (1994). No obstante, las conclusiones obtenidas con dichas técnicas, en particular mediante la comparación léxica masiva, se han validado al aplicarse sobre idiomas cuya clasificación es bien conocida (caso de las lenguas indoeuropeas) o al proporcionar clasificaciones verificadas posteriormente de manera tradicional (caso de las lenguas africanas).

Así pues, y a pesar de la polémica, se han implementado distintos algoritmos que utilizan información meramente léxica para clasificar automáticamente distintos lenguajes humanos, por ejemplo (Dyen, Kruskal y Black 1992), (Kessler 1995), (Huffman 1998) o (Nerbonne y Heeringa 1997). Muchas de estas técnicas se basan en el clásico modelo vectorial, caso de Huffman que utiliza *Acquaintance* (Damashek 1995), o aplican la distancia de Levenshtein (1966) directamente sobre cadenas de caracteres (Kessler 1995) o fonemas (Nerbonne y Heeringa 1997). Recientemente Gray y Atkinson (2003) han analizado la evolución de las lenguas indoeuropeas con técnicas biocomputacionales y Warnow *et al.* (2004) y Evans, Ringe y Warnow (2004) han estudiado modelos estocásticos de evolución lingüística.

Por todo ello el autor de este trabajo consideró interesante la aplicación de *blindLight* al problema de la clasificación genética de lenguajes. La principal diferencia entre los trabajos anteriores y los experimentos que se llevaron a cabo radicó, además de en la técnica empleada, en el tipo de información lingüística utilizada. En primer lugar, se trabajó no sobre listas de palabras traducidas (como las listas de Swadesh[2]) sino sobre documentos

[1] También términos cognados. Palabras de distintos idiomas de significado y sonido similar. Como ejemplos de palabras afines pueden citarse *night*, *nuit* o *noche* en inglés, francés y español, respectivamente o *shalom* y *salaam* en hebreo y árabe.

[2] Una lista de Swadesh (1950) es una secuencia de palabras que constituyen un vocabulario básico (y presumiblemente antiguo). Todas las listas de Swadesh contienen el mismo vocabulario de tal modo que están alineadas a nivel de palabra constituyendo un *corpus* paralelo. Por ejemplo, la lista de Swadesh en castellano comienza con las palabras `{yo, tú/usted, él/ella, nosotros/nosotras, vosotros/vosotras, ellos/ellas, este, ese/aquel, …}` mientras la correspondiente al inglés contiene `{I, you/thou, he, we, you, they, this, that, …}`

completos[1]. Por otro lado además de información léxica también se utilizaron transcripciones fonéticas ya que ésta es una de las debilidades frecuentemente achacadas a las técnicas estadísticas.

Así, el primer experimento se realizó sobre vectores de *n*-gramas obtenidos a partir de los tres primeros capítulos del Libro del Génesis y, aplicando la versión no incremental y jerárquica del algoritmo de clasificación, se obtuvo un árbol con 14 idiomas (véase Fig. 60).

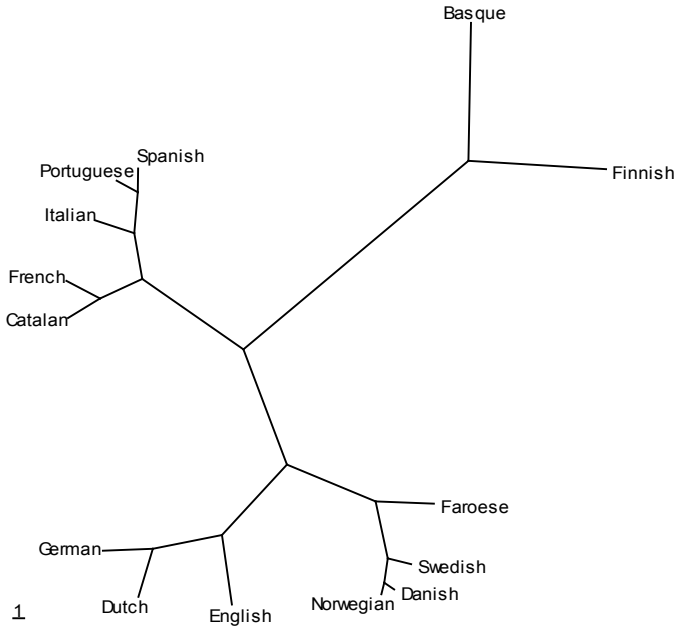


Fig. 60 Dendrograma mostrando las distancias entre muestras escritas de 14 idiomas europeos (tres primeros capítulos del Libro del Génesis).

En el segundo experimento se utilizaron vectores construidos a partir de transcripciones fonéticas (véase Fig. 61) de la fábula "El viento del norte y el sol" que se obtuvieron del *Handbook of the International Phonetic Association* (Manual de la Asociación Fonética International) (IPA 1999). Los idiomas que participaron en este último experimento fueron alemán, catalán, español, francés, gallego, holandés, inglés, portugués y sueco. En el árbol obtenido en esta ocasión se encontraban 8 idiomas presentes en el primer experimento (véase Fig. 62).

Al analizar los resultados obtenidos es necesario ser cauto, en primer lugar porque el autor no es lingüista y, en segundo, porque los lingüistas no dan demasiado crédito a los resultados obtenidos a partir de análisis estadísticos de datos puramente léxicos. No obstante, es preciso señalar que la segunda de las experiencias no ha empleado datos léxicos sino fonológicos producidos por expertos en el campo y los resultados obtenidos en dicha prueba son coherentes con los alcanzados en la primera.

Por otra parte, los resultados obtenidos en ambos casos clasifican de manera adecuada las lenguas indoeuropeas e incluso la estrecha relación manifestada (tanto léxica como fonéticamente) entre catalán y francés encuentra apoyo en ciertos autores, por ejemplo, Pere Verdaguer (1999).

---

[1] Recientemente el autor ha tenido noticia de un trabajo en el que utilizando la *Declaración Universal de los Derechos Humanos* se ha establecido una clasificación automática de 52 idiomas (Li *et al.* 2004, p. 3260-3261). La técnica empleada difiere de la utilizada por el autor pero los resultados son muy similares.

ðə ˈnɔɹθ ˌwɪnd ən ə ˈsʌn wɚ dɪsˈpjutɪŋ ˈwɪtʃ wəz ðə ˈstɹɑŋgɚ, wɛn ə ˈtɹævlɚ kem əˈlɑŋ ˈɹæpt ɪn ə ˈwɔɹm ˈklok. ðə əˈgɹid ðət ðə ˈwʌn hu ˈfɚst səkˈsidəd ɪn ˈmekɪŋ ðə ˈtɹævlɚ ˈtek ɪz ˈklok ˌɑf ʃʊd bi kənˈsɪdɚd ˈstɹɑŋgɚ ðən ðɪ ʌðɚ. ˈðɛn ðə ˈnɔɹθ ˌwɪnd ˈblu əz ˈhɑɹd əz hi ˈkʊd, bət ðə ˈmɔɹ hi ˈblu ðə ˈmɔɹ ˈklosli dɪd ðə ˈtɹævlɚ ˈfold hɪz ˈklok əˈɹaʊnd hɪm; ˌæn ət ˈlæst ðə ˈnɔɹθ ˌwɪnd ˌgev ˈʌp ði əˈtɛmpt. ˈðɛn ðə ˈsʌn ˈʃaɪnd ˌaʊt ˈwɔɹmli, ən ɪˈmidiətli ðə ˈtɹævlɚ ˌtʊk ˈɑf ɪz ˈklok. ən ˈso ðə ˈnɔɹθ ˌwɪnd wəz əˈblaɪʒ tɪ kənˈfɛs ðət ðə ˈsʌn wəz ðə ˈstɹɑŋgɚ əv ðə ˈtu.

The North Wind and the Sun were disputing which was the stronger, when a traveller came along wrapped in a warm cloak. They agreed that the one who first succeeded in making the traveller take his cloak off should be considered stronger than the other. Then the North Wind blew as hard as he could, but the more he blew, the more closely did the traveller fold his cloak around him; and at last the North Wind gave up the attempt. Then the Sun shone out warmly, and immediately the traveller took off his cloak. And so the North Wind was obliged to confess that the Sun was the stronger of the two.

Fig. 61 Transcripción fonética de la versión inglesa de la fábula de Esopo (*IPA* 1999, p. 44).

Debe quedar claro que en modo alguno se está aventurando ninguna hipótesis sobre la evolución de estos lenguajes, tan sólo se presentan unos resultados que, al ajustarse a clasificaciones construidas por humanos, y que junto con los obtenidos en el capítulo anterior para la clasificación de los mini-*corpora* paralelos apoyan la afirmación del autor acerca de que la utilización de *blindLight* como método de *clustering* permite obtener clasificaciones bastante "naturales" desde el punto de vista del usuario.

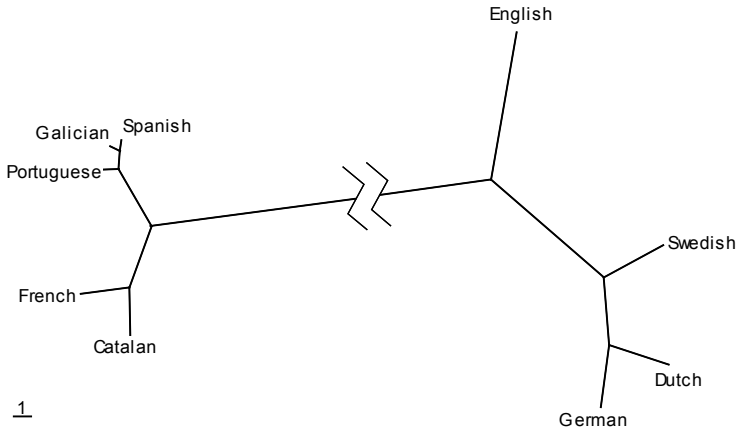


Fig. 62 Dendrograma mostrando las distancias entre muestras orales de 9 idiomas europeos (transcripciones fonéticas de la fábula "El viento del norte y el sol").

La distancia entre los sub-árboles Galo-Ibérico (izquierda) y Germánico es 23.985, más del doble de la distancia mostrada en la figura.

3.2 Comparación de *blindLight* con *SOM*

Ya se ha mencionado antes la aplicación de los Mapas Auto-Organizativos (*Self-Organizing Maps* o *SOM*) al problema de la clasificación automática de documentos. Este tipo de mapas, también denominados mapas de Kohonen (1982) por su inventor, pueden interpretarse como una distribución (generalmente bi o tridimensional) de neuronas situadas en posiciones fijas que se entrenan con vectores de características en un proceso competitivo. Para cada vector hay una única neurona ganadora que ajustará sus pesos para

aproximarse al vector de entrada. No obstante, el resto de neuronas también ajustan parcialmente sus pesos de forma inversamente proporcional a la distancia a que se encuentren de la vencedora. De este modo, se van vinculando los vectores a diferentes coordenadas del mapa y en caso de que estén etiquetados se asociarán sus etiquetas a las distintas zonas del mismo.

Ya el propio Kohonen propuso la utilización de mapas auto-organizativos para visualizar y explorar colecciones de artículos *USENET* (Honkela *et al.* 1996) –véase Fig. 63– y la misma técnica fue aplicada con objetivos similares a una serie de colecciones entre las que se incluía la colección *TIME* (Rauber y Merkl 1999) y la edición de 1990 de *The World Fact Book* de la *CIA* (Merkl y Rauber 1998).

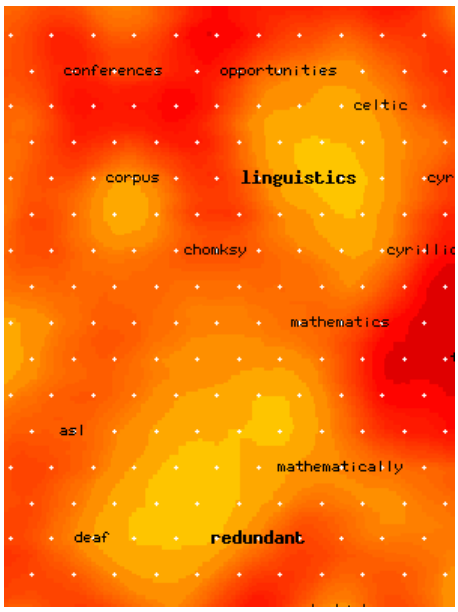


Fig. 63 Una zona de un mapa auto-organizativo para artículos publicados en los grupos `sci.lang.*`.

La colección *TIME* fue elaborada por Gerald Salton (1972) y consiste en 423[1] artículos publicados durante 1963 en la sección internacional de dicha revista. Los documentos abarcan toda una serie de temas tales como la situación en Vietnam, la guerra fría o escándalos políticos. El objetivo original de la colección fue el de servir de prueba estandarizada para la evaluación de sistemas de recuperación de información pero se ha utilizado con frecuencia para comparar distintos métodos de clasificación y categorización.

Por lo que respecta a *The World Fact Book* se trata de una publicación anual de la *CIA* que describe en términos geográficos, sociales, económicos y políticos distintas regiones y países del planeta. Merkl y Rauber (1998) utilizaron la edición de 1990 para elaborar una colección que consta de 246 documentos. Algunos ejemplos de documentos serían `argentin`, `soviet_u`, `gazastri`, `world` o `atlantic`.

En este apartado se describirán los resultados alcanzados al aplicar *blindLight* a la clasificación de ambas colecciones y se compararán con los obtenidos por Rauber y Merkl utilizando *SOM*[2] mientras que en el siguiente se comparará con técnicas como *k*-medias o *UPGMA*. Recuérdese que el autor afirma que esta técnica es capaz de ofrecer en todas sus aplicaciones, incluida la clasificación de documentos, resultados similares a los de técnicas específicas.

Al disponer de los grupos producidos por *SOM* la forma más sencilla de comparar esos resultados con los alcanzados con *blindLight* es calculando la similitud promedio. Para ello hay que calcular la media de las similitudes entre todos los posibles pares de

[1] Según Salton (1972, p. 5) la colección consta de 425 artículos pero las versiones actualmente disponibles sólo incluyen 423.

[2] *CIA World FactBook* <http://www.ifs.tuwien.ac.at/~andi/somlib/data/wfb90/wfb7a_1_1_0_0.html>, *TIME* <http://www.ifs.tuwien.ac.at/~andi/somlib/data/time60/time-map10x15_labels.html>

documentos de un grupo (Steinbach *et al.* 2000, p. 6) y posteriormente combinar estas medidas individuales en una medida única ponderando la similitud promedio de cada grupo en función del porcentaje de documentos de la colección incluidos en el mismo. Cuanto más elevada sea el valor de esta medida más cohesivos serán los grupos y, en consecuencia, mejor la técnica de clasificación.

En el caso de la primera colección se empleó la versión incremental de *blindLight* que dividió la colección *TIME* en 177 grupos de los cuales 78 estaban formados por un único documento. Para aplicar *SOM* Rauber y Merkl utilizaron un mapa bidimensional de 10x15 celdas[1]. De cara a la comparación de ambas técnicas se ha considerado cada celda como un grupo lo que supone 150 grupos de los que 41 fueron *singletons*. Los resultados obtenidos para esta colección aplicando *blindLight* y *SOM* se muestran en la Tabla 4.

blindLight		***SOM***	
Incluyendo *singletons*	Sin incluir *singletons*	Incluyendo *singletons*	Sin incluir *singletons*
0,597	**0,506**	0,547	0,498

Tabla 4. Similitud promedio de los resultados obtenidos con *blindLight* y *SOM* al clasificar la colección *TIME*.

Por lo que respecta a la segunda colección se empleó el método no incremental y se estableció sobre el dendrograma un punto de corte que proporcionase un número de grupos similar al obtenido con *SOM*. Al aplicar mapas auto-organizativos la colección queda dividida en 84 grupos de los cuales 26 estaban formados por un único documento. Los grupos obtenidos con *blindLight* fueron 86 incluyendo 24 *singletons*. Los resultados obtenidos en este segundo experimento se muestran en la Tabla 5.

blindLight		***SOM***	
Incluyendo *singletons*	Sin incluir *singletons*	Incluyendo *singletons*	Sin incluir *singletons*
0,731	**0,702**	0,712	0,678

Tabla 5. Similitud promedio de los resultados obtenidos con *blindLight* y *SOM* al clasificar la colección *CIA*.

Como se puede ver, los resultados obtenidos con *blindLight* son ligeramente mejores que los alcanzados por *SOM*. No obstante, el criterio general establece que una mejora de rendimiento es "apreciable" tan sólo si se encuentra entre el 5 y el 10% y "sustancial" si supera el 10% (Spärck-Jones 1974, citado por Rasmussen 2002) y no es ésta la situación. En el caso de la colección *TIME* hay una mejora de *blindLight* sobre *SOM* del 1,61% sin contar los *singletons* y del 9,14% contándolos. Por lo que respecta a la colección de la *CIA* el incremento en la similitud promedio es del 3,54% y del 2,67%, respectivamente. Así pues, se puede afirmar que las diferencias entre *blindLight* y *SOM* respecto a la clasificación de ambas colecciones no son relevantes.

Tales resultados son acordes con el objetivo de ofrecer una efectividad similar a los de técnicas específicas para una serie de tareas PLN por lo que permiten sostener dicha afirmación en lo que se refiere al problema de la clasificación automática y la técnica de mapas auto-organizativos. En el siguiente apartado se procederá a comparar *blindLight* con las técnicas de *k*-medias, *k*-medias bisecante (Steinbach, Karypis y Kumar 2000) y *UPGMA*[2] (Jain y Dubes 1988, citado por Zhao y Karypis 2002).

---

[1] El mapa está disponible en `http://www.ifs.tuwien.ac.at/~andi/somlib/data/time60/time-map10x15_labels.html`. Posteriormente los autores del mismo lo analizaron manualmente para reducir el número de grupos; no obstante, no se han empleado estos últimos debido a la intervención humana requerida para su construcción.

[2] *Unweighted Pair-Group Method with Arithmetic Mean.*

No obstante, no queremos concluir este apartado sin ofrecer algunas muestras de los resultados obtenidos con la nueva técnica propuesta a fin de permitir al lector valorar desde un punto de vista más práctico su utilidad. Después de todo, no se puede olvidar que el objetivo de la clasificación automática de documentos es señalar a un usuario final grupos con similitudes "interesantes".

Así, en Fig. 64 se presentan algunos grupos de documentos encontrados al clasificar con *blindLight* la colección *TIME* y que resultan particularmente representativos del tono de la colección. El primer grupo contiene una serie de documentos en torno a Nikita Khruschev, el segundo trata acerca de las elecciones en el Reino Unido, el siguiente sobre un escándalo político-sexual también en el Reino Unido y el último contiene documentos que tocan el problema de la tensión religiosa y política en Vietnam.

Cluster	T026, T032, T135, T248, T346, T539, **T542**, T558
Etiquetas	Khrushchev, Russia, Nikita, Soviet, party, Moscow, West, Kremlin, foreign…
Medoide	...Yet who should be serving up lemonade last week than that old realist **Nikita Khrushchev**. In the **Kremlin**'s marble-halled palace of the congresses, addressing the communist **party** central committee and more than 5,000 other comrades, **Nikita** promised that one great force would miraculously straighten out the **Soviet** economic mess: big chemistry... More important, he admitted that **Russia** would need credit and supplies, including entire factories, from the **West** but not, he fumed, at "fabulous profits" to the capitalists... Moreover, new products must show better design, because it is "no longer possible to tolerate" Russian consumer goods that "look less smart than **foreign** articles...

Cluster	**T105**, T240, T265, T270, T324, T430, T493, T497, T503, T512
Etiquetas	labor, party, minister, government, Britain, election, leader…
Medoide	...The **labor party** last week chose a new **leader** to carry its banner against the Tories in **Britain**'s coming general **election**... Some opposed his pro-common market views; others among **labor**'s intellectual center and right flinched at the thought of a working-class, up-from-the-ranks prime **minister**, and preferred to go to the country with an Oxford graduate and economics don like Wilson… The feuding has faded, and **labor** finds itself in the best shape in years to topple the **government** of Prime **Minister** Harold Macmillan...

Cluster	T170, T301, T315, **T337**, T342, T354
Etiquetas	Christine, Ward, Profumo, Keeler, British, government, girl, party, London, Britain, flat, Stephen, Russian…
Medoide	...The moral decay surrounding the **Profumo** affair, he tried hard to suggest, must be blamed on the Tories. Referring to **Christine Keeler**'s reported $14,000-a-week nightclub contract, Wilson declared: "there is something utterly nauseating about a system of society which pays a harlot 25 times as much as it pays its prime minister." For the rest, Harold Wilson stuck to the security issue and the **government**'s handling of the **Profumo** case, which he attacked as either dishonest or incompetent, or both... First, there was the **Christine**-**Profumo** affair itself, which, according to **Profumo**, lasted only a few months, from July to December 1961, but by other evidence possibly, lasted longer. During those same months, **Christine** also entertained **Russian** assistant naval attache Evgeny Ivanov, who had been pals for some time with her mentor, Dr. **Stephen Ward**...

Cluster	**T418**, T434, T464, T480, T498
Etiquetas	Diem, government, Viet, Saigon, Nhu, Buddhist, Nam, Dinh, south, Ngo, troops, army, Cong, brother, regime, war, president, Vietnamese, communist, crackdown, city, law, Mme, aid, martial, catholic, jail, roman, radio, Thuc …
Medoide	...to the **Diem government**, the **crackdown** obviously seemed necessary to protect the **regime** and enforce the **law** of the land against **Buddisht** defiance... it also put U.S. policy in **south Viet Nam**, which involves the lives and safety of 14,000 U.S. **troops**, into an agonizing dilemma... **roman catholic Diem** may finally have shattered his own political usefulness. he also opened up the possibilities of coups, counter-coups, and even civil **war** from all of which only the **communist Viet Cong** could benefit... Pagodas, sporting protest signs in **Vietnamese** and English... appeals for **aid** were broadcast to **President** Kennedy... At a grisly, well-organized press conference in **Saigon**... his **brother** and sister-in-**law**, **Ngo Dinh Nhu** and **Mme**... ten truckloads of bridge defenders were carted away to **jail**... and an estimated 500 people were arrested throughout the **city**... under the **martial law** proclamation, the **army** was given blanket search-and-arrest powers and empowered to forbid all public gatherings...

Fig. 64 Algunos grupos obtenidos al clasificar la colección *TIME*.

Para cada grupo se muestran los documentos que forman parte del mismo, el medoide del grupo (en negrita y extractado) y una serie de palabras que etiquetan el grupo obtenidas automáticamente al extraer aquellas que aparecen en un porcentaje elevado de documentos del grupo.

En Fig. 65 se muestran sobre una serie de mapas algunos de los grupos encontrados al analizar la colección de documentos de la *CIA* con *blindLight* (como se recordará, dicha colección está básicamente constituida por descripciones de países). Nótese cómo los grupos han incluido parámetros tanto geográficos (países africanos y americanos) como, en muchos casos, políticos (grupos localizados en oriente medio, el grupo EEUU-URSS), socio-económicos (el grupo formado por RFA, Suiza y Austria) y/o ideológicos (países del "Telón de Acero").

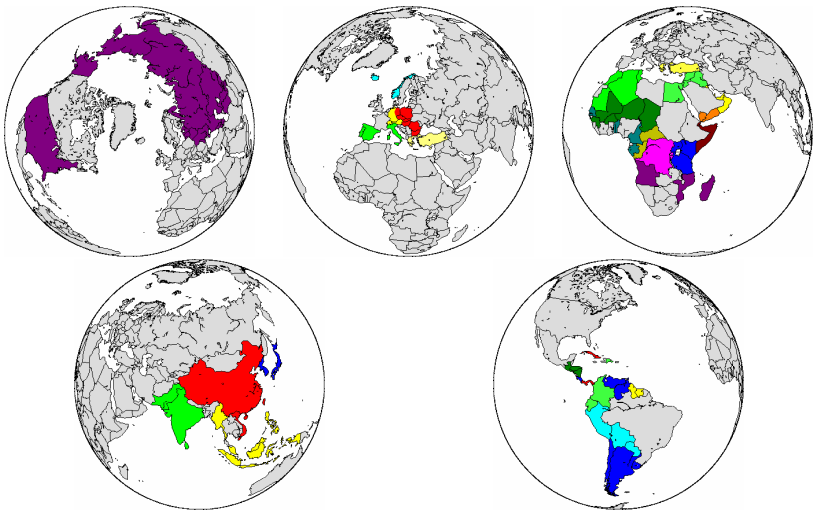


Fig. 65 Grupos de regiones localizados por *blindLight* analizando los textos de la *CIA*.

(De izquierda a derecha y de arriba abajo. Recuérdese que los datos son previos al desmoronamiento del bloque comunista). El primer grupo incluye tan sólo a EEUU y la URSS. El segundo mapa muestra a Portugal, España e Italia en un grupo, los países "germánicos" (RFA, Suiza y Austria) en otro, Grecia, Chipre y Turquía en un tercero, el "Telón de Acero" como cuarto grupo (señalado en rojo) y dos países nórdicos en un quinto. En tercer lugar se muestra una segmentación de África que parece responder a criterios geográficos, resulta interesante el grupo constituido por Egipto, Israel, Siria, Jordania, Líbano e Iraq. El siguiente gráfico divide Asia en cuatro grupos. El último mapa muestra una segmentación de América central y del sur por criterios fundamentalmente geográficos, nótese la ausencia de Brasil.

En la Tabla 6 se presentan algunos grupos particularmente interesantes y un análisis *post mortem* del "criterio" subyacente a cada clasificación. En Fig. 66 se muestran otros grupos obtenidos al "cortar" el dendrograma en un nivel superior, aquí tienen mayor peso los factores geográficos (fundamentalmente en Asia, África y América) y económicos (el primer mapa muestra los países más industrializados).

Por lo que respecta a grupos que contienen un único documento se han encontrado, en el "primer corte", Vaticano, Libia, Mongolia, Nigeria, Etiopía, Sudán, Ghana, Malta, Haití, México, Brasil, Arabia Saudí, Tailandia, Francia, S. Kitts, Georgia del Sur y las Islas Sandwich, Gibraltar, los territorios franceses en el Antártico, la Antártida y Jan Mayen (un territorio noruego). En cuanto al "segundo corte" los *singletons* fueron el Vaticano, Libia y Nigeria. Dada la naturaleza de la colección estos grupos podrían interpretarse como regiones "aisladas" de algún modo de sus vecinos, lo cual parece especialmente aplicable en los casos del Vaticano (único estado monárquico-sacerdotal del mundo), la Antártida o Gibraltar (único territorio colonial en suelo europeo).

Grupo	World, Svalbard, Gaza Strip, West Bank
Post mortem	El primer documento describe la situación mundial en términos generales. Svalbard es un territorio de soberanía noruega pero explotable por otros países, con presencia soviética y disputas marítimas entre ambos países. La franja de Gaza y Cisjordania son territorios ocupados por Israel y sin una situación política definida.

Grupo	Arctic ocean, Atlantic ocean, Indian ocean, Pacific ocean
Post mortem	Océanos

Grupo	Andorra, San Marino, Liechtenstein, Monaco
Post mortem	Microestados con relaciones especiales con otras naciones, en particular en términos de defensa: Liechtenstein con Suiza, Mónaco con Francia y Andorra con España y Francia.

Grupo	Iraq-Saudi Aurabia Neutral Zone, Paracel Islands, Spratly Islands
Post mortem	Territorios sin presencia humana permanente cuya defensa es responsabilidad de varios países o cuya soberanía es disputada por varios países.

Grupo	American Samoa, Guam, Northern Mariana Islands, Puerto Rico, Virgin Islands
Post mortem	Islas que constituyen territorios de EEUU (American Samoa, Guam y Virgin Islands) o estados asociados a EEUU (Northern Mariana Islands y Puerto Rico) y cuya defensa es responsabilidad de este pais.

Grupo	Marshall Islands, Federated States of Micronesia, Palau
Post mortem	Estados libres asociados a EEUU, situados en el Pacífico y cuya defensa es responsabilidad de este pais.

Grupo	Navassa Island, Kingman Reef, Palmyra Atoll
Post mortem	Territorios deshabitados de EEUU.

Grupo	Falkland Islands (Malvinas), St. Helena
Post mortem	Islas que constituyen territorios dependientes del Reino Unido y cuya defensa es responsabilidad de este país.

Grupo	Ashmore and Cartier Islands, Coral Sea Islands
Post mortem	Islas que constituyen territorio de Australia y carecen de población autóctona aunque pueden estar pobladas estacionalmente.

Grupo	Clipperton Island, Bassas da India, Europa Island, Tromelin Island, Glorioso Islands, Juan de Nova Island
Post mortem	Islas deshabitadas posesión de Francia.

Grupo	French Guiana, Reunion, Guadeloupe, Martinique, Mayotte, St. Pierre and Miquelon
Post mortem	Territorios y “colectividades” francesas de ultramar.

Grupo	French Polynesia, New Caledonia, Wallis and Futuna
Post mortem	Territorios franceses de ultramar situados en el Pacífico.

Tabla 6. Grupos producidos por *blindLight* y análisis *post mortem* de los mismos.

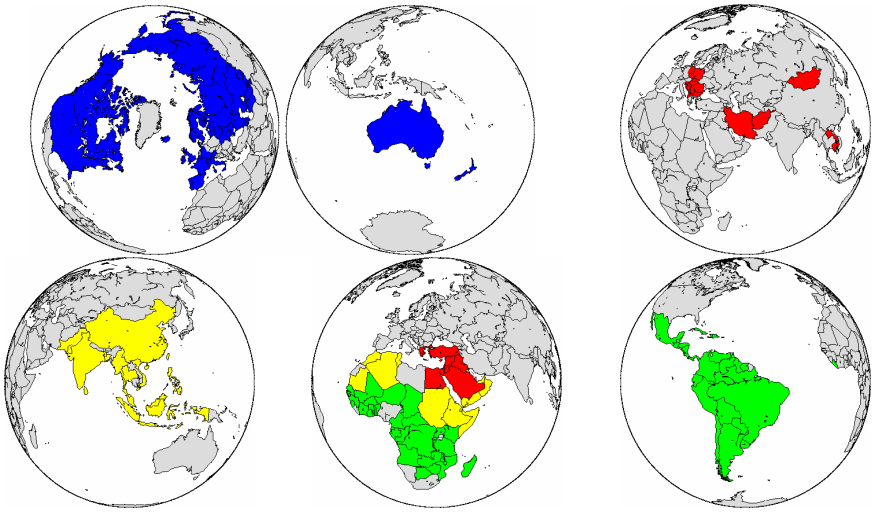


Fig. 66 Grupos de regiones obtenidos al "cortar" el dendrograma en un nivel superior.

El primer gráfico muestra el denominado "Norte Rico" que incluye a la URSS, Australia y Nueva Zelanda. El segundo mapa contiene países vinculados a la URSS ideológica, económica, militar y/o geográficamente: el "Telón de Acero", Irán (fronterizo) o Afganistan (fronterizo y ocupado). El tercer gráfico agrupa la mayor parte de Asia mientras el siguiente divide África en tres grandes grupos, destacando nuevamente la zona de Oriente Próximo. El último mapa coincide con Latinoamérica (nótese la ausencia de la Guyana Francesa) aunque incluye Liberia en un curioso salto transatlántico.

3.3 Comparación de *blindLight* con *k*-medias, *k*-medias bisecante y *UPGMA*

En el apartado anterior se han presentado los resultados obtenidos al aplicar *blindLight* y *SOM* sobre dos colecciones de documentos. Dichos resultados mostraron que las diferencias entre ambas técnicas no son relevantes. En este apartado se aprovechará el trabajo llevado a cabo por Steinbach, Karypis y Kumar (2000) para comparar la técnica propuesta por el autor con otras más "tradicionales".

En su trabajo Steinbach *et al.* presentaron un estudio experimental acerca del rendimiento de distintas técnicas de clasificación automática de documentos, en particular *k*-medias, una modificación de la misma denominada *k*-medias bisecante y un método jerárquico y aglomerativo clásico, *UPGMA*. El objetivo fundamental de dicho trabajo era determinar si, efectivamente, los métodos jerárquicos producen mejores clasificaciones que los métodos particionales. Para ello, utilizaron una serie de colecciones de documentos y obtuvieron un conjunto de medidas de la "calidad" de los resultados a fin de comparar las distintas técnicas. Del mismo modo, aplicando *blindLight* sobre una o más de dichas colecciones podrían obtenerse resultados análogos y comparar la técnica propuesta por el autor con las anteriores.

No obstante, se dan toda una serie de circunstancias que hicieron muy difícil la utilización de la mayor parte de las colecciones empleadas por Steinbach *et al.* En primer lugar, aun cuando los datos están accesibles en el sitio web de uno de los autores[1] no se ofrecen como texto plano sino procesados para su utilización con el paquete de *software CLUTO* haciéndolos inútiles para *blindLight* (véase Fig. 67).

[1] `http://www-users.cs.umn.edu/~karypis/cluto/files/datasets.tar.gz`

```
                                                    edition
                                                    dewei
4663 41681 83181                                    decim
1430 1 476 1 514 1 38 1 1024 1                      classif
13255 1 8549 1 2460 1 4987 1 175 1 249 1            studi
2186 1 1279 1 182 1 257 1 4515 1                    histori
…                                                   decimalclassif
                                                    ddc
                                                    …
```

Fig. 67 Datos procesados para ser empleados por *CLUTO*.

Las colecciones de documentos se representan mediante vectores de términos (a la izquierda). Cada fila del archivo se corresponde al vector de un documento donde los términos han sido reemplazados por índices enteros. A la derecha se muestran algunos de los términos empleados en la colección, su posición dentro del fichero sirve como índice del término para la representación vectorial. Este tipo de representación resulta inútil para *blindLight* que trabaja sobre el texto plano original de los documentos.

Por otro lado, al intentar localizar los textos originales se comprobó que varias de las colecciones son particiones utilizadas en la *Text REtrieval Conference* (*TREC*) que ni son libres ni gratuitas. Aún peor, la descripción para obtener dos particiones de la colección *Reuters-21578* es incompleta y el autor fue incapaz de reproducirlas aun disponiendo de ella. La única colección que pudo encontrarse y comprobarse que se correspondía con la descripción dada por Steinbach *et al.* fue la denominada *wap*[1] (*WebACE Project*).

Así pues, se aplicó *blindLight* (en su versión incremental) sobre esta colección (Han *et al.* 1998) que consta de 1560 páginas web extraídas de *Yahoo!* y asignadas a una única categoría de 20 posibles. Esto permitió no sólo que se pudiese calcular la similitud promedio (véase pág. 84) de los resultados obtenidos sino también la entropía de los mismos. Sin embargo, es necesario decir que en la bibliografía examinada se han encontrado dos definiciones de entropía distintas, Zhao y Karypis (2002) y Steinbach *et al.* (2000), y que a lo largo de este apartado se utilizará la segunda definición a fin de poder comparar adecuadamente los resultados obtenidos por Steinbach *et al.* y los alcanzados con *blindLight*.

Según Zhao y Karypis (2002) dado un grupo S_r de tamaño n_r la entropía del mismo se define como:

$$E(S_r) = -\frac{1}{\log q}\sum_{i=1}^{q}\frac{n_r^i}{n_r}\log\frac{n_r^i}{n_r}$$

donde q es el número de clases en la colección, n^i_r es el número de documentos de la clase i-ésima que fueron asignados al grupo r-ésimo. La entropía de la clasificación final se define como la suma de las entropías de todos los grupos ponderadas de acuerdo a su tamaño, es decir:

$$Entropia = \sum_{r=1}^{k}\frac{n_r}{n}E(S_r)$$

En cambio, según Steinbach *et al.* (2000, p. 7) la entropía de un grupo individual sería:

$$E(S_r) = -\sum_{i=1}^{q} p_{ir}\log p_{ir} = -\sum_{i=1}^{q}\frac{n_r^i}{n_r}\log\frac{n_r^i}{n_r}$$

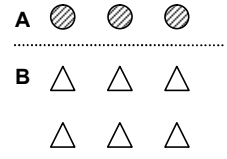
[1] `ftp://ftp.cs.umn.edu/dept/users/boley/PDDPdata/doc-K/`

Como se puede ver, la diferencia entre ambas definiciones es mínima y tan sólo cambia los valores numéricos y no la interpretación de la medida: a menor entropía mejor clasificación (véase Fig. 68).

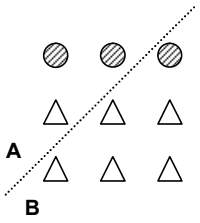
En las siguientes tablas se muestran los resultados obtenidos por *blindLight* para la colección *wap* y se comparan con los obtenidos por otras técnicas de clasificación (Steinbach *et al.* 2000, p. 14 y 15). Es necesario señalar que mientras dichas técnicas requieren la especificación del número de grupos a encontrar, *k*, la técnica propuesta por el autor no requiere tal parámetro por lo que los datos ofrecidos para *k*-medias, *k*-medias bisecante y *UPGMA* se han proyectado para *k*=70 (el número de grupos en que *blindLight* parte la colección) a partir de los publicados originalmente para *k*=16, 32 y 64.

blindLight	*k-medias*	*k-medias bisecante*	*k-medias bisecante "refinado"*	*UPGMA*	*UPGMA "refinado"*
1,1907	1,2230	**1,0888**	**1,0397**	1,3486	1,2561
Diferencia respecto a *blindLight*	-2,64% Inapreciable (A favor de *bL*)	9,36% Apreciable	14,52% Sustancial	-11,71% Sustancial (A favor de *bL*)	-5,21% Apreciable (A favor de *bL*)

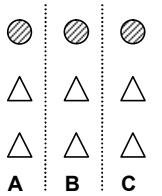
Tabla 7. Entropía de las distintas clasificaciones de la colección *wap*.



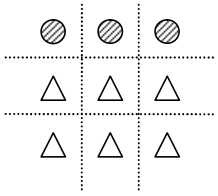
$$E(A) = -\left(\tfrac{3}{3}\cdot\log\tfrac{3}{3} + \tfrac{0}{3}\cdot\log\tfrac{0}{3}\right) = -(1\cdot\log 1 + 0\cdot\log 0) = 0$$
$$E(B) = -\left(\tfrac{0}{6}\cdot\log\tfrac{0}{6} + \tfrac{6}{6}\cdot\log\tfrac{6}{6}\right) = -(0\cdot\log 0 + 1\cdot\log 1) = 0$$
$$E = 0$$



$$E(A) = -\left(\tfrac{2}{3}\cdot\log\tfrac{2}{3} + \tfrac{1}{3}\cdot\log\tfrac{1}{3}\right) = -(-0{,}117 - 0{,}159) = 0{,}276$$
$$E(B) = -\left(\tfrac{1}{6}\cdot\log\tfrac{1}{6} + \tfrac{5}{6}\cdot\log\tfrac{5}{6}\right) = -(-0{,}130 - 0{,}066) = 0{,}196$$
$$E = \tfrac{3}{9}\cdot 0{,}276 + \tfrac{6}{9}\cdot 0{,}196 = 0{,}223$$



$$E(A) = E(B) = E(C) =$$
$$= -\left(\tfrac{1}{3}\cdot\log\tfrac{1}{3} + \tfrac{2}{3}\cdot\log\tfrac{2}{3}\right) = -(-0{,}159 - 0{,}117) = 0{,}276$$
$$E = 0{,}276$$



$$E(A) = E(B) = \ldots = E(H) = E(I) =$$
$$= -\left(\tfrac{1}{1}\cdot\log\tfrac{1}{1} + \tfrac{0}{1}\cdot\log\tfrac{0}{1}\right) = -(1\cdot\log 0 + 0\cdot\log 0) = 0$$
$$E = 0$$

Fig. 68 Cálculo del valor de la entropía para cuatro soluciones de agrupamiento en relación con una clasificación externa.

Cuanto mayor es la semejanza de la solución obtenida y la clasificación externa menor es la entropía de dicha solución. El valor mínimo posible es 0 que supondría una clasificación idéntica a la externa o bien una solución trivial consistente en la división del conjunto de elementos en grupos formados por un único *ítem* (último caso).

blindLight	*k-medias*	*k-medias bisecante*	*k-medias bisecante "refinado"*	*UPGMA*	*UPGMA "refinado"*
0,4270	0,3943	0,3914	0,3988	0,3634	0,3728
Diferencia respecto a *blindLight*	8,29% Apreciable	9,10% Apreciable	7,07% Apreciable	17,50% Sustancial	14,54% Sustancial

Tabla 8. Similitud promedio de las distintas clasificaciones de la colección *wap*.

Como se puede ver en las tablas anteriores, hay dos técnicas (*k*-medias bisecante y *k*-medias bisecante refinado) que obtienen una entropía menor que *blindLight* y tres técnicas que obtienen peores resultados que la técnicas propuesta por el autor. No obstante, tan sólo hay una técnica que mejora sustancialmente los resultados obtenidos por *blindLight*. Por lo que se refiere a la similitud promedio (o lo que es lo mismo, la cohesión) de los grupos en que *blindLight* divide la colección parece ser ligeramente mejor que la de las soluciones encontradas por otros algoritmos.

En resumen, a la vista de los resultados obtenidos en los experimentos descritos en este apartado y en el anterior puede concluirse que, al menos en lo que se refiere a las colecciones *TIME*, *CIA* y *wap*, al aplicar la técnica propuesta por el autor al problema de clasificar automáticamente una colección de documentos es posible obtener unos resultados semejantes, si no mejores, que los de técnicas específicas como mapas auto-organizativos, métodos particionales y métodos jerárquicos.

4 Influencia del tamaño de los *n*-gramas en la clasificación

Al describir los experimentos anteriores no se ha hecho mención alguna al tamaño de *n*-grama utilizado. Así, en el caso de las colecciones *TIME* y *CIA* se utilizaron 4-gramas mientras que para la colección *wap* se emplearon 2-gramas. Las razones que llevaron a utilizar tamaños diferentes fueron de índole práctica: un texto cualquiera de 700 palabras (alrededor de 3400 caracteres) contiene más de 400 2-gramas diferentes, alrededor de 1.400 3-gramas y más de 2.000 4-gramas. Obtener tales vectores con sus correspondientes significatividades así como combinarlos para calcular los valores de Π y P puede resultar muy costoso para colecciones grandes. Por tanto, es necesario determinar qué influencia tiene el tamaño de *n*-grama en los resultados obtenidos por *blindLight* al clasificar un conjunto de documentos a fin de evaluar en cada caso cuál es el más idóneo.

A este fin se prepararon dos subconjuntos de las colecciones *CIA* y *wap* que contenían 50 y 156 documentos seleccionados al azar lo que suponía, respectivamente, el 20% y el 10% de las colecciones originales. En ambos casos se obtuvieron vectores de 2-, 3- y 4-gramas para los documentos y se aplicó el algoritmo de clasificación no incremental[1] a cada uno de los seis conjuntos de datos obtenidos.

En el caso del subconjunto de la colección *CIA*, a partir de ahora *CIA-50*, se procedió a calcular el valor de la similitud promedio para distintas clasificaciones obtenidas con cada una de las tres versiones obteniéndose el gráfico que se muestra en Fig. 69.

Este gráfico parece sugerir que al utilizar 3-gramas para construir los vectores de documentos se obtienen clasificaciones con una similitud promedio superior, por tanto más cohesivas y, en teoría, preferibles. No obstante, puesto que para cada versión de la colección

[1] Debido al pequeño tamaño de estos subconjuntos es posible utilizar el algoritmo no incremental que siempre produce la misma clasificación. No obstante, como se verá después, las agrupaciones obtenidas con el algoritmo incremental y con el algoritmo no incremental son básicamente equivalentes.

CIA-50 (esto es, para cada tamaño de *n*-grama) se había obtenido una clasificación diferente era posible evaluar su cohesión sobre cada una de las versiones de los datos. Por ejemplo, dada la clasificación obtenida a partir de los vectores de 2-gramas es posible calcular la similitud promedio no sólo para dichos vectores sino también para los vectores de 3- y 4-gramas aun cuando éstos no se hubiesen utilizado para obtener dicha clasificación. Lo mismo es aplicable a las otras dos clasificaciones.

Al hacer esto se comprobó (véase Fig. 70) que las tres clasificaciones obtenían aproximadamente los mismos valores de similitud promedio cuando eran evaluadas sobre la misma colección de vectores para la colección (esto es, para el mismo tamaño de *n*-grama). No obstante, deducir de esto que las clasificaciones obtenidas son equivalentes parece arriesgado al tratarse la similitud promedio de una medida de calidad "interna" (Steinbach *et al.* 2000, p. 6).

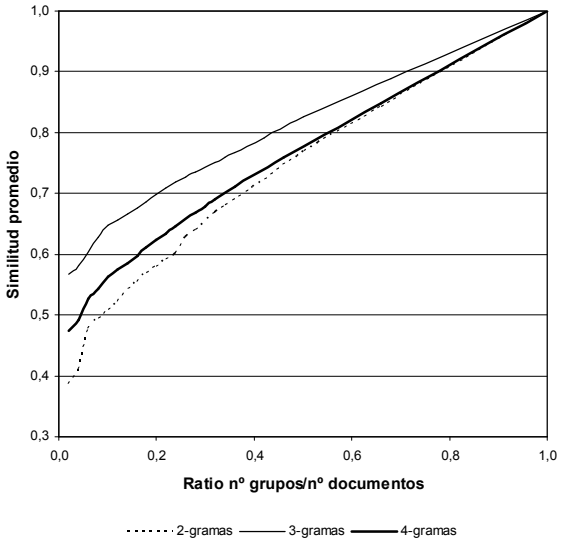


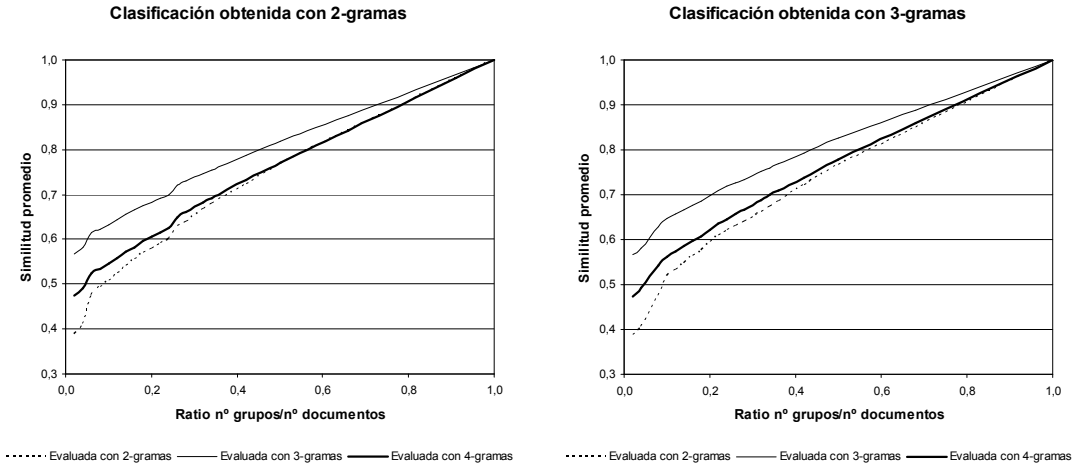
Fig. 69 Similitud promedio obtenida para cada una de las versiones de la colección *CIA-50* y clasificaciones con distinto números de grupos.

Por este motivo resultó tremendamente interesante el subconjunto *wap-156* puesto que se disponía de una clasificación externa respecto a la cual calcular la entropía (Steinbach *et al.* 2000, p. 7) y la pureza[1] (Zhao y Karypis 2002, p. 11) de las clasificaciones obtenidas aplicando *blindLight*. Los resultados obtenidos se muestran en Fig. 71. Como se puede ver, a medida que se emplean *n*-gramas de mayor tamaño para construir los vectores las clasificaciones obtenidas tienen menor entropía y una pureza superior. Las diferencias respecto al uso de 2-gramas son sensibles con independencia del número de grupos obtenidos y en el caso de 3-gramas y 4-gramas disminuyen cuando el número de grupos es del orden del número de documentos.

Así pues, como era de esperar a medida que aumenta el tamaño de *n*-grama utilizado mejora la calidad de la clasificación obtenida. Además de esto, se realizó un experimento con el subconjunto de la colección *wap* y el algoritmo incremental a fin de determinar si existen diferencias sensibles en los resultados obtenidos con ambas versiones del método. Puesto que el método incremental es estocástico se realizaron 20 ejecuciones del mismo sobre los vectores de 4-gramas (véase Tabla 9). Estos datos fueron promediados y representados junto con los resultados del algoritmo no incremental (véase Fig. 72) llegándose a la conclusión de que ambas versiones son básicamente equivalentes.

En conclusión, la técnica *blindLight* puede emplearse como método de clasificación automática de documentos obteniéndose resultados comparables e incluso mejores que los obtenidos al aplicar técnicas específicas.

[1] Recuérdese que la pureza evalúa en qué medida un grupo de una clasificación contiene documentos de una única clase.



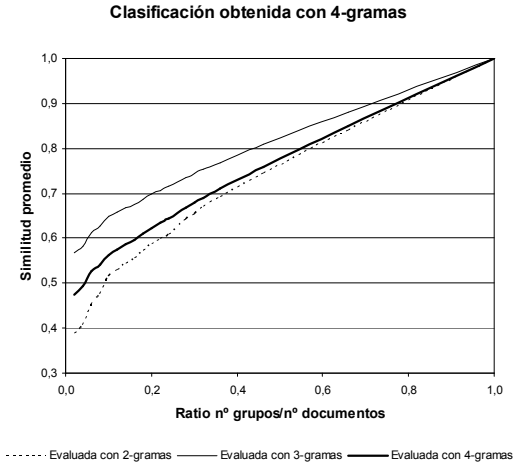


Fig. 70 Similitud promedio de las tres clasificaciones evaluadas sobre las tres versiones de la colección.

Grupos obtenidos	Entropía	Pureza
15	1,6040	0,4423
16	1,4882	0,4936
16	1,5042	0,4808
16	1,4104	0,4936
17	1,0654	0,5962
17	1,5967	0,4167
17	1,4087	0,4872
17	1,4694	0,5000
17	1,5058	0,4808
17	1,4620	0,4808
18	1,3639	0,4936
18	1,4449	0,4744
18	1,2991	0,5256
19	1,2248	0,5513
19	1,3884	0,4744
19	1,1151	0,6029
19	1,3815	0,5000
19	1,5627	0,4359
20	1,3761	0,4936
20	1,0529	0,5897

Tabla 9. Resultados obtenidos al ejecutar 20 veces el algoritmo incremental sobre el subconjunto de la colección *wap-156* elaborada con 4-gramas.

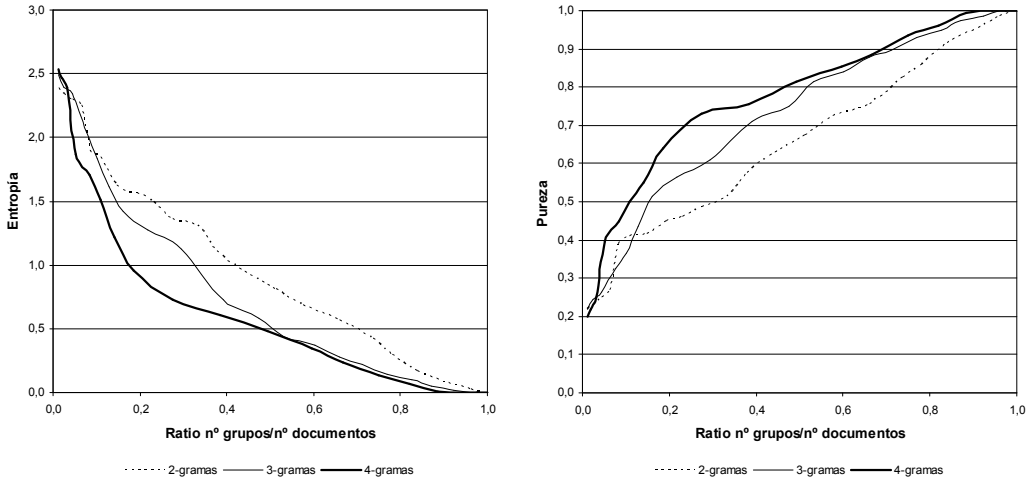


Fig. 71 Entropía y pureza de las clasificaciones obtenidas al utilizar distintos tamaños de *n*-grama en la clasificación de la colección *wap-156*.

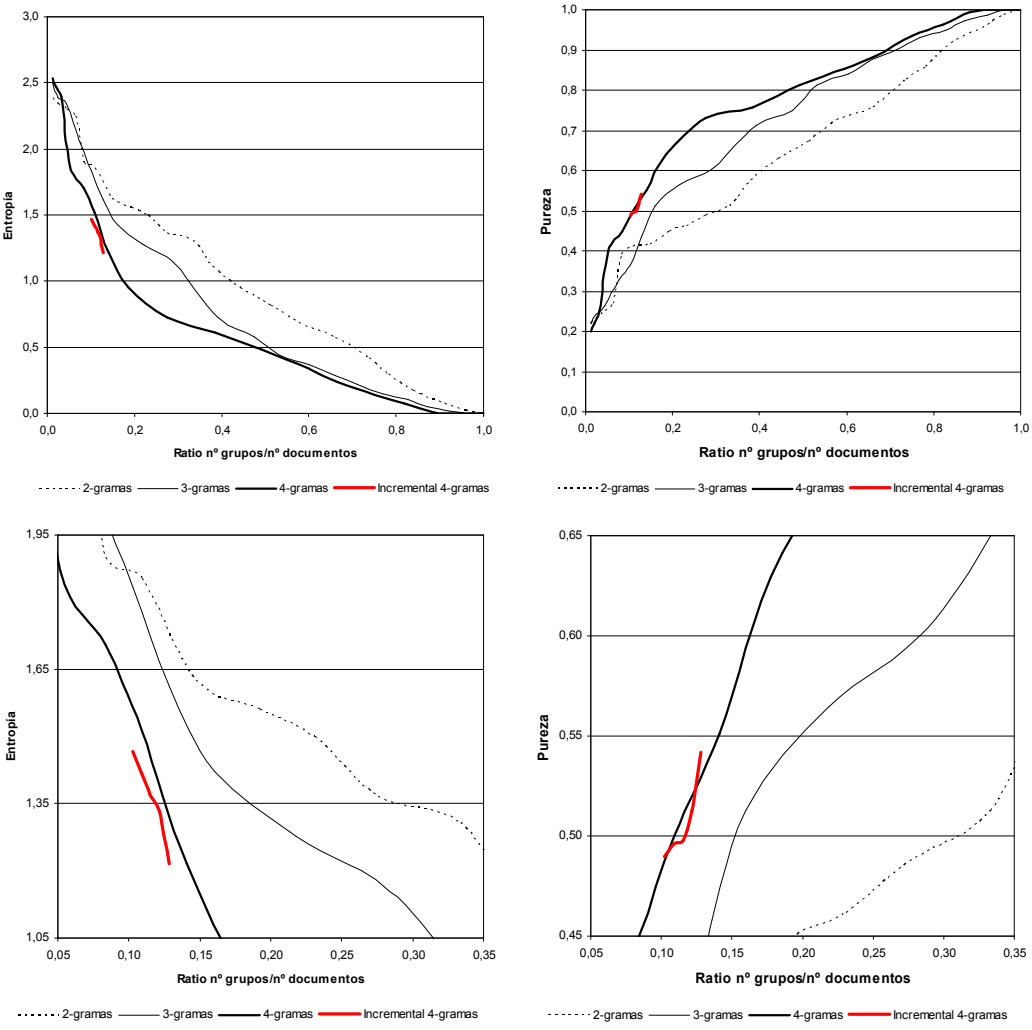


Fig. 72 Resultados de entropía y pureza obtenidos por el algoritmo incremental comparados con los obtenidos por la versión no incremental para la colección *wap-156*.