

Descubrimiento de Información en Textos

Tarea del Tema 6: Técnicas de minería de textos. (Clustering)

Jose Alberto Benítez Andrades

71454586A

Descubrimiento de Información en Textos

Máster en Lenguajes y Sistemas Informáticos - Tecnologías del Lenguaje en la Web

UNED

25/05/2011

0.Enunciado

Tarea obligatoria del Tema 6. Técnicas de Minería de Textos.*Clustering*

En esta tarea se pretende que los alumnos utilicen un software de *clustering* y se familiaricen con este tipo de programas, sus requerimientos y el tipo de salida que generan. En concreto, se propone utilizar **CLUTO - Family of Data Clustering Software Tools** que se encuentra en <http://glaros.dtc.umn.edu/gkhome/views/cluto/>.

De las tres familias de herramientas que aparecen en esa página habrá que utilizar: **CLUTO - Software for Clustering High-Dimensional Datasets** <http://glaros.dtc.umn.edu/gkhome/cluto/cluto/overview>. En la pestaña "Download" se puede descargar el software. Se recomienda la descarga de: Latest stable release (2.1.1). En esta misma página se puede encontrar información de la instalación y en "Documentation", se puede encontrar el enlace al **manual de uso** del programa.

Los programas de CLUTO tienen dos formas de ser utilizados: como un programa en sí mismo, al que se le llama desde la línea de comando, o como parte de un programa C. En nuestro caso, se utilizará de la primera manera y en el manual de uso habrá que consultar: **Using CLUTO via its Stand-Alone Program** (Apartado 3 del manual de uso).

Se van a utilizar como datos de prueba los que proporciona CLUTO que se encuentran accesibles en esa misma página como "Datasets": [datasets.tar.gz](http://glaros.dtc.umn.edu/gkhome/node/165). La descripción de lo que contienen los conjuntos de datos se puede encontrar en el artículo "**Criterion Functions for Document Clustering**" <http://glaros.dtc.umn.edu/gkhome/node/165>, en el apartado 4.1 Document Collections. En ese artículo también se definen algunos conceptos básicos de clustering, y se tratan algunas de las variables que intervienen en el clustering, como por ejemplo las funciones criterio.

La realización de la práctica consiste en lo siguiente:

- Descarga del software necesario, del manual de uso, conjunto de datos con los que se va a experimentar, su descripción y toda la información adicional que se considere necesaria.
- Seleccionar de los datasets de CLUTO aquellos con los que se va a realizar la experimentación. Serán necesarios 2 datasets con el fin de poder realizar comparaciones de comportamiento.
- A no ser que algún alumno tenga un interés especial en otras colecciones, se recomienda seleccionar las colecciones "re0" y "re1". La descripción de los datasets seleccionados se encuentra en el artículo "**Criterion Functions for Document Clustering**" y en la web. Por ejemplo, uno de los dataset de la agencia de noticias reuters, re0, es un subconjunto de la colección Reuters- 21578, tiene 1504 documentos, 2886 términos o rasgos y 13 clases o clusters. (ver tabla 1 en el artículo), en la web se puede encontrar información adicional de esta colección, por ejemplo en <http://www.daviddlewis.com/resources/testcollections/reuters21578/>.
- Utilizar el programa *vcluster* que usa como entrada una matriz que corresponde a un fichero con extensión .mat de los que aparecen en [datasets.tar.gz](http://glaros.dtc.umn.edu/gkhome/node/165). Además, mediante

25 de mayo de 2011

parámetros, se le pueden indicar que haga el clustering con diferentes algoritmos, funciones criterio, etc. (Ver apartado 3.1.1 del manual de uso).

- Hacer el análisis de los resultados del clustering con las dos colecciones con al menos:
 - Un algoritmo de partición y el algoritmo aglomerativo.
 - Dos funciones de similitud.
 - Dos funciones criterio.
- Preparar una memoria en la que:
 - Se describan los datasets seleccionados.
 - Se indiquen los parámetros de *vcluster* que se han utilizado indicando el tipo de algoritmo, funciones de similitud y criterio, así como cualquier otra variación de los parámetros por defecto que se haya utilizado.
 - Se añada la información de salida de las ejecuciones de *vcluster*. (En la sección 3.2 del manual de uso se explica la información que se obtiene). La calidad del clustering se mide con medidas internas y externas. Para obtener información de la calidad estadística del clustering comparando la solución del algoritmo con una solución manual (medida externa) se debe utilizar el parámetro – *rclassfile* en la ejecución de *vcluster* junto con los ficheros con extensión *.rclass* de [datasets.tar.gz](#).
 - Se analicen los resultados en términos de calidad de las medidas internas y externas con los diferentes tipos de algoritmos, funciones de similitud y criterio.

La fecha tope de entrega de esta tarea será el 27 de mayo de 2011.

1.Descripción de los dataset utilizados

Los dataset utilizados han sido *re0.mat* y *re1.mat*, que son los propuestos en el enunciado de este trabajo.

El fichero *re0.mat* tiene: 1504 documentos, 2886 términos y 13 clases.

El fichero *re1.mat* tiene: 1657 documentos, 3758 términos y 25 clases.

A continuación detallo las pruebas realizadas con distintos algoritmos, funciones de similitud y funciones de criterio a estos dos ficheros.

25 de mayo de 2011

2.Pruebas realizadas y análisis de resultados.

He utilizado la versión del programa CLUTO para Linux, creando un script que generaba los ficheros .txt que me interesaban, acorde con la información que deseaba guardar para realizar su posterior estudio.

El script se llama **ejecutor.sh** y contiene el siguiente código:

```
#!/bin/bash

./vcluster -clmethod='agglo' -sim='cos' -crfun='i1' re0.mat 10 > r0-agglo-cos-i1.txt
./vcluster -clmethod='agglo' -sim='cos' -crfun='i2' re0.mat 10 > r0-agglo-cos-i2.txt

./vcluster -clmethod='graph' -sim='dist' -crfun='i1' re0.mat 10 > r0-graph-dist-i1.txt
./vcluster -clmethod='graph' -sim='jacc' -crfun='i1' re0.mat 10 > r0-graph-jacc-i1.txt

./vcluster -clmethod='agglo' -sim='cos' -crfun='i1' re1.mat 10 > r1-agglo-cos-i1.txt
./vcluster -clmethod='agglo' -sim='cos' -crfun='i2' re1.mat 10 > r1-agglo-cos-i2.txt

./vcluster -clmethod='graph' -sim='dist' -crfun='i1' re1.mat 10 > r1-graph-dist-i1.txt
./vcluster -clmethod='graph' -sim='jacc' -crfun='i1' re1.mat 10 > r1-graph-jacc-i1.txt
```

```
# CON ESTADISTICAS

./vcluster -clmethod='agglo' -sim='cos' -crfun='i1' -rclassfile=estadisticas/re0.mat.rclass re0.mat 10 > r0-agglo-cos-i1-EST.txt
./vcluster -clmethod='agglo' -sim='cos' -crfun='i2' -rclassfile=estadisticas/re0.mat.rclass re0.mat 10 > r0-agglo-cos-i2-EST.txt

./vcluster -clmethod='graph' -sim='dist' -crfun='i1' -rclassfile=estadisticas/re0.mat.rclass re0.mat 10 > r0-graph-dist-i1-EST.txt
./vcluster -clmethod='graph' -sim='jacc' -crfun='i1' -rclassfile=estadisticas/re0.mat.rclass re0.mat 10 > r0-graph-jacc-i1-EST.txt
./vcluster -clmethod='agglo' -sim='cos' -crfun='i1' -rclassfile=estadisticas/re1.mat.rclass re1.mat 10 > r1-agglo-cos-i1-EST.txt
./vcluster -clmethod='agglo' -sim='cos' -crfun='i2' -rclassfile=estadisticas/re1.mat.rclass re1.mat 10 > r1-agglo-cos-i2-EST.txt
./vcluster -clmethod='graph' -sim='dist' -crfun='i1' -rclassfile=estadisticas/re1.mat.rclass re1.mat 10 > r1-graph-dist-i1-EST.txt
./vcluster -clmethod='graph' -sim='jacc' -crfun='i1' -rclassfile=estadisticas/re1.mat.rclass re1.mat 10 > r1-graph-jacc-i1-EST.txt
```

Siguiendo el orden del script, a continuación mostraré las salidas obtenidas y haré un breve análisis, explicando distintos elementos.

25 de mayo de 2011

VCluster aplicado al fichero re0.mat

Con un algoritmo aglomerativo, función de similitud cosine y función de criterio i1

Línea de comando:

```
./vcluster -clmethod='agglo' -sim='cos' -crfun='i1' re0.mat 10
```

Salida obtenida

```

root@licantropo: /home/jabenitez/master/dit/cluto-2.1.1/Linux
root@licantropo:/home/jabenitez/master/dit/cluto-2.1.1/Linux# ./vcluster -clmethod='agglo' -sim='cos' -crfun='i1' re0.mat 10
*****
vcluster (CLUTO 2.1.1) Copyright 2001-03, Regents of the University of Minnesota

Matrix Information -----
  Name: re0.mat, #Rows: 1504, #Columns: 2886, #NonZeros: 77808

Options -----
  CLMethod=AGGLO, CRfun=I1, SimFun=Cosine, #Clusters: 10
  RowModel=None, ColModel=IDF, GrModel=SY-DIR, NNbrs=40
  Colprune=1.00, EdgePrune=-1.00, VtxPrune=-1.00, MinComponent=5
  CStype=Best, AggloFrom=0, AggloCRFun=I1, NTrials=10, NIter=10

Solution -----

10-way clustering: [I1=2.36e+02] [1504 of 1504]

-----
cid  Size  ISim  ISdev  ESim  ESdev  |
-----
  0   103  +0.528 +0.100 +0.034 +0.006 |
  1    51  +0.595 +0.080 +0.036 +0.006 |
  2   104  +0.160 +0.049 +0.045 +0.017 |
  3    78  +0.269 +0.060 +0.045 +0.012 |
  4   134  +0.159 +0.049 +0.033 +0.014 |
  5   604  +0.048 +0.016 +0.037 +0.014 |
  6   174  +0.141 +0.044 +0.034 +0.011 |
  7    48  +0.254 +0.064 +0.026 +0.012 |
  8    35  +0.301 +0.048 +0.029 +0.009 |
  9   173  +0.093 +0.029 +0.028 +0.010 |
-----

Timing Information -----
  I/O:                                0.064 sec
  Clustering:                          1.148 sec
  Reporting:                           0.008 sec
*****
root@licantropo:/home/jabenitez/master/dit/cluto-2.1.1/Linux#

```

En primer lugar, comentar de forma muy breve el significado de cada columna:

- cid: Representa la id de cada cluster creado a partir del documento que hemos querido empaquetar mediante esta técnica de clustering.
- Size: cantidad de objetos que contiene cada uno de los 10 clusters que hemos creado.
- ISim: Número que muestra la media de similitud entre los objetos del cluster.
- ISdev: Media de la desviación de las similitudes entre objetos del cluster.
- ESim: Similitud de los objetos de cada cluster y del resto.
- ESdev: Desviación de las similitudes de los objetos de cada cluster y del resto.

25 de mayo de 2011

Aplicando la siguiente línea de comando:

```
./vcluster -clmethod='agglo' -sim='cos' -crfun='i1' -rclassfile=estadisticas/re0.mat.rclass re0.mat 10
```

```

root@licantropo: /home/jabenitez/master/dit/cluto-2.1.1/Linux
*****
root@licantropo: /home/jabenitez/master/dit/cluto-2.1.1/Linux# ./vcluster -clmethod='agglo' -sim='cos' -crfun='i1' -rclassfile=estadisticas/re0.mat.rclass re0.mat 10
*****
vcluster (CLUTO 2.1.1) Copyright 2001-03, Regents of the University of Minnesota

Matrix Information -----
Name: re0.mat, #Rows: 1504, #Columns: 2886, #NonZeros: 77808

Options -----
CLMethod=AGGLO, CRfun=I1, SimFun=Cosine, #Clusters: 10
RowModel=None, ColModel=IDF, GrModel=SY-DIR, NNbrs=40
Colprune=1.00, EdgePrune=-1.00, VtxPrune=-1.00, MinComponent=5
CSType=Best, AggloFrom=0, AggloCRFun=I1, NTrials=10, NIter=10

Solution -----

10-way clustering: [I1=2.36e+02] [1504 of 1504], Entropy: 0.488, Purity: 0.549

cid Size ISim ISdev ESim ESdev Entpy Purty | hous mone trad rese cpi inte gnp reta ipi jobs lei bop wpi
-----
0 103 +0.528 +0.100 +0.034 +0.006 0.051 0.971 | 0 100 0 0 0 3 0 0 0 0 0 0 0 0 0
1 51 +0.595 +0.080 +0.036 +0.006 0.253 0.647 | 0 33 0 0 0 18 0 0 0 0 0 0 0 0 0
2 104 +0.160 +0.049 +0.045 +0.017 0.677 0.317 | 1 17 33 20 1 2 1 1 1 0 2 24 1
3 78 +0.269 +0.060 +0.045 +0.012 0.064 0.962 | 0 75 0 0 0 3 0 0 0 0 0 0 0 0 0
4 134 +0.159 +0.049 +0.033 +0.014 0.142 0.910 | 0 122 6 0 0 6 0 0 0 0 0 0 0 0 0
5 604 +0.048 +0.016 +0.037 +0.014 0.699 0.336 | 1 203 114 22 17 132 69 8 16 11 0 11 0
6 174 +0.141 +0.044 +0.034 +0.011 0.838 0.236 | 13 30 2 0 41 1 8 10 18 27 7 3 14
7 48 +0.254 +0.064 +0.026 +0.012 0.000 1.000 | 0 0 0 0 0 48 0 0 0 0 0 0 0 0
8 35 +0.301 +0.048 +0.029 +0.009 0.369 0.543 | 0 19 12 0 0 4 0 0 0 0 0 0 0 0
9 173 +0.093 +0.029 +0.028 +0.010 0.231 0.879 | 1 9 152 0 1 2 2 1 2 1 2 0 0 0

Timing Information -----
I/O: 0.068 sec
Clustering: 1.160 sec
Reporting: 0.012 sec
*****
root@licantropo: /home/jabenitez/master/dit/cluto-2.1.1/Linux#

```

Se obtienen 2 columnas más de información:

- Entpy (Entropy): Muestra el índice de objetos de distinto tipo empaquetados en un clúster.
- Purty (Purity): Nos da un valor que varía entre 0 y 1 en función de la similitud de la clase de los objetos empaquetados en un clúster concreto.

Sabiendo esto, en este primer caso se puede destacar que:

- El clúster 7 realiza un empaquetado perfecto: posee una pureza de valor 1 debido a que contiene 48 objetos que son del mismo tipo 'inte'.
- Los clústers 0, 3 y 4 tienen también un alto nivel de pureza mientras que por ejemplo, el clúster 6 tiene muchos paquetes de distinto tipo, con lo que la pureza es bastante pequeña.

25 de mayo de 2011

Con un algoritmo aglomerativo, función de similitud cosine y función de criterio i2

Línea de comando:

`./vcluster -clmethod='agglo' -sim='cos' -crfun='i2' re0.mat 10`

Salida obtenida

```

root@licantropo: /home/jabenitez/master/dit/cluto-2.1.1/Linux
root@licantropo:/home/jabenitez/master/dit/cluto-2.1.1/Linux# ./vcluster -clmethod='agglo' -sim='cos' -crfun='i2' re0.mat 10
*****
vcluster (CLUTO 2.1.1) Copyright 2001-03, Regents of the University of Minnesota

Matrix Information -----
  Name: re0.mat, #Rows: 1504, #Columns: 2886, #NonZeros: 77808

Options -----
  CLMethod=AGGLO, CRfun=I2, SimFun=Cosine, #Clusters: 10
  RowModel=None, ColModel=IDF, GrModel=SY-DIR, NNbrs=40
  Colprune=1.00, EdgePrune=-1.00, VtxPrune=-1.00, MinComponent=5
  CStype=Best, AggloFrom=0, AggloCRFun=I2, NTrials=10, NIter=10

Solution -----

10-way clustering: [I2=5.47e+02] [1504 of 1504]

-----
cid  Size  ISim  ISdev  ESim  ESdev  |
-----
  0   104 +0.512 +0.115 +0.034 +0.005 |
  1   255 +0.060 +0.021 +0.033 +0.013 |
  2   317 +0.093 +0.034 +0.034 +0.013 |
  3   214 +0.081 +0.027 +0.033 +0.015 |
  4    57 +0.548 +0.095 +0.036 +0.007 |
  5    96 +0.199 +0.045 +0.034 +0.015 |
  6    82 +0.194 +0.058 +0.038 +0.014 |
  7   176 +0.091 +0.029 +0.027 +0.010 |
  8    74 +0.168 +0.036 +0.043 +0.013 |
  9   129 +0.136 +0.052 +0.045 +0.015 |
-----

Timing Information -----
  I/O:                                0.068 sec
  Clustering:                          1.200 sec
  Reporting:                           0.008 sec
*****
root@licantropo:/home/jabenitez/master/dit/cluto-2.1.1/Linux#

```

25 de mayo de 2011

Aplicando la siguiente línea de comando:

```
./vcluster -clmethod='agglo' -sim='cos' -crfun='i2' -rclassfile=estadisticas/re0.mat.rclass re0.mat 10
```

```

root@licantropo: /home/jabenitez/master/dit/cluto-2.1.1/Linux
*****
root@licantropo:/home/jabenitez/master/dit/cluto-2.1.1/Linux# ./vcluster -clmethod='agglo' -sim='cos' -crfun='i2' -rcl
sticas/re0.mat.rclass re0.mat 10
*****
vcluster (CLUTO 2.1.1) Copyright 2001-03, Regents of the University of Minnesota

Matrix Information -----
  Name: re0.mat, #Rows: 1504, #Columns: 2886, #NonZeros: 77808

Options -----
  CLMethod=AGGLO, CRfun=I2, SimFun=Cosine, #Clusters: 10
  RowModel=None, ColModel=IDF, GrModel=SY-DIR, NNbrs=40
  ColPrune=1.00, EdgePrune=-1.00, VtxPrune=-1.00, MinComponent=5
  CStype=Best, AggloFrom=0, AggloCRFun=I2, NTrials=10, NIter=10

Solution -----

10-way clustering: [I2=5.47e+02] [1504 of 1504], Entropy: 0.481, Purity: 0.548

cid  Size  ISim  ISdev  ESim  ESdev  Entpy  Purty  |  hous  mone  trad  rese  cpi  inte  gnp  reta  ipi  jobs  lei  bop  wpi
-----
0    104 +0.512 +0.115 +0.034 +0.005 0.037 0.981 | 0 102 0 0 0 2 0 0 0 0 0 0
1    255 +0.060 +0.021 +0.033 +0.013 0.532 0.447 | 0 114 87 3 3 28 12 1 2 4 1 0 0
2    317 +0.093 +0.034 +0.034 +0.013 0.908 0.177 | 14 49 19 5 53 15 56 16 28 32 7 9 14
3    214 +0.081 +0.027 +0.033 +0.015 0.337 0.631 | 0 68 3 1 0 135 6 0 0 0 0 0 1 0
4    57 +0.548 +0.095 +0.036 +0.007 0.248 0.667 | 0 38 0 0 0 19 0 0 0 0 0 0 0 0
5    96 +0.199 +0.045 +0.034 +0.015 0.091 0.938 | 0 90 0 0 0 6 0 0 0 0 0 0 0 0
6    82 +0.194 +0.058 +0.038 +0.014 0.567 0.646 | 1 53 4 3 3 2 0 3 2 2 2 6 1
7   176 +0.091 +0.029 +0.027 +0.010 0.221 0.858 | 0 17 151 0 1 0 2 0 3 1 1 0 0
8    74 +0.168 +0.036 +0.043 +0.013 0.568 0.500 | 0 37 7 11 0 10 3 0 0 0 0 0 6 0
9   129 +0.136 +0.052 +0.045 +0.015 0.576 0.372 | 1 40 48 19 0 2 1 0 2 0 0 16 0

Timing Information -----
  I/O:                                0.072 sec
  Clustering:                          1.184 sec
  Reporting:                           0.008 sec
*****
root@licantropo:/home/jabenitez/master/dit/cluto-2.1.1/Linux#

```

A diferencia de los parámetros usados en el punto anterior, aplicando estos criterios se observa que la Pureza general (que anteriormente era de 0.549) es prácticamente igual a la obtenida con esta variación, habiendo utilizado la función de criterio i2, que ha sido de 0.548.

Sin embargo, los clústers que se han formado son totalmente distintos, destacan el 0 y el 5 que tienen una pureza muy alta, pero sin embargo en el resto de clústers, en lugar de tener 4 clústers con una pureza alta y 6 con pureza algo más baja, tenemos una pureza más o menos acorde con la media general en cada clúster.

25 de mayo de 2011

Con un algoritmo de separación, función de similitud de distancia euclidea y función de criterio i1

Línea de comando:

`./vcluster -clmethod='graph' -sim='dist' -crfun='i1' re0.mat 10`

Salida obtenida

```

root@licantropo: /home/jabenitez/master/dit/cluto-2.1.1/Linux
root@licantropo:/home/jabenitez/master/dit/cluto-2.1.1/Linux# ./vcluster -clmethod='graph' -sim='dist' -crfun='i1' re0.mat 10
*****
vcluster (CLUTO 2.1.1) Copyright 2001-03, Regents of the University of Minnesota

Matrix Information -----
  Name: re0.mat, #Rows: 1504, #Columns: 2886, #NonZeros: 77808

Options -----
  CLMethod=GRAPH, CRfun=Cut, SimFun=EuclDist, #Clusters: 10
  RowModel=None, ColModel=IDF, GrModel=SY-DIR, NNbrs=40
  Colprune=1.00, EdgePrune=-1.00, VtxPrune=-1.00, MinComponent=5
  CStype=Best, AggloFrom=0, AggloCRFun=SLINK_W, NTrials=10, NIter=10

Solution -----

29-way clustering: [Cut=3.08e+03] [458 of 1504]

-----
cid  Size  ISim  ISdev  ESim  ESdev  |
-----
  0    20 +0.934 +0.102 +0.000 +0.000 |
  1     7 +0.995 +0.003 +0.000 +0.000 |
  2     7 +0.982 +0.004 +0.000 +0.000 |
  3     6 +0.997 +0.001 +0.000 +0.000 |
  4     5 +0.997 +0.002 +0.000 +0.000 |
  5     9 +0.881 +0.159 +0.000 +0.000 |
  6    17 +0.804 +0.201 +0.014 +0.008 |
  7    37 +0.689 +0.213 +0.005 +0.010 |
  8    11 +0.672 +0.200 +0.023 +0.018 |
  9     5 +0.666 +0.261 +0.000 +0.000 |
 10    48 +0.539 +0.193 +0.009 +0.012 |
 11     6 +0.598 +0.219 +0.000 +0.000 |
 12     5 +0.593 +0.218 +0.000 +0.000 |
 13    36 +0.484 +0.233 +0.000 +0.000 |
 14    22 +0.491 +0.234 +0.002 +0.007 |
 15    25 +0.479 +0.220 +0.026 +0.019 |
 16    11 +0.471 +0.236 +0.000 +0.000 |
 17     5 +0.474 +0.179 +0.000 +0.000 |
 18    15 +0.361 +0.167 +0.003 +0.003 |
 19     6 +0.391 +0.214 +0.000 +0.000 |
 20     6 +0.386 +0.214 +0.000 +0.000 |
 21     6 +0.361 +0.167 +0.000 +0.000 |
 22    15 +0.304 +0.141 +0.001 +0.001 |
 23    14 +0.307 +0.100 +0.006 +0.005 |
 24     7 +0.319 +0.129 +0.000 +0.000 |
 25    35 +0.280 +0.157 +0.010 +0.014 |
 26     8 +0.299 +0.166 +0.000 +0.000 |
 27    36 +0.196 +0.122 +0.003 +0.004 |
 28    28 +0.089 +0.066 +0.000 +0.000 |
-----

Timing Information -----
  I/O:                                0.060 sec
  Clustering:                          0.340 sec
  Reporting:                           -0.000 sec
*****
root@licantropo:/home/jabenitez/master/dit/cluto-2.1.1/Linux#

```

25 de mayo de 2011

Comando:

```
./vcluster -clmethod='graph' -sim='dist' -crfun='i1' -rclassfile=estadisticas/re0.mat.rclass
re0.mat 10
```

Salida obtenida:

```

root@licantropo: /home/jabenitez/master/dit/cluto-2.1.1/Linux

29-way clustering: [Cut=3.08e+03] [458 of 1504], Entropy: 0.264, Purity: 0.723

cid  Size  ISim  ISdev  ESIm  ESdev  Entpy  Purty | hous  mone  trad  rese  cpi  inte  gnp  reta  ipi  jobs  lei  bop  wpi
-----
0    20 +0.934 +0.102 +0.000 +0.000 0.000 1.000 | 0    20    0    0    0    0    0    0    0    0    0    0    0
1     7 +0.995 +0.003 +0.000 +0.000 0.160 0.857 | 0     6    0    0    0    1    0    0    0    0    0    0    0
2     7 +0.982 +0.004 +0.000 +0.000 0.000 1.000 | 0     7    0    0    0    0    0    0    0    0    0    0    0
3     6 +0.997 +0.001 +0.000 +0.000 0.000 1.000 | 0     0    0    0    0    6    0    0    0    0    0    0    0
4     5 +0.997 +0.002 +0.000 +0.000 0.000 1.000 | 0     5    0    0    0    0    0    0    0    0    0    0    0
5     9 +0.881 +0.159 +0.000 +0.000 0.000 1.000 | 0     9    0    0    0    0    0    0    0    0    0    0    0
6    17 +0.804 +0.201 +0.014 +0.008 0.000 1.000 | 0    17    0    0    0    0    0    0    0    0    0    0    0
7    37 +0.689 +0.213 +0.005 +0.010 0.000 1.000 | 0    37    0    0    0    0    0    0    0    0    0    0    0
8    11 +0.672 +0.200 +0.023 +0.018 0.234 0.818 | 0     9    1    0    0    0    0    0    0    1    0    0    0
9     5 +0.666 +0.261 +0.000 +0.000 0.000 1.000 | 0     5    0    0    0    0    0    0    0    0    0    0    0
10   48 +0.539 +0.193 +0.009 +0.012 0.185 0.854 | 0    41    1    0    6    0    0    0    0    0    0    0    0
11     6 +0.598 +0.219 +0.000 +0.000 0.338 0.667 | 0     4    1    0    1    0    0    0    0    0    0    0    0
12     5 +0.593 +0.218 +0.000 +0.000 0.000 1.000 | 0     0    0    0    5    0    0    0    0    0    0    0    0
13   36 +0.484 +0.233 +0.000 +0.000 0.261 0.611 | 0    22    0    0    0    14    0    0    0    0    0    0    0
14   22 +0.491 +0.234 +0.002 +0.007 0.000 1.000 | 0    22    0    0    0    0    0    0    0    0    0    0    0
15   25 +0.479 +0.220 +0.026 +0.019 0.736 0.360 | 0     1    0    1    4     9    3    0    3    2    0    1    1
16   11 +0.471 +0.236 +0.000 +0.000 0.000 1.000 | 0     0    0    0    0    11    0    0    0    0    0    0    0
17     5 +0.474 +0.179 +0.000 +0.000 0.000 1.000 | 0     0    5    0    0    0    0    0    0    0    0    0    0
18   15 +0.361 +0.167 +0.003 +0.003 0.386 0.467 | 0     6    7    0    2    0    0    0    0    0    0    0    0
19     6 +0.391 +0.214 +0.000 +0.000 0.394 0.500 | 0     0    3    2    0    0    0    0    0    0    0    1    0
20     6 +0.386 +0.214 +0.000 +0.000 0.394 0.500 | 0     0    3    1    0    0    0    0    0    0    0    2    0
21     6 +0.361 +0.167 +0.000 +0.000 0.176 0.833 | 0     0    1    0    0    0    0    0    0    0    0    5    0
22   15 +0.304 +0.141 +0.001 +0.001 0.493 0.333 | 0     5    1    0    0    0    5    0    0    0    0    4    0
23   14 +0.307 +0.100 +0.006 +0.005 0.481 0.500 | 0     2    0    3    0    7    0    0    0    2    0    0    0
24     7 +0.319 +0.129 +0.000 +0.000 0.000 1.000 | 0     7    0    0    0    0    0    0    0    0    0    0    0
25   35 +0.280 +0.157 +0.010 +0.014 0.179 0.829 | 0    29    0    0    0    6    0    0    0    0    0    0    0
26     8 +0.299 +0.166 +0.000 +0.000 0.422 0.375 | 0     3    0    2    0    3    0    0    0    0    0    0    0
27   36 +0.196 +0.122 +0.003 +0.004 0.863 0.194 | 1     3    5    2    2     7    0    3    2    0    3    7    1
28   28 +0.089 +0.066 +0.000 +0.000 0.432 0.536 | 0     0    0    0    15    0    0    1    0    5    0    0    7

Timing Information -----
I/O:                                0.036 sec
Clustering:                         0.284 sec
Reporting:                          -0.000 sec
*****
root@licantropo: /home/jabenitez/master/dit/cluto-2.1.1/Linux#

```

En esta aplicación, se observa que el programa genera un mayor número de clústers, con un menor número de objetos almacenados en cada uno, pero con una pureza bastante alta, la media es de 0.723, frente a 0.548 y 0.549 de la anterior aplicación.

Esto indica que siguiendo un algoritmo de separación y una función de selección mediante distancia euclídea, se obtiene un mejor empaquetado mediante clustering.

25 de mayo de 2011

Con un algoritmo de separación, función de similitud del coeficiente de Jaccard y función de criterio i1

Línea de comando:

```
./vcluster -clmethod='graph' -sim='jacc' -crfun='i1' re0.mat 10
```

Salida obtenida

```

root@licantropo: /home/jabenitez/master/dit/cluto-2.1.1/Linux
*****
root@licantropo:/home/jabenitez/master/dit/cluto-2.1.1/Linux# ./vcluster -clmethod='gra
ph' -sim='jacc' -crfun='i1' re0.mat 10
*****
vcluster (CLUTO 2.1.1) Copyright 2001-03, Regents of the University of Minnesota

Matrix Information -----
  Name: re0.mat, #Rows: 1504, #Columns: 2886, #NonZeros: 77808

Options -----
  CLMethod=GRAPH, CRfun=Cut, SimFun=ExtJaccard, #Clusters: 10
  RowModel=None, ColModel=IDF, GrModel=SY-DIR, NNbrs=40
  Colprune=1.00, EdgePrune=-1.00, VtxPrune=-1.00, MinComponent=5
  CStype=Best, AggloFrom=0, AggloCRFun=SLINK_W, NTrials=10, NIter=10

Solution -----

-----
10-way clustering: [Cut=4.06e+03] [1502 of 1504]
-----
cid  Size  ISim  ISdev  ESim  ESdev  |
-----
  0    20 +0.693 +0.097 +0.001 +0.001 |
  1    57 +0.347 +0.142 +0.000 +0.001 |
  2    23 +0.287 +0.118 +0.001 +0.001 |
  3    97 +0.101 +0.084 +0.000 +0.000 |
  4   140 +0.032 +0.020 +0.001 +0.000 |
  5   111 +0.031 +0.018 +0.000 +0.000 |
  6   171 +0.022 +0.014 +0.001 +0.000 |
  7   245 +0.015 +0.008 +0.000 +0.000 |
  8   239 +0.012 +0.008 +0.000 +0.000 |
  9   399 +0.010 +0.005 +0.001 +0.001 |
-----

Timing Information -----
  I/O:                                0.068 sec
  Clustering:                          0.660 sec
  Reporting:                           0.004 sec
*****
root@licantropo:/home/jabenitez/master/dit/cluto-2.1.1/Linux#

```

25 de mayo de 2011

Comando:

```
./vcluster -clmethod='graph' -sim='jacc' -crfun='i1' -rclassfile=estadisticas/re0.mat.rclass
re0.mat 10
```

Salida obtenida:

```

root@licantropo: /home/jabenitez/master/dit/cluto-2.1.1/Linux

Options -----
CLMethod=GRAPH, CRfun=Cut, SimFun=ExtJaccard, #Clusters: 10
RowModel=None, ColModel=IDF, GrModel=SY-DIR, NNbrs=40
ColPrune=1.00, EdgePrune=-1.00, VtxPrune=-1.00, MinComponent=5
CSType=Best, AggloFrom=0, AggloCRFun=SLINK_W, NTrials=10, NIter=10

Solution -----

10-way clustering: [Cut=4.06e+03] [1502 of 1504], Entropy: 0.500, Purity: 0.531

cid  Size  ISim  ISdev  ESim  ESdev  Entpy  Purty | hous mone trad rese  cpi  inte  gnp  reta  ipi  jobs  lei  bop  wpi
-----
0    20 +0.693 +0.097 +0.001 +0.001 0.000 1.000 |  0  20  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
1    57 +0.347 +0.142 +0.000 +0.001 0.000 1.000 |  0  57  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
2    23 +0.287 +0.118 +0.001 +0.001 0.000 1.000 |  0  23  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
3    97 +0.101 +0.084 +0.000 +0.000 0.347 0.660 |  0  64  6  1  0  25  0  0  0  1  0  0  0  0  0  0  0
4   140 +0.032 +0.020 +0.001 +0.000 0.801 0.407 |  2  57 11  7  8  14  9  4  6  5  3 12  2
5   111 +0.031 +0.018 +0.000 +0.000 0.335 0.712 |  0  79 19  0  1 11  1  0  0  0  0  0  0  0  0  0
6   171 +0.022 +0.014 +0.001 +0.000 0.826 0.257 | 12  4 14  0 44  1  8 13 26 28  8  0 13
7   245 +0.015 +0.008 +0.000 +0.000 0.309 0.498 |  0 118  3  0  0 122  2  0  0  0  0  0  0  0  0
8   239 +0.012 +0.008 +0.000 +0.000 0.323 0.762 |  0  36 182  1  2 12  3  1  1  1  0  0  0  0
9   399 +0.010 +0.005 +0.001 +0.001 0.686 0.376 |  2 150 84 33  5 32 57  2  3  5  0 26  0

Timing Information -----
I/O:                                0.048 sec
Clustering:                          0.644 sec
Reporting:                           0.008 sec
*****
root@licantropo: /home/jabenitez/master/dit/cluto-2.1.1/Linux#

```

En esta aplicación, se observa que el programa genera los clústers con una pureza más alta y con un menor número de objetos insertados en los primeros clústeres, pero en los últimos clústers almacena un mayor número de objetos y además de distinto tipo, haciendo que la pureza general sea peor que las anteriores.

Esto es producido por el cambio en la función de similitud.

25 de mayo de 2011

VCluster aplicado al fichero re1.mat

Con un algoritmo aglomerativo, función de similitud cosine y función de criterio i1

Línea de comando:

```
./vcluster -clmethod='agglo' -sim='cos' -crfun='i1' re1.mat 10
```

Salida obtenida

```

root@licantropo: /home/jabenitez/master/dit/cluto-2.1.1/Linux
*****
root@licantropo:/home/jabenitez/master/dit/cluto-2.1.1/Linux# ./vcluster -clmethod='agglo' -sim='cos' -crfun='i1' re1.mat 10
*****
vcluster (CLUTO 2.1.1) Copyright 2001-03, Regents of the University of Minnesota

Matrix Information -----
  Name: re1.mat, #Rows: 1657, #Columns: 3758, #NonZeros: 87328

Options -----
  CLMethod=AGGLO, CRfun=I1, SimFun=Cosine, #Clusters: 10
  RowModel=None, ColModel=IDF, GrModel=SY-DIR, NNbrs=40
  Colprune=1.00, EdgePrune=-1.00, VtxPrune=-1.00, MinComponent=5
  CStype=Best, AggloFrom=0, AggloCRfun=I1, NTrials=10, Niter=10

Solution -----

-----
10-way clustering: [I1=1.52e+02] [1657 of 1657]
-----
cid  Size  ISim  ISdev  ESim  ESdev  |
-----
  0    75 +0.223 +0.059 +0.024 +0.008 |
  1   105 +0.174 +0.043 +0.025 +0.009 |
  2    70 +0.213 +0.055 +0.016 +0.006 |
  3    44 +0.271 +0.048 +0.020 +0.009 |
  4    29 +0.487 +0.103 +0.021 +0.008 |
  5   325 +0.061 +0.018 +0.023 +0.008 |
  6    72 +0.203 +0.043 +0.028 +0.008 |
  7    32 +0.325 +0.047 +0.022 +0.005 |
  8    60 +0.166 +0.046 +0.017 +0.005 |
  9   845 +0.025 +0.008 +0.022 +0.010 |
-----

Timing Information -----
  I/O:                                0.076 sec
  Clustering:                          1.424 sec
  Reporting:                           0.008 sec
*****
root@licantropo:/home/jabenitez/master/dit/cluto-2.1.1/Linux#

```

25 de mayo de 2011

Aplicando la siguiente línea de comando:

```
./vcluster -clmethod='agglo' -sim='cos' -crfun='l1' -rclassfile=estadisticas/re1.mat.rclass re1.mat 10
```

```

root@licantropo: /home/jabenitez/master/dit/cluto-2.1.1/Linux
Options -----
CLMethod=AGGLO, CRfun=l1, SimFun=Cosine, #Clusters: 10
RowModel=None, ColModel=IDF, GrModel=SV-DIR, NNbrs=40
Colprune=1.00, EdgePrune=1.00, VtxPrune=1.00, MinComponent=5
CSType=Best, AggloFrom=0, AggloCRfun=l1, NTrials=10, NIter=10
Solution -----
10-way clustering: [Ii=1.52e+02] [1457 of 1657], Entropy: 0.548, Purity: 0.508
-----
cid  Size  ISim  ISdev  ESim  ESdev  Entpy  Purity | coco  grai  veg  whea  copp  coff  suga  ship  cott  carc  crud  nat  meal  alum  oils  gold  tin  live  iron  rubb  zinc  oran  pet  dir  gas
0  75 +0.223 +0.059 +0.024 +0.008 0.874 0.947 | 3  0  0  0  0  71  0  0  0  0  0  0  0  0  0  1  0  0  0  0  0  0  0  0  0  0  0  0
1  105 +0.174 +0.043 +0.025 +0.009 0.158 0.819 | 0  18  0  0  0  0  0  86  0  0  0  1  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
2  70  +0.213 +0.055 +0.016 +0.006 0.000 1.000 | 0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  70  0  0  0  0  0  0  0  0  0  0  0
3  44  +0.271 +0.048 +0.020 +0.009 0.111 0.909 | 0  0  0  0  0  0  0  0  0  0  0  10  1  0  0  0  0  0  0  0  0  0  0  0  0  0  0  3
4  29  +0.487 +0.103 +0.021 +0.008 0.000 1.000 | 29  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
5  325 +0.061 +0.018 +0.023 +0.008 0.419 0.643 | 4  209  44  14  0  1  3  7  0  4  4  0  2  0  22  0  0  7  1  0  0  0  0  0  0  1
6  72  +0.203 +0.043 +0.028 +0.008 0.023 0.984 | 0  0  0  0  0  0  0  0  0  0  0  71  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  1
7  32  +0.325 +0.047 +0.022 +0.005 0.416 0.562 | 0  18  3  1  0  0  0  0  0  0  0  3  0  2  0  0  5  0  0  0  0  0  0  0  0  0  0  0
8  60  +0.146 +0.046 +0.017 +0.005 0.376 0.633 | 0  0  0  1  0  0  0  0  38  0  0  5  0  0  0  6  5  5  0  0  0  0  0  0  0  0  0  0
9  843 +0.025 +0.008 +0.022 +0.010 0.838 0.247 | 12  126  12  4  87  27  15  92  17  16  209  26  10  31  19  12  13  30  30  32  10  13  18  20  14
-----
Timing Information -----
I/O: 0.040 sec
Clustering: 1.420 sec
Reporting: 0.000 sec
-----
root@licantropo: /home/jabenitez/master/dit/cluto-2.1.1/Linux#

```

Tiene una pureza general de 0.508, es un algoritmo que consigue unas purezas medias un poco aleatorias, sin seguir un patrón fácil de explicar.

25 de mayo de 2011

Con un algoritmo aglomerativo, función de similitud cosine y función de criterio i2

Línea de comando:

`./vcluster -clmethod='agglo' -sim='cos' -crfun='i2' re1.mat 10`

Salida obtenida

```

root@licantropo: /home/jabenitez/master/dit/cluto-2.1.1/Linux
*****
root@licantropo:/home/jabenitez/master/dit/cluto-2.1.1/Linux# ./vcluster -clmethod='agglo' -sim='cos' -crfun='i2' re1.mat 10
*****
vcluster (CLUTO 2.1.1) Copyright 2001-03, Regents of the University of Minnesota

Matrix Information -----
  Name: re1.mat, #Rows: 1657, #Columns: 3758, #NonZeros: 87328

Options -----
  CLMethod=AGGLO, CRfun=I2, SimFun=Cosine, #Clusters: 10
  RowModel=None, ColModel=IDF, GrModel=SY-DIR, NNbrs=40
  Colprune=1.00, EdgePrune=-1.00, VtxPrune=-1.00, MinComponent=5
  CStype=Best, AggloFrom=0, AggloCRfun=I2, NTrials=10, NIter=10

Solution -----

-----
10-way clustering: [I2=4.64e+02] [1657 of 1657]
-----
cid  Size  ISim  ISdev  ESim  ESdev  |
-----
  0   204 +0.049 +0.018 +0.016 +0.006 |
  1   276 +0.039 +0.013 +0.022 +0.009 |
  2   155 +0.074 +0.038 +0.019 +0.008 |
  3   392 +0.055 +0.020 +0.021 +0.008 |
  4    64 +0.234 +0.054 +0.017 +0.007 |
  5   120 +0.152 +0.045 +0.025 +0.008 |
  6    82 +0.203 +0.060 +0.023 +0.008 |
  7   180 +0.070 +0.025 +0.022 +0.008 |
  8    78 +0.151 +0.060 +0.019 +0.008 |
  9   106 +0.129 +0.048 +0.025 +0.008 |
-----

Timing Information -----
  I/O:                                0.072 sec
  Clustering:                          1.476 sec
  Reporting:                           0.012 sec
*****
root@licantropo:/home/jabenitez/master/dit/cluto-2.1.1/Linux#

```

25 de mayo de 2011

Aplicando la siguiente línea de comando:

```
./vcluster -clmethod='agglo' -sim='cos' -crfun='i2' -rclassfile=estadisticas/re1.mat.rclass re1.mat 10
```

```

root@licantropo:/home/jabenitez/master/dit/cluto-2.1.1/Linux
Options -----
CLMethod=AGGLO, CRfun=I2, SimFun=Cosine, #Clusters: 10
RowModel=None, ColModel=IDF, GrModel=SV-DIR, NNdrs=40
ColPrune=1.00, EdgePrune=1.00, VtxPrune=1.00, MinComponent=5
CSType=Best, AggloFrom=0, AggloCRfun=I2, NTrials=10, NIter=10
Solution -----
10-way clustering: [I2=4.64e+02] [1457 of 1657], Entropy: 0.470, Purity: 0.552
cid Size ISim ISdev ESim ESdev Entpy Purity | coco grai veg whea copp coff suga ship cott carc crud nat meal alum oile gold tin live iron rubb zinc oran pet dir gas
-----
0 204 +0.049 +0.018 +0.016 +0.006 0.482 0.490 | 0 27 2 0 3 0 2 100 0 7 47 0 2 0 6 0 0 2 0 0 1 0 3 0 2
1 276 +0.039 +0.013 +0.022 +0.009 0.491 0.105 | 3 23 29 0 2 9 7 18 1 8 18 0 5 23 16 10 0 28 27 3 7 13 8 18 0
2 155 +0.074 +0.038 +0.019 +0.008 0.697 0.232 | 36 10 4 0 27 7 5 4 2 1 5 1 0 0 1 7 17 0 0 27 1 0 0 0 0
3 392 +0.055 +0.020 +0.021 +0.008 0.457 0.673 | 5 264 9 19 4 6 3 12 14 3 4 0 8 0 21 0 0 11 2 2 1 0 0 2 2
4 64 +0.234 +0.054 +0.017 +0.007 0.000 1.000 | 0 0 0 0 0 0 0 0 0 0 0 0 0 0 64 0 0 0 0 0 0 0 0 0
5 120 +0.132 +0.043 +0.023 +0.008 0.200 0.742 | 0 29 0 0 0 0 0 89 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0
6 82 +0.203 +0.060 +0.023 +0.008 0.081 0.939 | 4 1 0 0 0 0 77 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
7 180 +0.070 +0.025 +0.022 +0.008 0.459 0.600 | 0 7 16 0 1 0 0 2 0 1 108 25 0 1 5 6 1 1 1 0 0 0 2 0 3
8 78 +0.151 +0.060 +0.019 +0.008 0.279 0.705 | 0 10 0 0 0 0 0 1 0 0 0 55 1 0 0 0 0 0 0 0 0 0 0 11
9 108 +0.125 +0.048 +0.023 +0.008 0.166 0.868 | 0 0 0 0 0 0 0 0 0 0 0 82 0 0 7 0 0 0 0 1 0 0 5 0 1
Timing Information -----
I/O: 0.080 sec
Clustering: 1.488 sec
Reporting: 0.000 sec
=====
root@licantropo:/home/jabenitez/master/dit/cluto-2.1.1/Linux#

```

Consigue una mejor pureza general, pero no se nota mucha diferencia al haber cambiado únicamente la función de criterio de i1 a i2.

25 de mayo de 2011

Con un algoritmo de separación, función de similitud de distancia euclidea y función de criterio i1

Línea de comando:

`./vcluster -clmethod='graph' -sim='dist' -crfun='i1' re0.mat 10`

Salida obtenida

```

root@licantropo: /home/jabenitez/master/dit/cluto-2.1.1/Linux
ph' -sim='dist' -crfun='i1' re1.mat 10
*****
vcluster (CLUTO 2.1.1) Copyright 2001-03, Regents of the University of Minnesota

Matrix Information -----
  Name: re1.mat, #Rows: 1657, #Columns: 3758, #NonZeros: 87328

Options -----
  CLMethod=GRAPH, CRfun=Cut, SimFun=EuclDist, #Clusters: 10
  RowModel=None, ColModel=IDF, GrModel=SY-DIR, NNbrs=40
  Colprune=1.00, EdgePrune=-1.00, VtxPrune=-1.00, MinComponent=5
  CStype=Best, AggloFrom=0, AggloCRFun=SLINK_W, NTrials=10, NIter=10

Solution -----

33-way clustering: [Cut=1.33e+03] [280 of 1657]

-----
cid  Size  ISim  ISdev  ESim  ESdev  |
-----
  0     8 +0.922 +0.107 +0.000 +0.000 |
  1     5 +0.999 +0.000 +0.000 +0.000 |
  2     4 +0.988 +0.002 +0.001 +0.002 |
  3     5 +0.874 +0.131 +0.000 +0.000 |
  4     7 +0.789 +0.173 +0.000 +0.000 |
  5     7 +0.752 +0.187 +0.000 +0.000 |
  6     4 +0.808 +0.186 +0.001 +0.002 |
  7    45 +0.622 +0.273 +0.018 +0.016 |
  8    10 +0.643 +0.168 +0.003 +0.004 |
  9     5 +0.719 +0.014 +0.000 +0.000 |
 10     5 +0.696 +0.273 +0.000 +0.000 |
 11     8 +0.633 +0.185 +0.000 +0.000 |
 12     5 +0.661 +0.259 +0.000 +0.000 |
 13     5 +0.654 +0.252 +0.000 +0.000 |
 14    23 +0.555 +0.224 +0.014 +0.017 |
 15     4 +0.653 +0.267 +0.002 +0.003 |
 16     4 +0.648 +0.265 +0.006 +0.006 |
 17     6 +0.573 +0.208 +0.000 +0.000 |
 18     5 +0.590 +0.221 +0.000 +0.000 |
 19     5 +0.585 +0.279 +0.001 +0.002 |
 20     8 +0.532 +0.196 +0.000 +0.000 |
 21    10 +0.532 +0.271 +0.033 +0.018 |
 22    11 +0.486 +0.230 +0.001 +0.002 |
 23     8 +0.487 +0.235 +0.000 +0.000 |
 24     5 +0.521 +0.196 +0.000 +0.000 |
 25    11 +0.430 +0.180 +0.000 +0.000 |
 26     4 +0.518 +0.137 +0.001 +0.002 |
 27     6 +0.461 +0.203 +0.000 +0.000 |
 28     5 +0.383 +0.214 +0.000 +0.000 |
 29     9 +0.324 +0.161 +0.001 +0.002 |
 30     8 +0.315 +0.194 +0.000 +0.000 |
 31    16 +0.249 +0.156 +0.004 +0.006 |
 32     9 +0.202 +0.084 +0.000 +0.000 |
-----

Timing Information -----
  I/O:                                0.068 sec
  Clustering:                          0.244 sec
  Reporting:                           0.004 sec
*****
root@licantropo:/home/jabenitez/master/dit/cluto-2.1.1/Linux#

```

25 de mayo de 2011

Comando:

```
./vcluster -clmethod='graph' -sim='dist' -crfun='i1' -rclassfile=estadisticas/re1.mat.rclass
re1.mat 10
```

Salida obtenida:

root@icantop:/home/yabentz/master/dlt/cuto-211/Unix

33-way clustering: [Cut=1.33e+03] [280 of 1657]. Entropy: 0.167, Purity: 0.796

cid	Size	LSim	LSdev	ESim	ESdev	Entropy	Purity	coco	grai	veg	whea	copp	coff	suga	ship	cott	carc	crud	nat	meal	alum	oils	gold	tin	live	iron	rubb	zinc	cran	pet	dir	gas
0	8	+0.922	+0.107	+0.000	+0.000	0.000	1.000	0	0	0	0	8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
1	5	+0.999	+0.000	+0.000	+0.000	0.000	1.000	0	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
2	4	+0.988	+0.002	+0.001	+0.002	0.000	1.000	0	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
3	5	+0.874	+0.131	+0.000	+0.000	0.155	0.800	0	0	0	0	0	0	0	0	0	0	4	0	0	0	0	0	0	0	0	0	0	0	0	1	
4	7	+0.789	+0.173	+0.000	+0.000	0.000	1.000	0	0	0	0	0	0	0	7	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
5	7	+0.752	+0.187	+0.000	+0.000	0.127	0.857	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	6	0	0	0	0	0	0	0	0	
6	4	+0.808	+0.186	+0.001	+0.002	0.000	1.000	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
7	45	+0.622	+0.273	+0.018	+0.016	0.484	0.467	5	21	1	0	0	1	2	0	0	0	9	0	0	0	0	0	0	0	0	0	0	0	0	5	
8	10	+0.643	+0.168	+0.003	+0.004	0.398	0.400	0	4	3	0	0	0	0	0	0	0	0	0	0	0	2	0	1	0	0	0	0	0	0	0	
9	5	+0.719	+0.014	+0.000	+0.000	0.000	1.000	0	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
10	5	+0.696	+0.273	+0.000	+0.000	0.000	1.000	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5	0	0	0	0	0	0	0	0	0	
11	0	+0.633	+0.185	+0.000	+0.000	0.000	1.000	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
12	5	+0.661	+0.259	+0.000	+0.000	0.155	0.800	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	
13	5	+0.684	+0.252	+0.000	+0.000	0.000	1.000	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
14	23	+0.555	+0.224	+0.014	+0.017	0.111	0.913	0	0	0	0	0	0	0	0	0	0	0	21	1	0	0	0	0	0	0	0	0	0	0	1	
15	4	+0.653	+0.267	+0.002	+0.003	0.000	1.000	0	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
16	4	+0.648	+0.265	+0.006	+0.006	0.175	0.750	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
17	6	+0.578	+0.208	+0.000	+0.000	0.000	1.000	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
18	5	+0.590	+0.221	+0.000	+0.000	0.209	0.600	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
19	5	+0.585	+0.279	+0.001	+0.002	0.000	1.000	0	0	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
20	8	+0.532	+0.196	+0.000	+0.000	0.464	0.375	0	1	2	0	1	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
21	10	+0.532	+0.271	+0.033	+0.018	0.232	0.700	0	7	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
22	11	+0.486	+0.230	+0.001	+0.002	0.000	1.000	11	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
23	8	+0.487	+0.233	+0.000	+0.000	0.173	0.750	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
24	5	+0.521	+0.196	+0.000	+0.000	0.000	1.000	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
25	11	+0.430	+0.180	+0.000	+0.000	0.000	1.000	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
26	4	+0.518	+0.197	+0.001	+0.002	0.000	1.000	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
27	4	+0.461	+0.203	+0.000	+0.000	0.140	0.833	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
28	5	+0.383	+0.214	+0.000	+0.000	0.000	1.000	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
29	9	+0.324	+0.161	+0.001	+0.002	0.291	0.556	0	5	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
30	8	+0.315	+0.194	+0.000	+0.000	0.000	1.000	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
31	16	+0.249	+0.156	+0.004	+0.006	0.175	0.750	0	4	0	0	0	0	12	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
32	9	+0.202	+0.084	+0.000	+0.000	0.000	1.000	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	

En esta aplicación, se observa que el programa genera un mayor número de clústers, con un menor número de objetos almacenados en cada uno, pero con una pureza bastante alta, la media es de 0.796.

Esto indica que siguiendo un algoritmo de separación y una función de selección mediante distancia euclídea, se obtiene un mejor empaquetado mediante clustering.

25 de mayo de 2011

Con un algoritmo de separación, función de similitud del coeficiente de Jaccard y función de criterio i1

Línea de comando:

```
./vcluster -clmethod='graph' -sim='jacc' -crfun='i1' re0.mat 10
```

Salida obtenida

```

root@licantropo: /home/jabenitez/master/dit/cluto-2.1.1/Linux
*****
root@licantropo:/home/jabenitez/master/dit/cluto-2.1.1/Linux# ./vcluster -clmethod='gra
ph' -sim='jacc' -crfun='i1' re1.mat 10
*****
vcluster (CLUTO 2.1.1) Copyright 2001-03, Regents of the University of Minnesota

Matrix Information -----
  Name: re1.mat, #Rows: 1657, #Columns: 3758, #NonZeros: 87328

Options -----
  CLMethod=GRAPH, CRfun=Cut, SimFun=ExtJaccard, #Clusters: 10
  RowModel=None, ColModel=IDF, GrModel=SY-DIR, NNbrs=40
  Colprune=1.00, EdgePrune=-1.00, VtxPrune=-1.00, MinComponent=5
  CStype=Best, AggloFrom=0, AggloCRFun=SLINK_W, NTrials=10, NIter=10

Solution -----

-----
10-way clustering: [Cut=3.17e+03] [1657 of 1657]
-----

cid  Size  ISim  ISdev  ESim  ESdev  |
-----
  0    90 +0.050 +0.027 +0.000 +0.000 |
  1    91 +0.049 +0.025 +0.000 +0.000 |
  2    90 +0.046 +0.034 +0.000 +0.000 |
  3   134 +0.030 +0.018 +0.000 +0.000 |
  4   148 +0.022 +0.018 +0.000 +0.000 |
  5   123 +0.020 +0.013 +0.000 +0.000 |
  6   168 +0.015 +0.010 +0.000 +0.000 |
  7   171 +0.015 +0.012 +0.001 +0.000 |
  8   244 +0.010 +0.006 +0.000 +0.000 |
  9   398 +0.009 +0.005 +0.000 +0.000 |
-----

Timing Information -----
  I/O:                                0.072 sec
  Clustering:                          0.716 sec
  Reporting:                           0.008 sec
*****
root@licantropo:/home/jabenitez/master/dit/cluto-2.1.1/Linux# █

```

25 de mayo de 2011

Comando:

```
./vcluster -clmethod='graph' -sim='jacc' -crfun='i1' -rclassfile=estadisticas/re1.mat.rclass
re1.mat 10
```

Salida obtenida:

```

root@icantropo:/home/jabenitez/master/dit/cluto-2.1.1/linux#
Options
-----
CLMethod=GRAPH, CRfun=Cut, SimFun=ExtJaccard, #Clusters: 10
RowModel=None, ColModel=IDF, GrModel=SV-DR, Nbrs=40
ColPrune=1.00, EdgePrune=1.00, VtxPrune=1.00, MinComponent=5
SType=Heat, AggloProc=0, AggloCRFun=SIMM_W, Mtrial=10, NIter=10
Solution
-----
10-way clustering: [Cut=3.17e+03] [1657 of 1657], Entropy: 0.454, Purity: 0.552
-----
cid  Size  ISim  ISdev  ESim  ESdev  Entpy  Purity  coco  grai  veg  ques  copp  coff  suga  ship  cott  carc  crud  nat  meal  alum  oils  gold  tin  live  iron  rubb  zinc  oran  pet  dir  gas
0  50 +0.050 +0.027 +0.000 +0.000 0.052 0.967 | 2 0 0 0 0 0 87 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
1  91 +0.049 +0.025 +0.000 +0.000 0.039 0.970 | 0 3 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
2  90 +0.046 +0.034 +0.000 +0.000 0.001 0.467 | 42 3 0 0 0 0 0 4 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
3  134 +0.030 +0.018 +0.000 +0.000 0.150 0.481 | 0 1 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
4  148 +0.022 +0.018 +0.000 +0.000 0.456 0.541 | 0 0 0 0 0 28 0 0 0 0 0 2 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
5  123 +0.020 +0.013 +0.000 +0.000 0.214 0.772 | 0 3 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
6  166 +0.015 +0.010 +0.000 +0.000 0.527 0.542 | 0 20 1 0 0 3 1 4 91 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
7  173 +0.015 +0.012 +0.000 +0.000 0.212 0.187 | 1 32 4 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
8  244 +0.010 +0.006 +0.000 +0.000 0.562 0.516 | 0 126 24 3 0 1 4 3 6 6 3 10 3 0 0 0 0 4 1 15 1 0 0 26 0 0 0 12 0 0 0 2
9  398 +0.009 +0.005 +0.000 +0.000 0.529 0.460 | 3 183 31 13 1 1 97 7 13 0 0 0 0 0 0 0 0 0 0 27 1 0 0 0 0 0 0 1 0 0 0
Timing Information
-----
I/O: 0.076 sec
Clustering: 0.724 sec
Reporting: 0.004 sec
-----
root@icantropo:/home/jabenitez/master/dit/cluto-2.1.1/linux#

```

En esta aplicación, se observa que el programa genera los clústers con una pureza más alta y con un menor número de objetos insertados en los primeros clústeres, pero en los últimos clústers almacena un mayor número de objetos y además de distinto tipo, haciendo que la pureza general sea peor que las anteriores.

Esto es producido por el cambio en la función de similitud.

3. Conclusiones generales sobre los distintos tipos de algoritmos a seguir.

Después de ver a las distintas salidas habiendo aplicado los distintos algoritmos, funciones de criterio y de selección, puedo observar lo siguiente:

- En la aplicación del algoritmo aglomerativo:

- No existe una gran diferencia de empaquetado por clustering utilizando la función de criterio i1 o i2.

- En la aplicación del algoritmo de partición:

- Cuando se aplica la función de similitud por distancia euclídea, se crea un mayor número de clusters y se consigue tener un mejor resultado.
- Al aplicar la función de similitud por el coeficiente de Jaccard, se puede observar perfectamente cómo los primeros clusters tienen una mayor pureza y un menor número de objetos en su interior, pero casi todos del mismo tipo. Sin embargo, los clusters creados al final, tienen un mayor número de objetos dentro y un menor coeficiente de pureza.

- En general: el que mejor resultado ha dado sin duda es el algoritmo de partición con la aplicación de la función de similitud por distancia euclídea.