Màster Interuniversitari en Seguretat en Tecnologies de la Informació y Comunicació

Avaluació de CluStream com a algorisme de microagregació per streams de dades

José Alberto Aguiló Escámez

Director: Guillermo Navarro-Arribas

Universitat Oberta de Catalunya
Universitat Autònoma de Barcelona
Universitat Rovira i Virgili

Palma de Mallorca, 12 de gener de 2014

ÍNDEX

ÍNDEX DE FIGURES

ÍNDEX DE TAULES

ACRÒNIMS

API	Application Programming Interface
ARFF	Attribute-Relation File Format
IDE	Integrated Development Environment
JDK	Java Development Kit
JRE	Java Runtime Environment
JVM	Java Virtual Machine
MOA	Massive Online Analysis
WEKA	Waikato Environment for Knowledge Analysis

RESUM

L'objectiu del treball final de màster s'enfoca en la realització d'un anàlisi sobre un conjunt de dades stream extretes de la xarxa social Twitter amb l'objectiu de garantir la privadesa de les dades i estudiar la pèrdua d'informació que suposa l'obtenció d'aquesta privadesa i com afecta a la qualitat. Per a tal d'acomplir aquesta tasca es realitza la implementació d'un algorisme d'agrupacions.

L'objectiu de l'algorisme en agrupacions de registres, anomenats clústers, que compleixen certa similitud en un nombre, facilitant un valor central conegut com centroide que marca el valor mitjà de l'agrupació específica.

Per a poder arribar a facilitar unes dades vàlides s'ha de filtrar la informació inicial no protegida per tal d'eliminar tota aquella informació que no aporta valor a l'estudi o bé que vulnera directament la privadesa per tal de poder aplicar l'algorisme amb fiabilitat. L'aplicació de l'algorisme amb diferents configuracions facilitarà dades amb diversos graus de privadesa. Posteriorment s'avaluarà amb operacions algorítmiques quina és la quantitat d'informació perduda per a cada un dels resultats.

Una vegada es disposa de les duples de grau de privadesa i del indicatiu de perduda d'informació, es presenta un estudi conclusiu sobre la necessitat de privadesa en front de la pèrdua de qualitat de les dades.

1. INTRODUCCIÓ

Les dades stream es caracteritzen per esser de gran volum i estar subjectes a continus canvis de forma molt dinàmica. Aquestes dades contenen en molts casos informació privada, que donat la particularitat del stream de dades, es comú que siguin monitoritzades.

Quan parlem de qualsevol conjunt de dades en general, aquest no deixa d'esser una agrupació d'un nombre de registres on cada registre té una sèrie d'atributs. Si tractem específicament amb grups de dades que són sensibles a facilitar informació que identifiqui qualque registre hem de poder avaluar els atributs en quatre gran grups segons expressa el estudi realitzat per Domingo-Ferrer i Torra [1].

Un atribut s'anomena identificador quan simplement coneixent el seu valor es pot identificar el registre al que pertany. Arguments com el D.N.I. o número de la seguretat social són clarament identificadors per sí.

També existeix l'argument quasi-identificador i es caracteritzen per poder arribar a facilitar la identificació del registre amb combinació d'altres arguments o d'informació externa al registre.

En un altre punt de vista els atributs també es poden diferenciar entre confidencials i no confidencials. Els confidencials aporten informació sensible del registre com ètnia, afiliació sindical, estat de salut o salari entre d'altres. Per contrapartida, les dades no confidencials són aquelles que en principi no faciliten informació sensible com podria esser per exemple el lloc de naixement, la ciutat de residència o l'edat.

L'existència del risc de pèrdua total o parcial de privacitat amb estudis sobre dades stream es deu en molts de casos a l'existència d'interessos econòmics, polítics o simplement d'investigació com és el cas que ens presenta al treball. Tenint en compte les caracteritzacions dels arguments esmentades i la necessitat de garantir la privacitat de les dades stream, s'han d'aplicar una sèrie de mesures per a tal de poder facilitar un estudi de les dades sense corrompre la identitat o privadesa.

Com a objectiu hem s'ha d'assolir una privadesa mínima, llavors donada aquesta premissa és lògic pensar que el primer a fer en el tractament de les dades no protegides és la identificació i supressió de tots aquells arguments que siguin identificadors així com els arguments confidencials, que en ocasions podria donar-se el cas de que un mateix argument compleixi ambdues característiques. El producte resultant és un grup de dades de quasi-identificadors no confidencials, però aquesta primera filtració no garanteix que no es pugui identificar determinats registres, ja que una combinació d'arguments podria identificar i inclús podria facilitar informació confidencial.

La solució oferta en aquest treball per a la cerca de la privadesa de les dades és emprar el mètode de micro-agregació sobre els quasi-identificadors no confidencials. Aquest mètode realitza agrupacions de dades per a tal d'aconseguir un grau

d'anonimat en les dades resultants. Si bé existeixen diferents algorismes de micro-agregació, s'ha elegit emprar *CluStream* [5]. Això s'aconsegueix agrupant un nombre finit de registres de certa similitud als quals són modificats amb un valor intermedi del grup. Cada agrupació és coneguda com a clúster, essent el valor mitjà elegit el centroide del clúster. Aquesta operació dona com a resultat una sèrie de registres amb idèntic valor, lo que facilita cert grau d'anonimat en funció del nombre de registres amb el mateix valor.

Un altre aspecte a tenir en compte és el fet de la relació existent entre el grau de privadesa assolit i la quantitat de pèrdua d'informació, per tant a major grau d'anonimat major serà la informació perduda. Per tal de poder quantificar aquesta pèrdua s'ha emprat la desviació estàndard de cada una de les agrupacions, ja que si bé no exactament el grau d'informació perduda sí que serveix com a clar indicatiu.

Totes aquestes valoracions s'han portat a terme per a la realització d'aquest treball, on s'han elegit dues fonts de dades stream de la xarxa social *Twitter*, recopilant informació d'usuaris per a posteriorment realitzar un anàlisi amb mètode algorítmic d'agrupació existents al programari *MOA* (*Massive Online Analysis*).

1.1. Organització del document

A continuació es detalla una breu introducció dels aspectes que es tractaran al document i la seva organització.

A la secció 2 es detallen les eines emprades per a la realització del treball.

La secció 3 servirà per a presentar els diferents algorismes d'agrupació que s'han emprat i les seves bases en quant a recerca d'anonimat.

La implementació es detalla a la secció 4, definint com s'ha realitzat el filtratge i sobre quins arguments. A més es detallarà com s'han configurat els diferents algorismes d'agrupació i el procediment de càlcul de la informació perduda. Posteriorment, a la secció 5 es presentarà un anàlisi amb els resultats obtinguts.

La conclusió del document es troba a la secció 6, on a més es detallen una sèrie de millores per a futures continuacions d'aquest estudi.

2. EINES UTILITZADES

En aquesta secció es mostraran les diferents eines emprades en el projecte. A l'apartat 2.1 es tracta el programari d'anàlisi de dades stream, a l'apartat 2.2 l'eina *IDE*. El llenguatge de programació que s'ha emprat es troba a l'apartat 2.3.

2.1. Anàlisi de dades stream

Per a l'execució del treball s'ha decidit emprar el programari *MOA*. Aquest software està orientat a l'extracció de dades stream i per això conté eines per a l'avaluació de les dades així com a conjunts d'algorismes. *MOA* està basat en *WEKA* (*Waikato Environment for Knowledge Analysis*) [2], un projecte encaminat a l'estudi i aplicació de mètodes d'aprenentatge màquina on gràcies a l'anàlisi automàtic de gran volum de dades s'aconsegueix decidir quin tipus d'informació té major rellevància. Al igual que *WEKA*, *MOA* està desenvolupat amb *Java*.

MOA funciona amb una seqüència de passes ben definides. En primera instància s'ha de facilitar una font de dades stream, ja sigui mitjançant una injecció de dades o bé amb un generador de dades, i una vegada es disposa de les dades aquestes seran configurades. Com a segon pas, s'elegeix un algorisme per al tractament de les dades i es configuraran els paràmetres del mètode elegit. Finalment, es determina un mètode avaluador o de mesura de dades.

2.2. Eina IDE

Les *IDE* (*Integrated Development Environment*) són eines de programació on el programador pot editar el codi en diferents llenguatges de programació i a més pot depurar, compilar i construir interfícies gràfiques. Les dues eines *IDE* més emprades i conegudes són *Eclipse* i *NetBeans* i serà *Eclipse* l'elegida per realitzar el treball.

Eclipse és un entorn de desenvolupament integrat de codi lliure multiplataforma desenvolupat en 2001 per *IBM* (*International Business Machines*) inicialment i actualment per *Eclipse Foundation*. *Eclipse* aporta el *Java IDE JDK* (*Java Development Toolkit*) i el compilador *ECJ* (*Eclipse Compiler for Java*).

Com a eina *IDE*, Eclipse possibilita la perfecta interacció amb *MOA* ja que comparteixen llenguatge de programació i a més és de fàcil configuració gràcies a la l'extensió d'*Eclipse* per a control de versions *Tortoise Hg* [3].

2.3. Llenguatge de programació

El llenguatge emprat per a la realització del treball és *Java* donat l'eina d'anàlisi de dades utilitzada.

Java [14] és un llenguatge de programació d'alt nivell orientat a objectes creat per *Sun Microsystems* a 1995. Si bé *Java* es coneix com a llenguatge de programació, en realitat es tracta d'un conjunt de tres elements on un d'ells sí que és el llenguatge de programació en si. Les altres dues parts són una màquina virtual (*JVM*) i la plataforma Java.

El llenguatge de programació *Java* té una clara influencia dels llenguatges C i C++ així com d'altres llenguatges. La seva sintaxi va esser dissenyada per ser familiar amb els

llenguatges que el varen influenciar i amb fermes principis d'orientació a objectes que es trobaven a C++.

Java està generalment pensat per tres tipus de plataformes: *SE* (*Standard Edition*), *EE* (*Enterprise Edition*) i *ME* (*Micro Edition*). Cada una d'elles descriu la combinació entre el llenguatge de programació, les llibreries estàndards i la màquina virtual per executar el codi. *EE* conté *SE* i per tant, qualsevol aplicació *EE* pot assumir que disposa de totes les llibreries *SE*. En el cas de *ME*, aquesta no es cap subgrup de SE*,* ja que disposa de llibreries exclusives que no té *SE* i per tant, tampoc *EE*.

Java precisa convertir el codi natiu en executable i per això necessita dues passes: el programador compila el codi en un codi de bytes de *Java* i després la màquina virtual *Java* (*JVM*) ho converteix de nou al codi natiu per a la plataforma amfitriona.

3. ALGORISMES D'AGRUPACIÓ

Per a la consecució de l'anonimat de les dades s'ha elegit emprar l'agrupació de dades. Seguint amb l'article de Domingo-Ferrer et al. ja introduït, es presenta l'anonimat d'un conjunt de dades amb un factor k com k-anonimat, on k registres formen una agrupació donada la similitud dels seus arguments quasi-identificadors, essent k sempre major que 1. Quan major sigui el valor de k, menor serà la probabilitat d'identificació de cada registre i per tant major serà el grau de privadesa, per contra s'augmenta la quantitat de pèrdua d'informació.

El programari *MOA* ofereix una sèrie de característiques per a la consecució d'aquest k-anonimat i així ve expressat a l'article de Bifet et al. [4]. Entre diverses funcionalitats que ofereix el programari es troben alguns algorismes d'agrupació ja coneguts on concretament ens focalitzarem en un d'ells com és *CluStream* [5].

3.1. CluStream

CluStream sorgeix enfront de qüestions de com facilitar suficient informació espacial i temporal per a processos online i offline, en quin moment s'ha de guardar la informació tenint en compte els requeriments d'agrupació temporal o com i quan els reports poden esser emprats per a agrupacions i evolucions. Per tal de donar solució a totes aquestes qüestions, *CluStream* presenta dos conceptes com són el de micro-clústers i marc de temps piramidal.

Els micro-clústers són extensions temporals dels vectors de configuració del clúster i es desen com *snapshots* en temps seguint un patró piramidal. Aquest patró ofereix un balanç eficient entre els requeriments d'emmagatzematge de la informació i la habilitat de sol·licitar reports a diferents intervals de temps. *CluStream* té la particularitat d'anar tractant les dades stream com un procés que va variant temporalment i no tracta de realitzar clústers sobre totes les dades a l'hora.

Un altre concepte que presenta aquest algorisme és macro-clústers. Aquest mètode guanya interès a l'hora de tractar informació de forma offline, ja que té com a base els micro-clústers calculats i tracta de cercar noves agrupacions de grau major, important per a veure l'evolució del tractament de les dades en funció del nombre de clústers.

Tal i com exposa l'estudi realitzat per Aggarwal et al., *CluStream* aporta major qualitat d'agrupacions que els sistemes d'agrupació tradicionals, que els conceptes de micro-clústers i marc de temps piramidal garanteixen una major precisió mantenint una alta eficiència i que disposa d'una gran escalabilitat en quant a grandària de dades, dimensionalitat i nombre de clústers.

4. IMPLEMENTACIÓ

Per a la realització del treball s'ha realitzat una implementació d'una sèrie de fases per a l'estudi de les dades stream de la xarxa social *Twitter*. Les diferents fases es separen en filtratge, conversió de dades i micro-agregació. Una vegada assolides aquestes tres passes s'obtenen les dades anonimitzades i llestes per al seu estudi.

La captació de les dades va esser realitzat per la investigadora de la UAB Cristina Pérez gràcies a un boot informàtic fent servir la API de *Twitter*.

4.1. Filtratge

L'extracció proporciona una taula d'informació. Aquesta facilita informació sobre els usuaris de la xarxa social. Inicialment la extracció presenta els següents arguments.

Dades d'usuari:

- *id*
- *name*
- *screen_name*
- *location*
- *description*
- *url*
- *protected*
- *followers_count*
- *friends_count*
- *created_at*
- *favourites_count*
- *utc_offset*
- *time_zone*
- *notifications*
- *geo_enabled*
- *verified*
- *following*
- *statuses_count*
- *lang*
- *contributors_enabled*

El primer pas a realitzar és l'elecció d'aquells arguments que aporten informació rellevant per a l'estudi. A més, s'han eliminat aquells que encara rellevants, no estan complets ja que són dades optatives que molt sovint l'usuari els bloqueja. Una vegada aplicat aquest primer filtre disposem de la següent taula amb el seus arguments.

Taula d'usuaris:

- *name*
- *screen_name*
- *followers_count*

- *friends_count*
- *favourites_count*

El segon pas de filtratge és l'eliminació d'aquells arguments que siguin identificador i/o confidencials. Si bé no hi ha cap argument confidencial, sí que trobem uns quants que son directament identificadors de l'usuari. Clarament tenim arguments identificadors com són el *name* i *screen_name*.

Aplicats aquets dos filtratges únicament tenim presents arguments quasi-identificadors i no confidencials als que s'han d'aplicar algorismes d'agrupament per tal de garantir l'anonimat.

Usuaris:

- *followers_count*
- *friends_count*
- *favourites_count*

4.2. Conversió de dades

Per tal de poder emprar el programari *MOA* amb les dades resultants del filtratge, aquestes s'han de convertir a un fitxer *ARFF* (*Attribute-Relation File Format*) [7]. Aquest tipus de fitxers estan composats d'una capçalera on s'identifica el nom del fitxer així com els arguments i les seves propietats seguit de les dades.

El fitxer amb la informació sobre els usuaris s'ha anomenat *moreInfo* i s'ha configurat el fitxer de la següent forma.

```
@RELATION moreInfo

@ATTRIBUTE followers_count NUMERIC
@ATTRIBUTE friends_count NUMERIC
@ATTRIBUTE favourites_count NUMERIC

@DATA
...
```

4.3. Micro-agregació

Una vegada es tenen les dades filtrades i amb format *ARFF*, podem tractar les dades a fi de realitzar agrupacions cercant l'anonimat d'aquestes. Per realitzar aquesta implementació s'ha emprat la llibreria de *MOA* per amb l'algorisme *CluStream*. Com a entrada, l'algorisme només requereix del nombre de clústers, que en el cas que ens presenta aquest valor de nombre d'agrupacions depèn del nombre total de registres i del grau d'anonimat que es desitja tenir amb la relació $c = n / 2k$ on c és el nombre de clústers, n el nombre total de registres i k el grau d'anonimat per a cada agrupació.

Com a resultat l'algorisme un llistat de tots el clústers amb el nombre de registres a cada un dels clústers, el seu centroide, el radi i la desviació estàndard.

5. RESULTATS

5.1. Registre d'usuaris

El fitxer emprat per a l'estudi dels usuaris de la xarxa social facilita un total de 598 registres. S'han cercat resultats per a graus d'anonimat de 2 a 7 registres com a mínim per clúster.

5.1.1. Grau d'anonimat 2

Per aquest cas s'ha realitzat l'estudi sobre un total de 124 clústers. A la figura 1 es pot observar el nombre de clústers en funció del nombre de registres que contenen cada un d'ells així com el nombre de registres per a cada associació de clústers. Clarament predominen les agrupacions de 1 i 2 registres i a més es pot observar un clúster que conté un valor considerable de registres, concretament amb 163 registres obté una quota del *32.7%* sobre el total de registres.

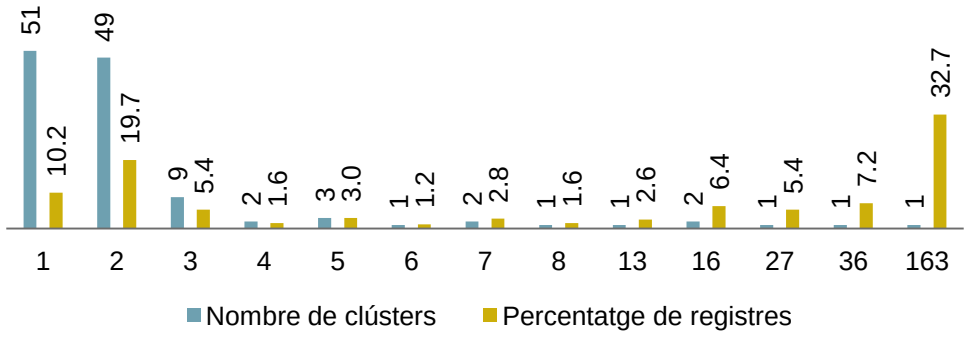

figura 1: Distribució de clústers per a k = 2

Donat que el grau d'anonimat és 2, s'ha de prescindir d'aquelles agrupacions d'un sol registre. La figura 2 mostra el percentatge útil del registres, havent de descartar 51 dels 498 registres.

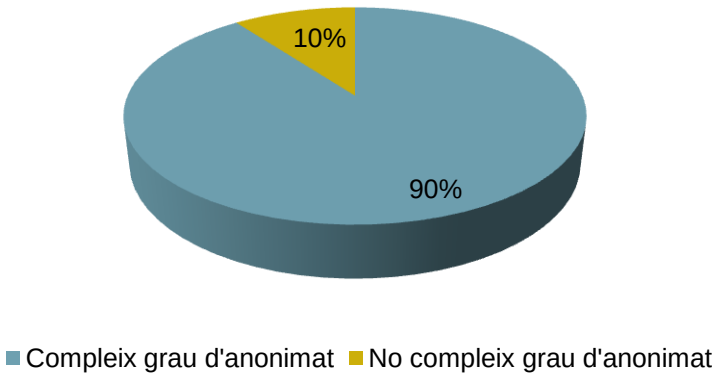

figura 2: Percentatge útil de registres per a k = 2

Un altre resultat a expressar és el grau de pèrdua d'informació que s'obté gràcies a la desviació estàndard de cada clúster. Per aquest estudi s'han descartat aquells clústers

amb menys registres que el mínim d'anonimat requerit. A la taula 1 es detallen les desviacions estàndard per a cada agrupació de registres de tal manera que es pugui comparar el nombre de registres del clúster amb la seva pèrdua d'informació.

Nº Registres	D. estàndard	Nº Registres	D. estàndard	Nº Registres	D. estàndard
2	0.0811	2	2.1713	3	0.0004
2	0.0129	2	56.0141	3	0.0048
2	0.0029	2	0.5039	3	6.2607
2	0.3351	2	0.0035	3	0.0087
2	0.1701	2	5.1687	3	0.0021
2	0.1672	2	0.8334	3	0.0066
2	0.1004	2	0.1671	3	0.0060
2	0.0055	2	0.1711	3	0.0034
2	0.6722	2	0.0007	4	0.0014
2	0.0002	2	0.0027	4	0.0019
2	0.0433	2	0.1683	5	0.0037
2	0.1695	2	0.0072	5	0.0054
2	0.1677	2	0.6704	5	0.0056
2	11.5722	2	0.0007	6	0.0055
2	0.0357	2	0.5042	7	0.0053
2	0.8335	2	0.8369	7	0.0056
2	3.5032	2	0.0003	8	0.0011
2	2.8340	2	0.5025	13	0.0038
2	0.0015	2	0.1069	16	0.0016
2	0.0059	2	0.1668	16	0.0053
2	0.5009	2	0.0163	27	0.0033
2	0.1697	2	0.1667	36	0.0040
2	0.0014	2	0.0015	163	0.0029
2	0.0264	2	0.1702		
2	0.1669	3	0.6204		

taula 1: Desviació estàndard per a k = 2

Es pot observar una sèrie de clústers que surten de la tònica general. Si entrem en detall en aquests clúster podem observar a la taula 2 que es tracten dels registres d'aquells usuaris amb gran nombre de *tweets* favorits, cosa no molt habitual a l'extracció de dades.

Clúster Nº	Nº Registres	Centroide 1er Arg	Centroide 2on Arg	Centroide 3er Arg	Desviació estàndard
31	2	0.0073	0.2275	1607.5	11.5722
37	2	0.0005	0.0194	668.5	3.5032
38	2	0.0014	0.0408	325.5	2.8340
42	3	0.0031	0.0045	1084.0	6.2607
56	2	0.0026	0.0146	415.5	2.1713
57	2	0.0010	0.0546	2240.0	56.0141
69	2	0.0024	0.0178	894.5	5.1687

taula 2: Centroides amb major desviació estàndard per a k = 2

5.1.2. Grau d'anonimat 3

En la recerca del grau d'anonimat 3 s'ha donat un estudi sobre 82 clústers. Seguint amb la tònica de l'anterior grau d'anonimat 2, es mostra la distribució dels registres i clústers.

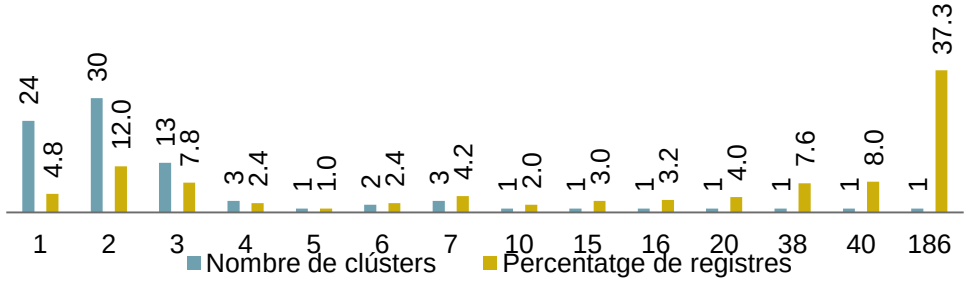

figura 3: Distribució de clústers per a k = 3

Si bé segueixen predominant les agrupacions de 1 i 2 registres, es nota una lleu reducció i per contra un increment de clústers de 3 registres. Aquests registres s'han distribuït en altres agrupacions i concretament en el clúster de major nombre de registres, que augmenta en 23 registres, passant a tenir el *37,3%* de la quota.

Per aquest grau d'anonimat s'observa un increment de registres que han d'esser exclosos en relació a l'anterior apartat de grau 2. Donat el increment de grau d'anonimat és lògic pensar que el nombre de registres que no arribin al llindar serà major.

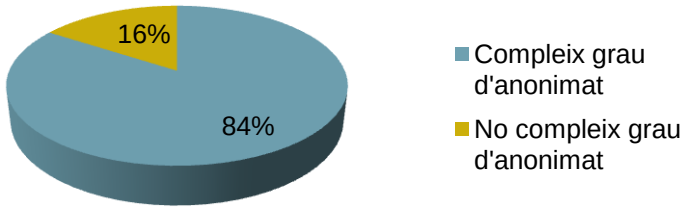

figura 4: Percentatge útil de registres per a k = 3

Si bé fins ara s'ha seguit una única relació $c = n / 2k$ entre el grau d'anonimat a assolir i el nombre de clústers, si per aquest nombre de clústers es volgués reduir el grau d'anonimat a 2, la comparació de registres que compleixen i no compleixen amb el nou grau d'anonimat es mostra a la figura.

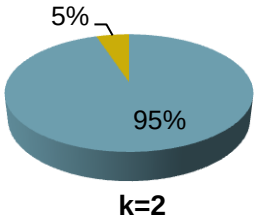

figura 5: Percentatge útil de registres amb variació de k = 3 a k = 2

Es pot observar una millora no només amb el grau d'anonimat 3, si no que també millora considerablement el resultat obtingut per l'anterior apartat amb relació per a grau 2 passant d'un *10%* a un *5%*.

D'igual forma que amb l'anterior apartat i una vegada descartats aquells clústers amb nombre de registres igual o superior al grau d'anonimat, es vol visualitzar la pèrdua d'informació.

Nº Registres	D. estàndard	Nº Registres	D. estàndard	Nº Registres	D. estàndard
3	0.0642	3	0.1573	7	0.0056
3	0.0574	3	0.0161	7	0.1660
3	19.4585	3	6.2607	10	0.0103
3	0.9740	4	0.0159	15	0.0180
3	0.6204	4	0.0113	16	0.0053
3	0.2736	4	0.0019	20	0.0249
3	0.0034	5	0.2291	38	0.0449
3	0.1831	6	0.1726	40	0.0192
3	4.7753	6	0.0132	186	0.0379
3	0.0060	7	0.0217		

taula 3: Desviació estàndard per a k = 3

Observant aquells clústers que obtenen una major desviació estàndard, es torna a veure que són aquells registres amb major nombre de *favorites*.

Clúster Nº	Nº Registres	Centroide 1er Arg	Centroide 2on Arg	Centroide 3er Arg	Desviació estàndard
17	3	0.0003	0.0112	692.66	19.4585
60	3	0.0007	0.0223	259.0	4.7753
75	3	0.0032	0.0045	1084.0	6.2607

taula 4: Centroides amb major desviació estàndard per a k = 3

5.1.3. Grau d'anonimat 4

Per a l'anonimat de grau 4 l'equació ens retorna un total de 62 clústers. Aquests clústers estan distribuïts segons mostra la figura.

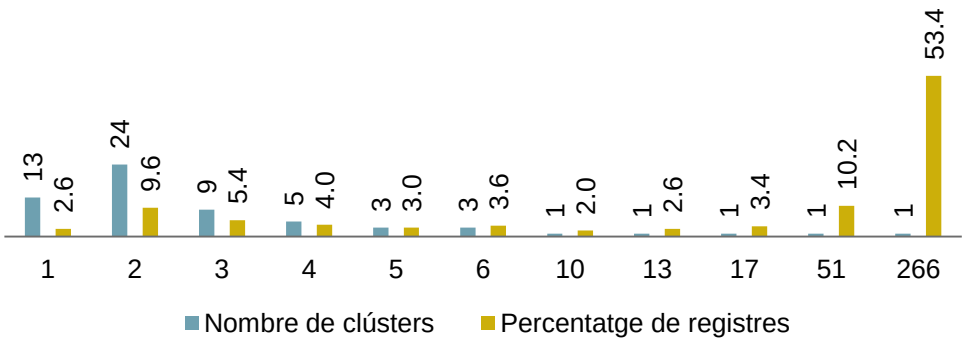

figura 6: Distribució de clústers per a k = 4

Seguint la tendència d'increment de grau es redueix el nombre de clústers per a pocs registres i no es pot deixar de banda el gran increment de registres per al clúster

majoritari que augmenta en 80 registres i es situa amb més de la mitat del total dels registres.

A la figura es pot visualitzar l'evidència de que a major grau de requeriment, major serà el percentatge de rebuig de dades que no arriben a complir amb el mínim.

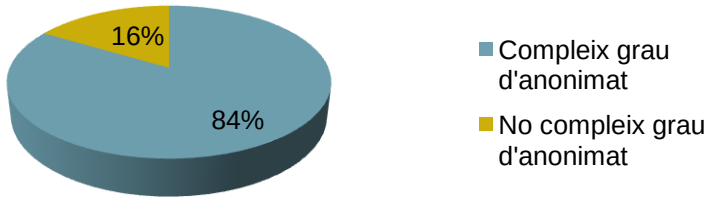

figura 7: Percentatge útil de registres per a k = 4

Si es manté la distribució de registres per clúster i baixem les expectatives d'anonimat obtenim les dades per a grau 3 i 2 tal i com mostren.

Grau d'anonimat	3	2
Nº registres descartats	61	13
Percentatge no vàlid	12.25%	2.61%
Percentatge vàlid	87.75%	97.39%

taula 5: Variació de registres útils amb menor grau de k = 4

D'igual forma que a l'anterior apartat per a grau d'anonimat 3, es veu com els percentatges de rebuig disminueixen.

Nº Registres	D. estàndard	Nº Registres	D. estàndard	Nº Registres	D. estàndard
4	0.2777	5	0.2500	10	0.2998
4	0.1595	5	0.2291	13	0.1867
4	137.4436	6	0.3871	17	0.1729
4	0.2495	6	0.1726	51	0.2806
4	0.2775	6	0.1714	266	0.3023
5	0.1668				

taula 6: Desviació estàndard per a k = 4

Únicament destaquem un registre amb una desviació estàndard excessivament elevada, que equival a un alt nombre de *favorites*. Com a dada a tenir en compte, la desviació estàndard de la resta dels clústers ha augmentat, ja que la mitja passa de *0.1213* a *0.2420*.

Clúster Nº	Nº Registres	Centroide 1er Arg	Centroide 2on Arg	Centroide 3er Arg	Desviació estàndard
42	4	0.0033	0.0563	2532.5	137.4436

taula 7: Centroides amb major desviació estàndard per a k = 4

5.1.4. Grau d'anonimat 5

Per aquest grau d'anonimat s'ha obtingut un total de 49 clústers. A la figura 8 s'observa la tendència a disminuir el nombre de registres als clústers de nombres baixos de registres, però per contrapartida el clúster majoritària disminueix la seva quota.

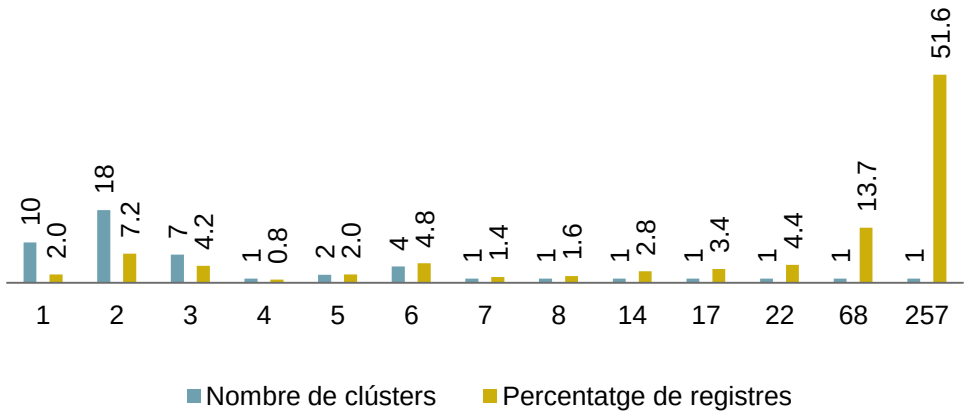

figura 8: Distribució de clústers per a k = 5

En quant a la relació de rebuig de registres, s'observa un canvi de tendència i es disminueix la quota de registres que no compleixen amb el grau d'anonimat.

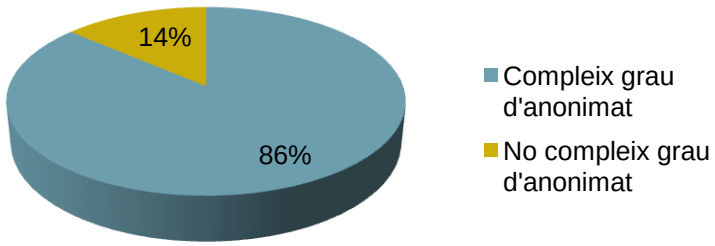

figura 9: Percentatge útil de registres per a k = 5

A la taula 8 es pot observar com evoluciona la relació de registres que compleixen l'anonimat en funció del grau aplicat.

Grau d'anonimat	4	3	2
Nº registres descartats	67	46	10
Percentatge no vàlid	13.45%	9.24%	2.01%
Percentatge vàlid	86.55%	90.76%	97.99%

taula 8: Variació de registres útils amb menor grau de k = 5

En quant a desviació estàndard, observem una estabilització si bé l'augment és generalitzat.

Nº Registres	D. estàndard	Nº Registres	D. estàndard	Nº Registres	D. estàndard
5	0.2291	6	0.4596	17	0.2868
5	0.4700	7	0.4670	22	0.5285
6	0.3871	8	0.3220	68	0.4191
6	1.7899	14	0.4920	257	0.2849
6	0.6299				

taula 9: Desviació estàndard per a k = 5

5.1.5. Grau d'anonimat 6

Per al grau 6 s'han calculat 41 clústers. A la figura 10 es pot apreciar que es reprèn la tendència general, augmentant considerablement el nombre de registres per al clúster majoritari on ja recull el *67,1%* del total del registres.

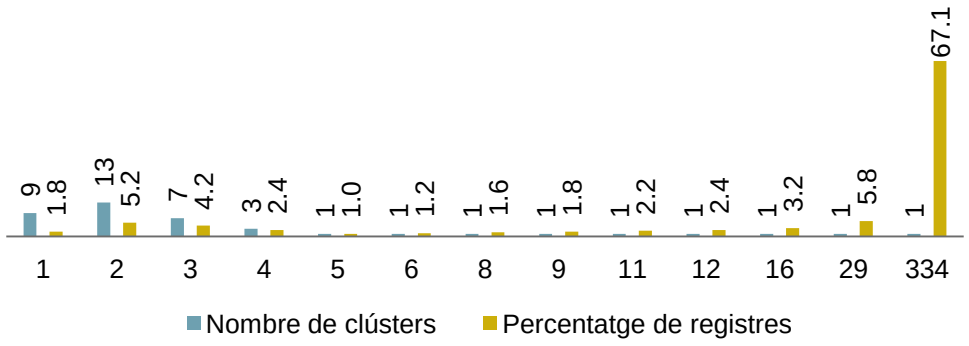

figura 10: Distribució de clústers per a k = 6

A partir d'aquest grau d'anonimat es pot observar una estabilització en el nombre de registres que no compleixen amb el grau d'anonimat.

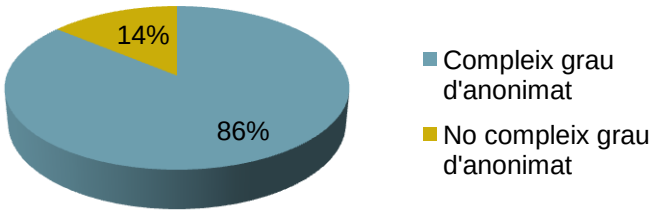

figura 11: Percentatge útil de registres per a k = 6

La taula 10 ens mostra com evoluciona aquesta relació en funció del nou grau d'anonimat.

Grau d'anonimat	5	4	3	2
Nº registres descartats	68	56	35	9
Percentatge no vàlid	13.65%	11.24%	7.03%	1.81%
Percentatge vàlid	86.55%	88.76%	92.97%	98.19%

taula 10: Variació de registres útils amb menor grau de k = 6

Mantenint l'evolució lògica, la desviació estàndard va augmentant a mida que també ho fa el grau d'anonimat mínim.

Nº Registres	D. estàndard	Nº Registres	D. estàndard
6	1.0517	12	0.6415
8	0.5441	16	0.6467
9	0.9715	29	0.7370
11	0.7114	334	0.6937

taula 11: Desviació estàndard per a k = 6

5.1.6. Grau d'anonimat 7

El darrer grau d'anonimat és el 7 i ens facilita un nombre total de 35 clústers. Les dades obtingudes per aquest grau són molt similars al grau anterior tal i com es pot apreciar a la figura 12.

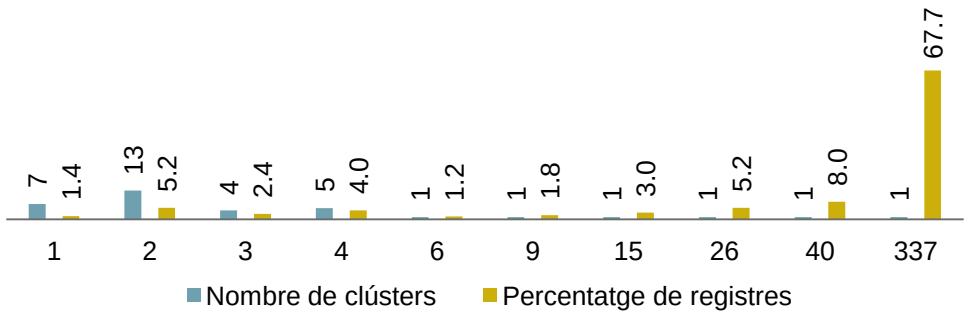

figura 12: Distribució de clústers per a k = 7

Aquesta mateixa estabilitat es segueix observant al nombre de registres a eliminar per no complir amb l'anonimat desitjat.

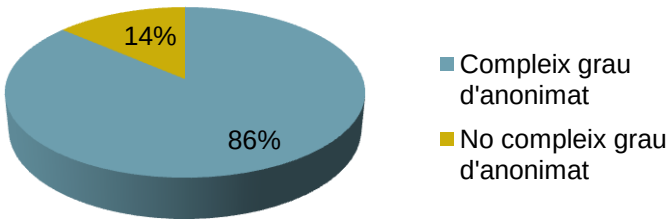

figura 13: Percentatge útil de registres per a k = 7

A la taula 12 es mostra l'evolució del percentatge segons el grau d'anonimat.

Grau d'anonimat	6	5	4	3	2
Nº registres descartats	65	65	45	33	7
Percentatge no vàlid	13.05%	13.05%	9.04%	6.63%	1.41%
Percentatge vàlid	86.95%	86.95%	90.96%	93.37%	98.59%

taula 12: Variació de registres útils amb menor grau de k = 7

Finalment, l'evolució de la desviació estàndard per al grau 7 es pot visualitzar a la taula 13.

Nº Registres	D. estàndard
9	0.9715
15	1.4647
26	1.5902
40	1.2182
337	0.7216

taula 13: Desviació estàndard per a k = 7

5.1.7. Valoracions generals

Per tal de veure com afecta el grau d'anonimat mínim en la pèrdua d'informació s'ha realitzat un gràfic amb el nombre de registres eliminats en funció del grau d'anonimat. A la figura 14 s'observa com es produeix un increment significatiu a partir del grau 3, però que aquest decreix a partir del grau 5 on s'estabilitza.

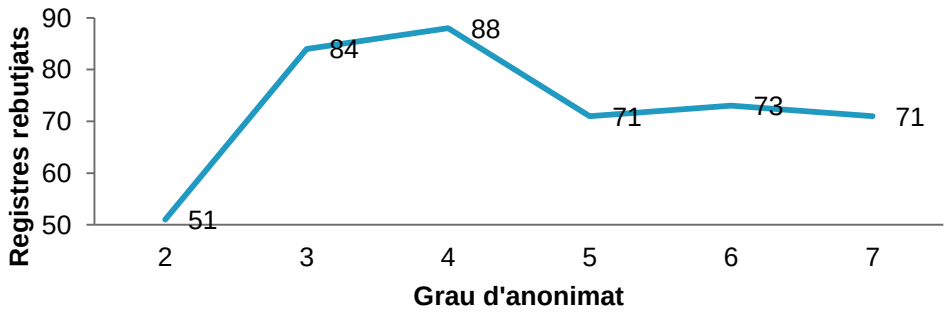

figura 14: Registres rebutjats en funció de k

Però un altre punt de vista sobre la quantitat d'informació útil perduda es percep amb la desviació estàndard. La figura 15 mostra la mitja de la desviació estàndard en funció de cada un dels graus d'anonimat. Aquesta gràfica no clarifica la relació entre l'anonimat i la pèrdua d'informació ja que existeixen valors extrems que desvirtuen la tònica general.

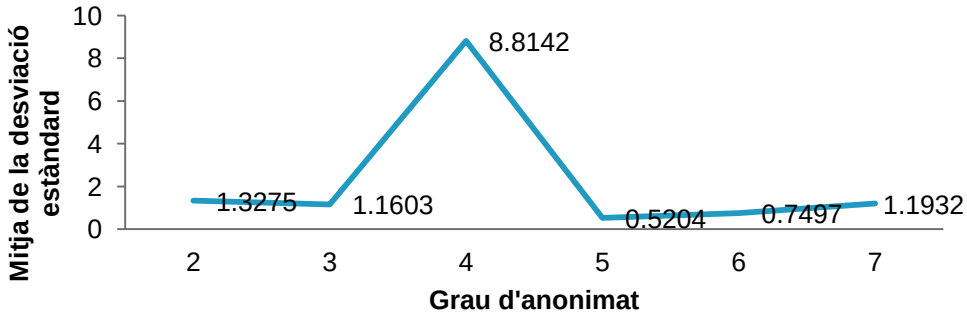

figura 15: Desviació estàndard mitjana en funció de k

Per tal de facilitar unes dades més clarificadores s'ha optat per suprimir aquells valors que sobrepassin un llindar superior al doble de la mitja de la desviació estàndard. Aquest nou filtratge genera la figura 16, on clarament s'observa com la desviació estàndard general va augmentant així com ho fa el grau d'anonimat.

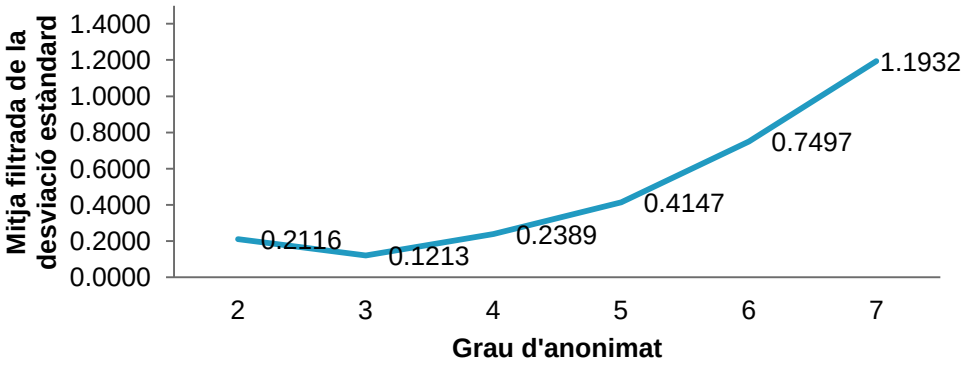

figura 16: Desviació estàndard filtrada mitjana en funció de k

6. CONCLUSIONS I PROPOSTA DE MILLORES

Aquest document mostra l'estudi sobre el tractament de dades stream en la recerca del seu anonimat per a propòsits de difusió estadística. Gràcies a l'estudi de micro-agregació i d'algorismes d'agrupació de dades s'ha arribat a conclusions en la relació directa entre el guany d'anonimat i la pèrdua d'informació.

L'algorisme d'agrupació que ha facilitat la implementació i la obtenció de resultats ha estat *CluStream,* que junt amb el filtratge de dades inicial ha donat un marc de treball per a la presentació de resultats. S'ha cercat un grup de dades offline com a mostra del potencial en quant a anonimització de dades stream, dades prevenients de la xarxa social *Twitter* i que representen el registre dels usuaris.

Amb els resultats obtinguts s'ha pogut corroborar que l'anonimització de dades suposa una pèrdua d'informació, on quan major és el grau de privadesa de les dades, major és també la quantitat d'informació perduda.

A la realització del treball s'han afrontat diferents problemes. El més important ha estat la impossibilitat de poder donar tractament a gran volum de dades, ja que l'algorisme *CluStream* sobre l'entorn *MOA* limitava el nombre de registres resultants, distorsionant així els resultats i les conclusions.

Durant el procés de micro-agregació s'han testejat altres algorismes com *CobWeb* per tal de comparar eficiència i precisió, però els resultats obtinguts no eren equivalents amb *CluStream*, ja que no existia la possibilitat de modificar el nombre d'agrupacions i més bé es podia modificar la distribució de registres per agrupació en funció de la pèrdua d'informació tolerable.

6.1 Proposta de millores

Com a proposta de millora del treball queda pendent el tractament d'un gran volum de dades comparable a un flux continu de dades stream tal i com succeeix de forma online. Un altre millora resideix en la utilització de diversos algorismes d'agrupació de dades i comprar la seva eficiència i precisió i com afecta a la privadesa i pèrdua d'informació.

6.2 Pla de treball

De cara a la realització del treball s'han seguit una sèrie de fases.

- Comprensió i estudi de conceptes rellevants al treball. *(1ª setmana)*
  - Anonimat i privadesa requerida a les dades stream.
  - Aprofundiment al concepte de micro-agregació.
- Identificar un conjunt de dades stream al que aplicar el mètode d'anonimització. *(1ª setmana)*
  - Elecció dels atributs definitius.
  - Estudiar i decidir proposta per protegir la informació.
- Estudi i millora del coneixements sobre el software necessari per la implementació del treball. *(2ª- 3ª setmana)*
  - Configuració de l'entorn *IDE* amb *MOA*.
  - Aprenentatge de la mecànica d'ús del software *MOA*.
- Disseny de l'aplicació del mètode al stream de dades. *(4ª – 7ª setmana)*

    - Elecció de l'algorisme *CluStream*
    - Configuració del mètode *Java* per a la micro-agregació.
    - Posta en pràctica de la proposta de protecció de la informació elegida.
- Implementació del mètode mitjançant el software *MOA* i recopilació de dades. *(8ª – 10ª setmana)*
    - Conversió de fitxer a format *ARFF*.
    - Generar taules i càlculs sobre els resultats obtinguts.
- Realització de una memòria del treball. *(11ª-12ª setmana)*
- Realització de la presentació del treball de fi de màster. *(12ª setmana)*

6.3 Viabilitat

Per la realització del treball s'ha disposat d'un període de 12 setmanes. Aquest temps ha estat suficient per a la realització del treball amb petites demores derivades de la implementació del programari *MOA* amb l'algorisme *CluStream*.

El treball no ha generat costos econòmics ja que el software emprat per a la implementació és open-source de llicència gratuïta. A més, no ha estat necessari la adquisició de recursos addicionals als existents. Els aspectes teòrics s'han extret de la base de dades acadèmica *IEEE Xplore* amb ús de llicència de la UOC, el que suposa no haver assumit despeses.

6.4 Valoració personal

A nivell personal aquest treball ha suposat un nou punt de vista en privadesa en les tecnologies de la informació i de la comunicació, que junt amb la resta de coneixements adquirits al llarg del màster s'afegeixen als obtinguts a la carrera de telecomunicacions i que de ben segur seran de gran profit per a futures metes professionals.

BIBLIOGRAFIA

[1] J. Domingo-Ferrer, V. Torra "*Ordinal, Continuous and Heterogeneous k-Anonymity Through Microaggregation.*" Data Mining and Knowledge Discovery, 11, 195–212, 2005

[2] B. Durratn, E. Frank, L. Hunt, G. Holmes, M. Mayo, B. Pfahringer, T. Smith, I. Witten. "*Machine Learning*". http://www.cs.waikato.ac.nz/ml/index.html

[3] "*TortoiseHg. Mercurial*". http://tortoisehg.bitbucket.org/

[4] A. Bifet, G. Holmes, B. Pfahringer, P. Kranen, H. Kremer, T. Jansen, T. Seidl. "*MOA: Massive Online Analysis, a Framework for Stream Classification and Clustering*."

[5] C. C. Aggarwal, J. Han, J.Wang, and P. S. Yu. "*A framework for clustering evolving data streams*." In VLDB, pages 81–92, 2003.

[6] http://www.cs.waikato.ac.nz/~ml/weka/arff.html