

Diseño Editorial Social en Portales de Información Utilizando Técnicas de Minería De datos

Daniel Gómez M.

DCC - Universidad de Chile

10 de agosto de 2007

1 Introducción

2 Implementación

- módulo: indexacion de documentos
- módulo: clustering batch
- módulo: clustering online

3 Resultados

4 Conclusiones

1 Introducción

2 Implementación

- módulo: indexacion de documentos
- módulo: clustering batch
- módulo: clustering online

3 Resultados

4 Conclusiones

Motivación

- Internet
... visto como un medio de comunicación

Motivación

- Internet
... visto como un medio de comunicación
- Periodismo
... cambio en la forma de compartir y difundir información

Motivación

- Internet
 - ... visto como un medio de comunicación
- Periodismo
 - ... cambio en la forma de compartir y difundir información
- Web 2.0
 - ... todos publican y comparten contenidos
 - ... los usuarios forman comunidades
 - ... cada consumidor de información asume un rol activo

Motivación



Motivación



Motivación

- Sindicación
... cada usuario puede decidir sus fuentes de información
- RSS
... estándar para la distribución de contenidos

Motivación

- Sindicación
... cada usuario puede decidir sus fuentes de información
- RSS
... estándar para la distribución de contenidos

<< todos tienen derecho a producir/difundir contenidos por igual >>

Motivación

- Sindicación
... cada usuario puede decidir sus fuentes de información
- RSS
... estándar para la distribución de contenidos

<< todos tienen derecho a producir/difundir contenidos por igual >>
- Gran flujo de artículos
... necesidad de mejorar formas de visualizar dicho contenido

Motivación

- Sindicación
... cada usuario puede decidir sus fuentes de información
- RSS
... estándar para la distribución de contenidos

<<todos tienen derecho a producir/difundir contenidos por igual>>
- Gran flujo de artículos
... necesidad de mejorar formas de visualizar dicho contenido
 - buscadores de noticias / blogs
 - portadas “sociales”

Motivación

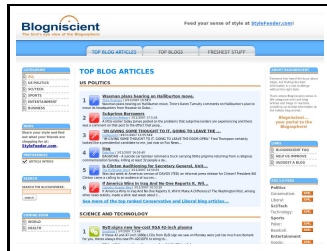
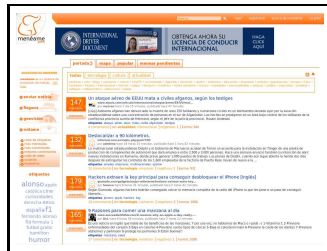
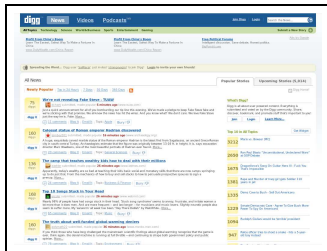
The screenshot shows the Technorati homepage with a search bar and a list of popular blogs. The 'Top 100' list includes blogs like 'The Daily Beast', 'The New York Times', and 'The Washington Post'. The 'Top 100' list is a collection of the most popular blogs on the platform, ranked by the number of links they have. The interface is green and white, with a sidebar on the left and a main content area on the right.

The screenshot shows the Topix homepage with a search bar and a list of popular blogs. The 'Top 100' list includes blogs like 'The Daily Beast', 'The New York Times', and 'The Washington Post'. The interface is blue and white, with a sidebar on the left and a main content area on the right.

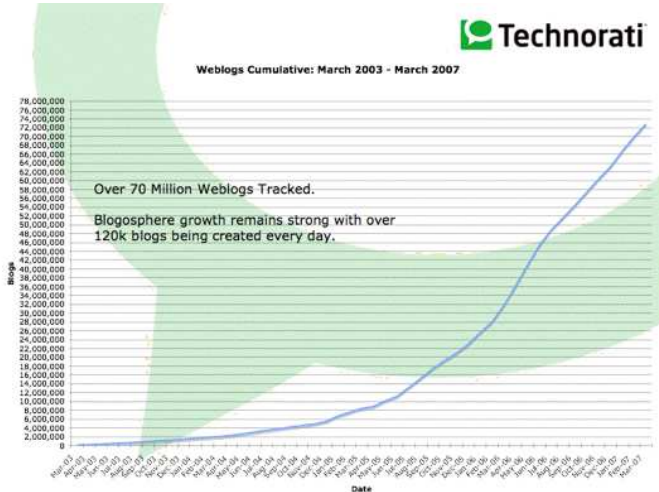
The screenshot shows the BlogPulse homepage with a search bar and a list of popular blogs. The 'Top 100' list includes blogs like 'The Daily Beast', 'The New York Times', and 'The Washington Post'. The interface is blue and white, with a sidebar on the left and a main content area on the right.

The screenshot shows the Google News homepage with a search bar and a list of popular blogs. The 'Top 100' list includes blogs like 'The Daily Beast', 'The New York Times', and 'The Washington Post'. The interface is blue and white, with a sidebar on the left and a main content area on the right.

Motivación



Estado de la Blogósfera



orbitando / chile / temas

Top 10

Blog

Noticias

Un Secreto...
Google Video - Video Search [Chile -chili] | Hace 9 horas
Un Secreto... katedivina1 min 40 sec - Aug 6, 2007La gringa habla castellano! A little nerve-wracking attempting to vlog in my non-native language, but practice makes perfect, right?...

[Comentar](#) [Promueve](#)

MARCEL CLAUDE... IRONIAS PARA LLORAR...
— | Hace 63 minutos
al margen de los monopolios...

[Comenta](#) [Promueve](#)

Siguen ajustes de la agenda legislativa
Senado de Chile - Portada | Hace 79 minutos
Los Presidentes del Senado, Eduardo Frei Ruiz-Tagle y de la Cámara de Diputados, Patricio Walker se reunirán hoy a las 12,30 horas con el Ministro Secretario General de la Presidencia, José Antonio Villalón...

[Comenta](#) [Promueve](#)

Guerra de cartas: Zaldívar contra Espejo
Hoy me interesa... | Hace 63 minutos
Hoy en Reportajes de El Mercurio, puedes leer una carta que el ex ministro Sergio Espejo le mandó a Adolfo Zaldívar y la respuesta de éste último. Lo interesante es que mientras Espejo escribe en forma...

[Comenta](#) [Promueve](#)

Copywriter 2007: Festival de Redacción Publicitaria v2.0
Zapping Digital 2.0 | Hace 58 minutos
La idea de Copywriter es poder sentir la esencia de ser un [copy de verdad]. El concurso está orientado a marcar un precedente de importancia no solo en Chile, si no también en Ibero América, consid...

[Comenta](#) [Promueve](#)

previa maga dep. santa cruz
Google Video - Video Search [Chile -chili] | Hace 9 horas
previa maga dep. santa cruzclubmagallanes5 min 34 sec - Aug 6, 2007previa maga dep. santa cruz...

[Comenta](#) [Promueve](#)

Mayne-Nicholls visitó nuevamente a Bielsa esta mañana
lasegunda.com - Google Noticias | 06-agosto-2007, 11:55
Mayne-Nicholls visitó nuevamente a Bielsa esta mañanaLa Segunda - hace 39 minutosTras esta cita el estratega trasandino estaría viajando a Argentina, por lo que se descartaría una presentación en hora...

[Comenta](#) [Promueve](#)

En Chile esperan la decisión de Marcelo Bielsa - Lanacion.com (Argentina)

SIGUEN LOS MISTERIOS SOBRE DT DE LA ROJA: MARCELO BIELSA REGRESO A ... - Terra Chile
Mayne-Nicholls Dijo que Aún No está Confirmado Marcelo Bielsa como ... - Radio Infinita

Mayne-Nicholls visitó nuevamente a Bielsa esta mañana

Harold Mayne-Nicholls tuvo un matinal encuentro con Marcelo Bielsa - Radio Cooperativa

temas [Ranking Blogs](#)

argentina august 4
bachelet bolivia brasil
codelco flickr fotolog
futbol oportunidad peru policia politico
radio salud santiago de chile
senador torneo
trabajadores
transantiago universidad de chile

Comentarios

Blogs

Videos

Fotografías

Podcast

Noticias

- Política
- Negocios
- Tecnología
- Deportes
- Cultura
- Espectáculos
- News English
- Avisos
- Foros

Blog Oficial

[RSS FEED](#)

Envíanos tu blog

Propuesta

agrupar tópicos. . .

. . .deseamos crear una nueva forma de visualizar los contenidos.

Propuesta

agrupar tópicos. . .

. . . deseamos crear una nueva forma de visualizar los contenidos.

- portada de titulares

una “headlines cloud”



Propuesta

agrupar tópicos. . .

. . . deseamos crear una nueva forma de visualizar los contenidos.

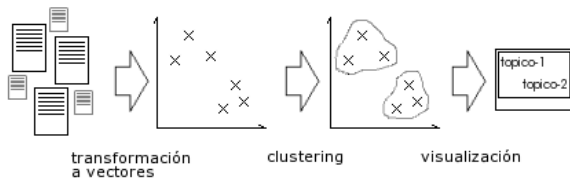
- portada de titulares

una “headlines cloud”



- casos de uso
 - consulta de los top-k tópicos
 - consulta de artículos relacionados

Metodología



Metodología

- indexación de documentos

Metodología

- indexación de documentos
- clustering

Metodología

- indexación de documentos
- clustering
- clustering online

Metodología

- conjunto de referencia (maqueta !!)
- indexación de documentos
- clustering
- clustering online

Conjunto de Referencia

- universo: artículos período enero-marzo 2007

Conjunto de Referencia

- universo: artículos período enero-marzo 2007
- muestra de 2000 artículos

Conjunto de Referencia

- universo: artículos período enero-marzo 2007
- muestra de 2000 artículos
- categorización manual de tópicos

Conjunto de Referencia

- universo: artículos período enero-marzo 2007
- muestra de 2000 artículos
- categorización manual de tópicos
- objetivos (para que?)
 - conocer los datos
 - disponer de un conjunto de testing

Conjunto de Referencia

los 10 tópicos más populares del período ene-mar 2007 en Chile:

tópico	# artículos
transantiago	52
sudamericano de futbol sub 20	37
michell bachelet	27
copa davis chile-rusia	24
resultados club u.de chile	22
conflicto limitrofe con peru	21
fernando gonzalez	20
chiledeportes	18
colo colo	18
copa america de futbol	18

Conjunto de Referencia

Conclusiones

- cuidado con artículos repetidos
... considerar sólo artículos únicos (post-procesados)
- detectar artículos que no aportan información
 - eliminar artículos en ingles (taxonomy = *english*)
 - eliminar artículos de flickr / fotolog
- se conoce forma de los artículos
... caso google-news
- se conoce dinámica de los tópicos
... comportamiento en el tiempo de los tópicos

1 Introducción

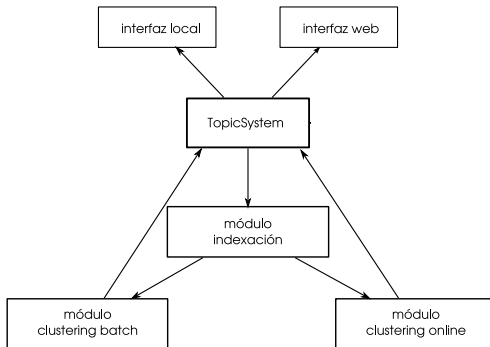
2 Implementación

- módulo: indexacion de documentos
- módulo: clustering batch
- módulo: clustering online

3 Resultados

4 Conclusiones

Implementación



Módulo: indexación de documentos

Objetivos:

- representar documentos como vectores
- extraer la semántica de los artículos
- dimensionalidad baja (o lo más baja posible)

Módulo: indexación de documentos

necesitamos traducir texto a algo que entienda el computador...

VSM: Vector Space Model

$$D_i \rightarrow d_{i1}, d_{i2}, \dots, d_{ik}$$

Módulo: indexación de documentos

necesitamos traducir texto a algo que entienda el computador...

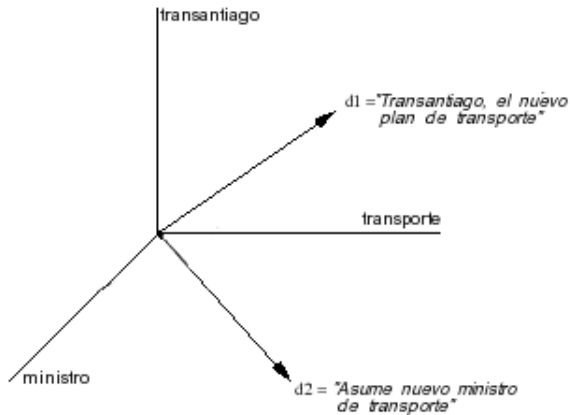
VSM: Vector Space Model

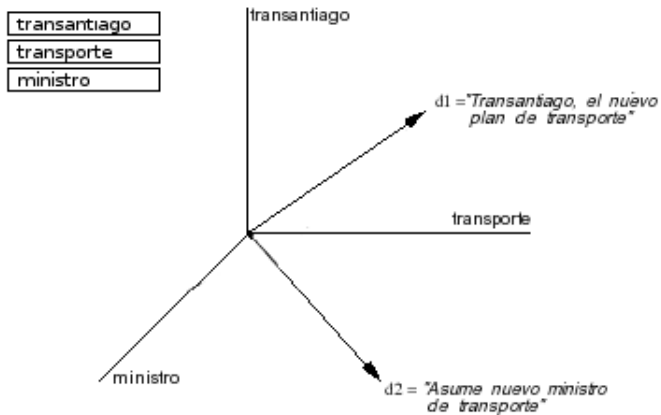
$$D_i \rightarrow d_{i1}, d_{i2}, \dots, d_{ik}$$

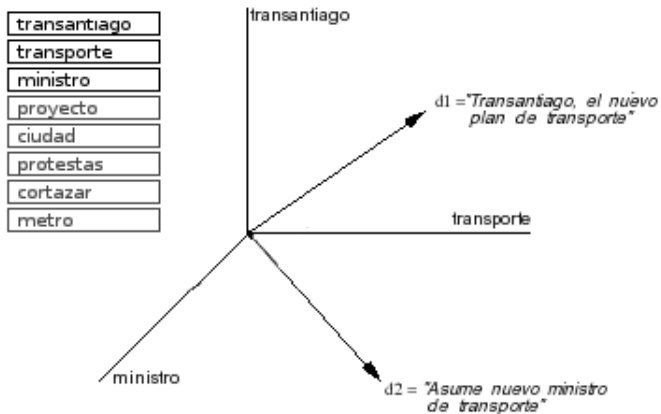
necesitamos relacionar términos con documentos...

Funciones de peso:

- binaria : $d_{ij} = 1$ si existe, 0 si no
- frecuencia del término : $d_{ij} = f_{ij}$
- frecuencia normalizada : $d_{ij} = \frac{f_{ij}}{n_i}$
- TF-IDF : $d_{ij} = TF * IDF = \frac{f_{ij}}{n_j} * -\log\left(\frac{n_i}{N}\right)$







Módulo: indexación de documentos

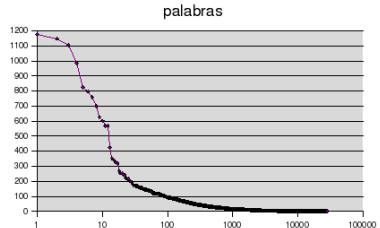
Principal desafío:

reducción de la dimensionalidad!!

Módulo: indexación de documentos

Principal desafío:

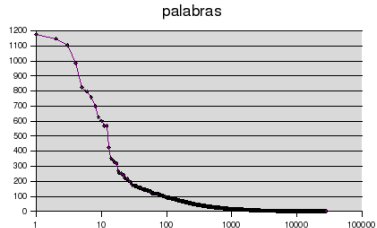
reducción de la dimensionalidad!!



Módulo: indexación de documentos

Principal desafío:

reducción de la dimensionalidad!!



- stemming: reducción a la raíz
- filtrado de stopwords
- poda de la lista de palabras

Módulo: indexación de documentos

Herramientas disponibles

Lucene - Weka - WVTool

Módulo: indexación de documentos

Herramientas disponibles

Lucene - **Weka** - WVTool

Módulo: indexación de documentos

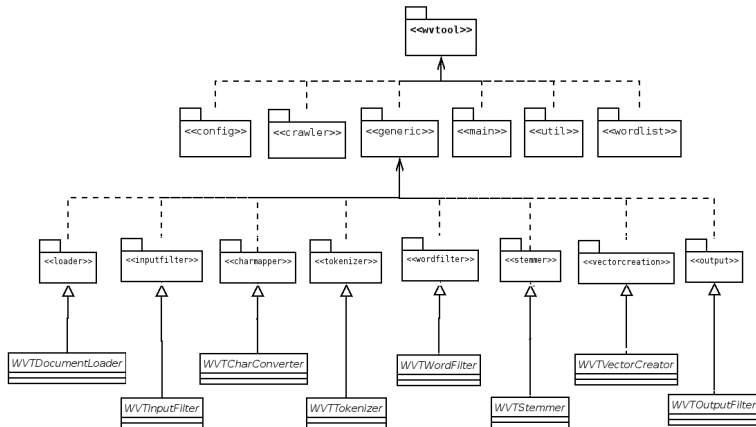
Herramientas disponibles

Lucene - Weka - **WVTool**

WVTool

The Word & Web Vector Tool is a flexible Java library for statistical language modeling and integration of Web and Webservice based data sources. It supports the creation of word vector representations of text documents in the vector space model.

Módulo: indexación de documentos



Módulo: indexación de documentos

Loader → InputFilter → Charnapper → Tokenizer → WordFilter → Stemmer

carga el documento desde la base de datos:
{titulo + descripcion}

Módulo: indexación de documentos

Loader → **InputFilter** → Charnapper → Tokenizer → WordFilter → Stemmer

primer filtro sobre la entrada:

- filtrado de tags-html
- eliminación de texto de enlaces
- eliminación de nombres de medios

Módulo: indexación de documentos

Loader → InputFilter → **Charmapper** → Tokenizer → WordFilter → Stemmer

conversión de caracteres:

reemplazos: {A, á, á}→{a}

Módulo: indexación de documentos

Loader → InputFilter → Charnapper → **Tokenizer** → WordFilter → Stemmer

separación en palabras (tokens)

Módulo: indexación de documentos

Loader → InputFilter → Charnapper → Tokenizer → **WordFilter** → Stemmer

filtrado de stopwords:

se eliminan palabras que no ayudan a discriminar.

- adjetivos - articulos- preposiciones - verbos
- stopwords ingles : the, be, have, has, that
- html : table, style, br
- rss : noticias, nacional, articulos, relacionados

Módulo: indexación de documentos

Loader → InputFilter → Charnapper → Tokenizer → WordFilter → Stemmer

reducción a la raíz

reemplazos: {politica, politicos, politico} → {politic}

Módulo: indexación de documentos

VectorCreator → OutputFilter →

creación del vector del documento

elección de la función de peso: TF-IDF, TF, u otra...

Módulo: indexación de documentos

VectorCreator → OutputFilter →

Idea...

ponderar palabras según su posición en el texto.



VectorCreator → OutputFilter →

ponderar palabras según su posición en el texto.



$$\text{función PF-IDF: } f(D_j, t_i) = PF * IDF = (\frac{f_{ij}}{|D_j|} * pos) * (-\log(\frac{f_j}{|N|}))$$

Módulo: indexación de documentos

VectorCreator → **OutputFilter** →

vector se agrega al conjunto de vectores del espacio.
... se guarda a disco.

Módulo: indexación de documentos

Implementación:

```
WVTString2Vector implements String2Vector  
    load(source)  
    createVSpace()  
    addVector(article)  
    updateWordList(action)
```

Módulo: clustering batch

Clustering

... el proceso de segmentar el espacio vectorial.

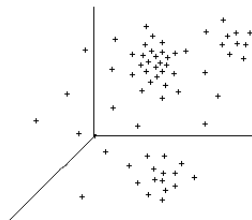
Buscamos identificar las agrupaciones
que se forman naturalmente en el
espacio de artículos.

Módulo: clustering batch

Clustering

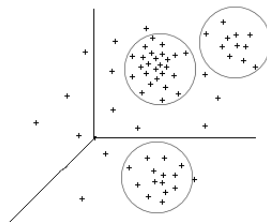
... el proceso de segmentar el espacio vectorial.

Buscamos identificar las agrupaciones
que se forman naturalmente en el
espacio de artículos.



Clustering

Buscamos identificar las agrupaciones que se forman naturalmente en el espacio de artículos.



Módulo: clustering batch

Algoritmos de clustering

- algoritmos de particionamiento
- algoritmos jerárquicos
- algoritmos basados en densidad
- algoritmos basados en grafos

Módulo: clustering batch

Algoritmos de clustering

- algoritmos de particionamiento
- algoritmos jerárquicos
- algoritmos basados en densidad
- algoritmos basados en grafos

Noción de distancia:
distancia coseno :

$$sim(u, v) = \cos\theta = \frac{\vec{v}_1 * \vec{v}_2}{\|\vec{v}_1\| \|\vec{v}_2\|}$$

Módulo: clustering batch

Cluto

CLUTO is a family of computationally efficient and high-quality data clustering and cluster analysis programs & libraries, that are well suited for low- and high-dimensional data sets.

Módulo: clustering batch

```
*****  
vcluster (CLUTO 2.1.1) Copyright 2001-03, Regents of the University of Minnesota
```

```
Matrix Information -----
```

```
Name: /tmp/ts0926.mat, #Rows: 100, #Columns: 284, #NonZeros: 28400
```

```
Options -----
```

```
CLMethod=RB, CRfun=I1, SimFun=Cosine, #Clusters: 10  
RowModel=None, ColModel=None, 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=1.89e+01] [100 of 100]  
-----
```

cid	Size	ISim	ISdev	ESim	ESdev	
0	3	+0.548	+0.024	+0.009	+0.003	
1	4	+0.469	+0.052	+0.026	+0.005	
2	5	+0.392	+0.045	+0.024	+0.006	
3	8	+0.245	+0.034	+0.025	+0.007	
4	9	+0.219	+0.032	+0.021	+0.006	
5	9	+0.203	+0.029	+0.017	+0.010	
6	10	+0.204	+0.033	+0.025	+0.007	
7	10	+0.179	+0.028	+0.027	+0.011	
8	21	+0.114	+0.031	+0.025	+0.017	
9	21	+0.071	+0.033	+0.018	+0.015	

```
-----  
10-way clustering solution - Descriptive & Discriminating Features...  
-----
```

Módulo: clustering batch

Parámetros del proceso:

- *función de similaridad/distancia*
- *función objetivo : i1, i2, e1, g1, g1p, h1, h2*
- *método de particionado*
- *criterio de bisección*
- *rowmodel, colmodel:*
- *número de intentos:*
- *número de iteraciones:*

Módulo: clustering batch

Problemas:

- función objetivo:
... medida de similaridad interna.

$$I_1 = \sum_{i=1}^k \frac{1}{n_i} \left(\sum_{u,v \in S_i} \text{sim}(u, v) \right)$$

$$I_2 = \sum_{i=1}^k \sqrt{\sum_{u,v \in S_i} \text{sim}(u, v)}$$

- número de clusters:
 $k = ?$

Módulo: clustering batch

Implementación:

```
ClutoVector2Cluster implements Vector2Cluster  
    load(source)  
    run()  
    saveClusters()
```

Módulo: clustering online

Motivación:

... necesitamos la capacidad de detectar nuevos tópicos.

Módulo: clustering online

Motivación:

... necesitamos la capacidad de detectar nuevos tópicos.

Clustering incremental !!

Módulo: clustering online

Motivación:

... necesitamos la capacidad de detectar nuevos tópicos.

Clustering incremental !!

alternativas:

- COBWEB
- DBSCAN
- SinglePass
- KNN: k-nearest neighbor
- k-means incremental

Módulo: clustering online

problemas

- wordlist variable

Aparecen constantemente nuevas palabras: agregarlas todas, no agregar ninguna, agregar “inteligentemente”.

- tópicos basura

A medida que pasa el tiempo algunos tópicos van quedando obsoletos.

Módulo: clustering online

Solución: Single Pass con Ventanas

- ventana de actualización de la wordlist
- ventana de purgado de artículos
- ventana del garbage collector

Módulo: clustering online

Solución: Single Pass con Ventanas

- ventana de actualización de la wordlist
- ventana de purgado de artículos
- ventana del garbage collector

Implementación:

```
SimpleBufferingClusteringOnline implements ClusteringOnline  
    addVector(vector)  
    updateClusters()
```

```
clusteringVector(Article article){  
    ItemVector vector = indexingModule.addVector(article);  
    boolean action = onlineModule.addVector(vector);  
    indexingModule.updateWordList(action);  
    onlineModule.updateClusters();  
}
```

1 Introducción

2 Implementación

- módulo: indexacion de documentos
- módulo: clustering batch
- módulo: clustering online

3 Resultados

4 Conclusiones

Resultados

BASE DE DATOS DE ARTICULOS RSS

inicio

#	id -	topico	#	...
1	77	col (0.553) --- carac (0.079) --- deport (0.053)	28	...
2	777	transantiago (0.339) --- recurs (0.078) --- transport (0.073)	22	...
3	737	festival (0.312) --- cin (0.310) --- internacional (0.121)	20	...
4	94	davis (0.322) --- rusi (0.257) --- cop (0.226)	18	...
5	760	diari (0.387) --- ecuador (0.082) --- xxi (0.057)	18	...
6	728	gonzalez (0.383) --- fern (0.169) --- tenis (0.153)	16	...
7	57	sudamericano (0.329) --- sub (0.281) --- paraguay (0.233)	16	...
8	729	bachelet (0.365) --- michell (0.174) --- president (0.163)	15	...
9	779	chiledeport (0.214) --- director (0.202) --- antecedent (0.063)	15	...
10	767	muse (0.2) --- artist (0.162) --- arte (0.157)	15	...
11	727	cop (0.316) --- amer (0.302) --- futbol (0.097)	14	...
12	97	tenis (0.301) --- mes (0.258) --- circuit (0.219)	14	...
13	764	univers (0.29) --- refuerz (0.132) --- futbol (0.107)	14	...
14	767	libr (0.365) --- softwar (0.081) --- edit (0.059)	14	...
15	778	aument (0.233) --- dolar (0.150) --- millon (0.136)	13	...
16	758	disc (0.312) --- error (0.200) --- band (0.090)	13	...
17	775	clas (0.189) --- alumn (0.169) --- horari (0.141)	13	...
18	772	antofagast (0.46) --- suspend (0.091) --- deport (0.082)	13	...
19	776	ley (0.264) --- proyect (0.095) --- acced (0.080)	13	...
20	23	puel (0.239) --- bolson (0.232) --- hit (0.226)	13	...

Resultados

BASE DE DATOS DE ARTICULOS RSS

inicio

#	id - topico	#	...
1	77 col (0.553) --- carac (0.079) --- deport (0.053)	28	...
2	777 transantiago (0.339) --- recurs (0.078) --- transport (0.073)	22	...
3	737 festival (0.312) --- cin (0.310) --- internacional (0.121)	20	...
4	94 davis (0.322) --- rusi (0.257) --- cop (0.226)	18	...
5	760 diari (0.387) --- ecuador (0.082) --- xxi (0.057)	18	...
6	728 gonzalez (0.383) --- fern (0.169) --- tenis (0.153)	16	...
7	57 sudamericano (0.329) --- sub (0.281) --- paraguay (0.233)	16	...
8	729 bachelet (0.365) --- michell (0.174) --- president (0.163)	15	...
9	779 chiledeport (0.214) --- director (0.202) --- antecedent (0.063)	15	...
10	767 muse (0.2) --- artist (0.162) --- arte (0.157)	15	...
11	727 cop (0.316) --- amer (0.302) --- futbol (0.097)	14	...
12	97 tenis (0.301) --- mes (0.258) --- circuit (0.219)	14	...
13	764 univers (0.29) --- refuerz (0.132) --- futbol (0.107)	14	...
14	767 libr (0.365) --- softwar (0.081) --- edit (0.059)	14	...
15	778 aument (0.233) --- dolar (0.150) --- millon (0.136)	13	...
16	758 disc (0.312) --- error (0.200) --- band (0.090)	13	...
17	775 clas (0.189) --- alumn (0.169) --- horari (0.141)	13	...
18	772 antofagast (0.46) --- suspend (0.091) --- deport (0.082)	13	...
19	776 lev (0.264) --- proyect (0.095) --- acced (0.080)	13	...
20	23 puel (0.239) --- bolson (0.232) --- hit (0.226)	13	...

colo colo

festival de cine

ecuador

sudamericano sub20

chile deportes

transantiago

copa davis

fdo.gonzalez

bachelet

exposiciones

Resultados

BASE DE DATOS DE ARTICULOS RSS

inicio

#	id - topico	#	...
colo colo	1 77 col (0.553) --- carac (0.079) --- deport (0.053)	28	...
	2 777 transantiago (0.339) --- recurs (0.078) --- transport (0.073)	22	...
festival de cine	3 737 festival (0.312) --- cin (0.310) --- internacional (0.121)	20	...
	4 94 davis (0.322) --- rusi (0.257) --- cop (0.226)	18	...
ecuador	5 760 diari (0.387) --- ecuador (0.082) --- xxi (0.057)	18	...
	6 728 gonzalez (0.383) --- fern (0.169) --- tenis (0.153)	16	...
sudamericano sub20	7 57 sudamericano (0.329) --- sub (0.281) --- paraguay (0.233)	16	...
	8 729 bachelet (0.365) --- michell (0.174) --- president (0.163)	15	...
chile deportes	9 779 chiledeport (0.214) --- director (0.202) --- antecedent (0.063)	15	...
	10 767 muse (0.2) --- artist (0.162) --- arte (0.157)	15	...
	11 727 cop (0.316) --- amer (0.302) --- futbol (0.097)	14	...
	12 97 tenis (0.301) --- mes (0.258) --- circuit (0.219)	14	...
	13 764 univers (0.29) --- refuerz (0.132) --- futbol (0.107)	14	...
	14 767 libr (0.365) --- softwar (0.081) --- edit (0.059)	14	...
	15 778 aument (0.233) --- dolar (0.150) --- millon (0.136)	13	...
	16 758 disc (0.312) --- error (0.200) --- band (0.090)	13	...
	17 775 clas (0.189) --- alumn (0.169) --- horari (0.141)	13	...
	18 772 antofagast (0.46) --- suspend (0.091) --- deport (0.082)	13	...
	19 776 lev (0.264) --- proyect (0.095) --- acced (0.080)	13	...
	20 23 puel (0.239) --- bolson (0.232) --- hit (0.226)	13	...
		transantiago	
		copa davis	
		fdo.gonzalez	
		bachelet	
		exposiciones	

1 Introducción

2 Implementación

- módulo: indexacion de documentos
- módulo: clustering batch
- módulo: clustering online

3 Resultados

4 Conclusiones

Conclusiones

- solución a un problema real : creación de una portada
- estudio y aplicación de técnicas conocidas (IR, datamining)
... para el caso particular de los artículos sindicados.
- aplicaciones. . . muchas!!