Una Aproximació d'Aprenentatge Automàtic per a Extracció d'Informació Adaptativa

Memòria de la recerca realitzada amb el suport d'una beca FI
de la Generalitat de Catalunya

Edgar Gonzàlez i Pellicer

Índex

Resum

Les tècniques de clustering poden ajudar a reduir la supervisió en processos d'obtenció de patrons per a Extracció d'Informació.

En aquest treball es comença per estudiar la representació de documents més adequada per a la tasca de clustering. Per tal d'evitar els biaixos dels mètodes individuals de clustering, es consideren mètodes de clustering conjunt. S'exploren diversos mètodes de combinació supervisada, i s'hi afegeixen estratègies automàtiques per a determinar el nombre de clusters de la combinació. També es consideren mecanismes per a obtenir clusterings conjunts ponderats, així com estratègies de combinació no supervisada. Finalment, els resultats del clustering s'utilitzen en un sistema d'adquisició de patrons per a substituir els elements de supervisió humana.

Totes aquestes estratègies i mètodes s'avaluen en tasques de clustering de documents i adquisició de patrons usant dades reals. Es comprova que els mots com representació de documents superen altres models per a la tasca de clustering, així com que el clustering conjunt supera les limitacions dels clusterings individuals, i que les estratègies no supervisades d'adquisició de patrons obtenen resultats competitius respecte a les estratègies supervisades.

Clustering techniques can help in reducing supervision in processes of pattern acquisition for Information Extraction.

In this work we start by studying the document representation most suitable for the task of clustering. To avoid the biases of individual clustering methods, ensemble clustering methods are considered. Several combination clustering methods are explored, and strategies to automatically determine the number of clusters in the combination. Weighted ensemble clusterings are also considered, as well as strategies of unsupervised combination. Finally, the results of clustering are used in a pattern acquisition system, to replace the elements of human supervision.

All these strategies and methods are evaluated in document clustering and pattern acquisition task using real data. It is shown that words outperform other models as document representation for the task of clustering, as well that ensemble clustering overcomes the limitations of individual clusterings, and that unsupervised strategies for pattern acquisition obtain competitive results with respect to supervised strategies.

Capítol 1

Introducció

1.1 Extracció d'Informació

La meva tesi s'emmarca en el Processament del Llenguatge Natural, dins l'àrea de la Intel·ligència Artificial, i té com a eix la incorporació de tècniques de clustering a metodologies d'obtenció de patrons per a Extracció d'Informació, amb l'objectiu de reduir-ne la supervisió.

L'Extracció d'Informació busca l'obtenir automàticament informació estructurada a partir de documents textuals. Així, si es té una notícia com la de la figura 1.1, corresponent a un partit de basquetbol, l'objectiu podria ser identificar-hi els equips, jugadors i entrenadors que hi apareixen, així com les relacions que existeixen entre ells, obtenint una estructura com la de la figura 1.2.

Els primers sistemes d'Extracció d'Informació es van desenvolupar durant la dècada dels 1970 (DeJong, 1979), però fins a la dècada dels 1990 es tractava de sistemes molt complexos i que requerien molt d'esforç humà, tant en la seva construcció com en la seva adaptació a diversos dominis. Tanmateix, a principis dels 1990 es va produir una simplificació de l'arquitectura dels sistemes d'Extracció d'Informació (Hobbs, 1993), fet que va permetre l'aplicació de mètodes d'Aprenentatge Automàtic. L'objectiu era trobar tècniques per adaptar els sistemes de manera eficaç, eficient i el menys supervisada possible (Riloff, 1993).

Encara que recentment s'han desenvolupat sistemes d'obtenció de patrons per a Extracció d'Informació que requereixen una quantitat molt petita de supervisió, com ara els basats en bootstrapping (Yangarber, 2003; Surdeanu et al., 2006), aquesta encara hi és present, sigui en la forma d'un conjunt inicial de patrons o en la forma d'una classificació dels documents d'un conjunt en rellevants i no rellevants a la tasca (figura 1.3a). La meva idea és reduir encara més aquesta supervisió utilitzant tècniques de clustering.

El terme clustering engloba el conjunt de tècniques que tenen com a objectiu agrupar objectes similars dins un conjunt, d'acord a una certa noció de semblança. La utilitat de les tècniques de clustering com a eines automàtiques per a l'anàlisi exploratòria de col·leccions de dades ha estat provada en repetides ocasions (Hartigan, 1975; Dimitriadou, 2003).

La proposta de la meva tesi és doncs millorar el procés d'aprenentatge de patrons per a Extracció d'Informació utilitzant tècniques de clustering de documents, per tal de reduir els elements de

Victòria del Sant Hipòlit a la pista del La Gleva
Els homes del Sant Hipòlit han demostrat la seva vàlua a domicili, aconseguint un resultat de 67 - 87 davant un La Gleva en què Joan Genís no ha brillat com habitualment, i no ha passat d'uns modestos 3 punts. En canvi, el base Francesc Veguer ha aconseguit 14 punts per als Hipolitencs, i ha estat una peça determinant en la victòria del conjunt de Jaume Forns.

Figura 1.1: Exemple de notícia

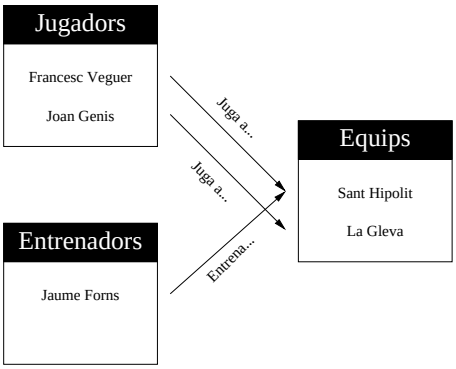


Figura 1.2: Informació extreta de la notícia anterior

supervisió humana que hi són presents. L'objectiu és el desenvolupament d'una metodologia tal que, a partir d'una col·lecció de documents sense cap mena d'anotació, i sense requerir cap mena de documents o patrons llavor manuals, obtingui patrons de bona qualitat i útils per a Extracció d'Informació i altres tasques de Processament del Llenguatge Natural.

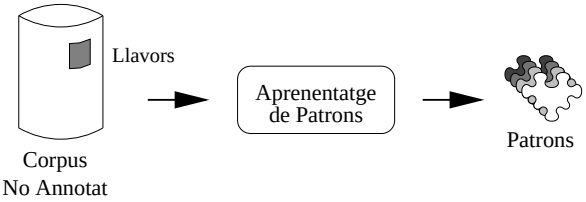
1.2 Aproximació

Als primers experiments que havíem dut a terme sobre el procés d'aprenentatge de patrons (descrits al meu projecte de tesi (Gonzàlez, 2007), juntament amb una introducció més detallada als temes de l'Extracció d'Informació i l'obtenció automàtica de patrons) havíem utilitzat un esquema seqüencial: els documents d'una col·lecció són primerament clusteritzats, i posteriorment s'identifiquen els patrons rellevants a cada clúster (figura 1.3b).

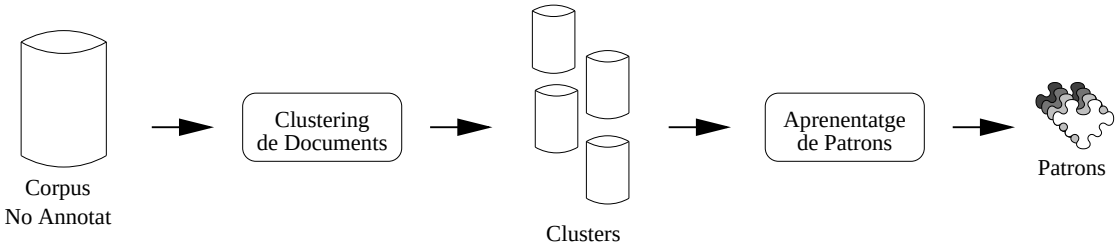
Tanmateix, al projecte de tesi també havia presentat dos esquemes alternatius a l'aprenentatge de patrons usant tècniques de clustering: un de col·laboratiu en què el clustering dels documents i l'aprenentatge dels patrons intercanvien informació per a obtenir una solució final (figura 1.3c); i un de conjunt en què el clustering i l'aprenentatge són resultat d'un procés únic (figura 1.3d).

La factibilitat de l'aplicació de tècniques de clustering a l'aprenentatge de patrons depèn clarament de l'existència de mètodes de clustering adequats. I per aquesta raó, el principal focus de recerca en una primera fase de la tesi va ésser el clustering de documents. Posteriorment, la recerca va desplaçar-se cap als procediments d'adquisició de patrons.

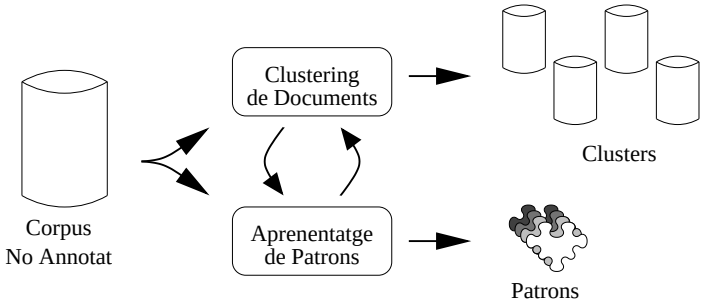
Aquest document recull les línies explorades i els resultats obtinguts durant el període de gaudi de la beca FI de la Generalitat de Catalunya. El capítol 2 recull la recerca referent als aspectes de clustering, i el capítol 3 en recull la referent a l'aprenentatge de patrons. Un darrer capítol 4 recull conclusions que es poden extreure de la recerca realitzada.



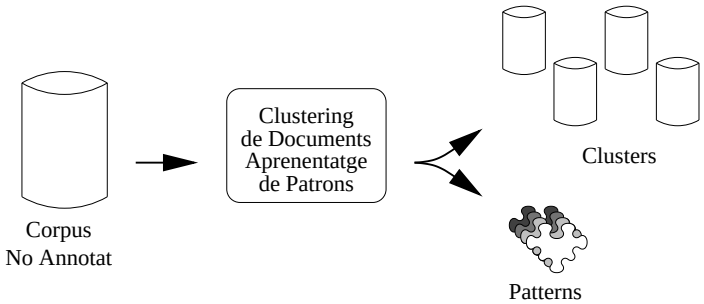
a) Aprenentatge de patrons amb selecció manual de llavors



b) Clustering i aprenentatge de patrons seqüencials



c) Clustering i aprenentatge de patrons col·laboratius



d) Clustering i aprenentatge de patrons conjunts

Figura 1.3: Esquemes per a l'aprenentatge de patrons d'Extracció d'Informació

Capítol 2

Clustering

La tasca de clustering es pot definir com *"el procés de particionar un conjunt de patrons en grups coherents, disjunts i homogenis, anomenats clusters"*[1] (Tasoulis i Vrahatis, 2004). Aquestes tècniques han estat estudiades i utilitzades en moltes àrees de l'Aprenentatge Automàtic i el Reconeixement de Patrons. Dos estudis del ventall de tècniques existents per a clustering es poden trobar a Jain et al. (1999) i Xu i Wunsch (2005).

Encara que el clustering de documents és principalment una tasca no supervisada, en molts mètodes de clustering romanen elements de supervisió. Sovint cal obtenir, de l'usuari o mitjançant un procés extern, el nombre de clusters k, o un conjunt de documents *llavor* a partir de què arrencar el procés de clustering (Zhao i Karypis, 2002, 2004). El clustering no supervisat es pot definir doncs com el procés d'agrupació de patrons sense conèixer a priori el nombre de clusters.

Donat que l'objectiu de la meva tesi és reduir els elements de supervisió en el procés d'aprenentatge de patrons, el nostre treball es va centrar en aquest plantejament no supervisat del problema de clustering.

Les aproximacions *clàssiques* a la tasca del clustering no supervisat consisteixen en aplicar repetidament un algorisme de clustering supervisat iteratiu (per exemple, k-Means (McQueen, 1967)) amb diferents nombres de clusters i/o condicions inicials; i, posteriorment, escollir el millor dels clusterings obtinguts utilitzant criteris de selecció de models (Calinski i Harabasz, 1974; Rissanen, 1978; Milligan i Cooper, 1985). Altres aproximacions estimen el nombre de clusters a priori usant propietats matemàtiques del conjunt de documents, i llavors apliquen un algorisme de clustering iteratiu (Li et al., 2004). També existeixen aproximacions basades en l'ús d'un algorisme jeràrquic (per exemple, Hierarchical Agglomerative Clustering (HAC) (Murty i Krishna, 1980) i un criteri de selecció per a determinar un punt de tall al dendrograma (Tibshirani et al., 2001). Recentment, s'han dut a terme experiments amb mètodes híbrids, en què la sortida generada per un algorisme de clustering s'utilitza per a inicialitzar-ne un altre (Fraley i Raftery, 1998; Surdeanu et al., 2005).

Tanmateix, cada algorisme proposat té el seu biaix intrínsec i particular, usa una certa representació dels documents i depèn d'una mesura de similitud diferent. Totes aquestes assumpcions porten el procés de clustering a una solució particular, que no té per què ser l'òptima. Per a intentar superar les limitacions dels algorismes individuals, tant en el camp del reconeixement de patrons en general, com en el del clustering de documents en particular, s'han utilitzat bastament mètodes de combinació, també anomenats de consens.

Des d'un punt de vista general, els mètodes de clustering de consens cerquen, *"donats diversos clusterings d'un conjunt de dades, trobar un clustering combinat de millor qualitat"*[2] (Topchy et al., 2005). En el cas en què la combinació es fa en base només als clusterings, sense accedir a la (o les) representació original de les dades, és habitual referir-se al problema com a clustering conjunt[3] (Strehl i Ghosh, 2002).

[1] *"The process of partitioning a set of patterns into disjoint and homogeneous meaningful groups, called clusters."*
[2] *"Given multiple clusterings of the data set, find a combined clustering with better quality."*
[3] *ensemble clustering*

La major part del treball en clustering conjunt de documents s'ha centrat en aproximacions supervisades (Strehl i Ghosh, 2002; Sevillano et al., 2006; Greene i Cunningham, 2006). Tanmateix, recentment han aparegut també aproximacions no supervisades (Gionis et al., 2005).

La nostra recerca en temes de clustering s'ha centrat en el problema del clustering conjunt. Inicialment es va estudiar la representació de documents més adequada a la tasca, i es van considerar diverses estratègies de clustering individual no supervisat, per a posteriorment combinar els seus resultats. Aquesta combinació es va plantejar utilitzant mètodes de clustering conjunt tant supervisat, afegint mecanismes externs per a reduir-ne la supervisió i per a introduir-hi una ponderació dels clusterings, com no supervisat.

Les següents seccions descriuen cadascun d'aquests aspectes. La secció 2.1 presenta la formalització utilitzada del problema de clustering. A continuació, la secció 2.3 presenta cadascun dels mètodes considerats: individuals, de clustering conjunt ponderat i de clustering conjunt no supervisat (seccions 2.3.1, 2.3.2 i 2.3.3, respectivament). Per últim, la secció 2.4 presenta un resum dels experiments realitzats i dels resultats obtinguts amb cadascuna de les aproximacions considerades.

Tot aquest treball es descriu amb més detall a Gonzàlez i Turmo (2005, 2006); Gonzàlez (2007); Gonzàlez i Turmo (2008a,b).

2.1 Formalització

El punt de partida del problema del clustering és una col·lecció d'elements, en la major part dels nostres casos documents, $\mathcal{D} = \{d\}$. Un clustering Π és una partició del conjunt $\mathcal{D}$ en conjunts disjunts (o clústers) $\Pi = \{\pi_1 \dots \pi_k\}$. El clustering Π també es pot veure com una funció que s'aplica als documents d per a obtenir etiquetes $\{1 \dots k\}$, corresponents als clústers:

$$\begin{aligned} \Pi\colon \mathcal{D} &\rightarrow \{1 \dots k\} \\ \Pi(d) = m &\leftrightarrow d \in \pi_m \end{aligned}$$

Donats un conjunt de clusterings $\{\Pi_j\} = \{\Pi_1 \dots \Pi_r\}$, l'objectiu del clustering de consens és trobar un clustering $\bar{\Pi}$, amb $\bar{k}$ clusters, que sigui el consens dels r clusterings d'entrada, mitjançant una certa funció de consens Γ:

$$\bar{\Pi} = \Gamma(\mathcal{D}; \{\Pi_1 \dots \Pi_r\})$$

Si la funció Γ no utilitza la representació original dels documents $\mathcal{D}$, parlem de clustering conjunt:

$$\begin{aligned} \bar{\Pi} &= \Gamma(\mathcal{D}; \{\Pi_1 \dots \Pi_r\}) \\ &= \Gamma(\{\Pi_1 \dots \Pi_r\}) \end{aligned}$$

En el cas d'una combinació ponderada de clusterings, per a cada clustering inicial Π_j es defineix un per w_j, i el clustering resultant $\bar{\Pi}$ ha d'estar *més* d'acord amb els clusterings inicials de major pes. El grau d'acord entre $\bar{\Pi}$ i els $\{\Pi_j\}$ s'expressa mitjançant una puntuació σ.

2.2 Representació

Tal com s'ha mencionat a la introducció del capítol, com a pas previ a la recerca sobre clustering de documents, es van dur a terme una sèrie d'experiments per tal de comparar diverses representacions dels documents de cara al procés de clustering. Amb aquesta fi, es va escollir un algorisme de clustering individual supervisat, i es van avaluar els clusterings obtinguts amb les diverses representacions a estudiar, inicialitzant manualment l'algorisme per tal d'obtenir el millor resultat assolible a priori.

L'aproximació al clustering usada és una adaptació a clustering del model probabilístic generatiu per a classificació de documents de Nigam et al. (2000), que ja ha estat usat per a clustering en altres treballs, com ara Surdeanu et al. (2005).

El punt de partida del mètode és representar cada document d de la col·lecció de documents $\mathcal{D}$ com una seqüència d'esdeveniments $d = \{\eta_1 \dots \eta_{|d|}\}$. Cada esdeveniment η_j correspon a una

paraula, patró, etc. Els esdeveniments pertanyen a un vocabulari d'esdeveniments E que considerarem enumerat: $E = \{\bar{\eta}_1 \dots \bar{\eta}_{|E|}\}$.

El model generatiu considerat modela la col·lecció usant una mescla (de l'anglès *mixture*) de k components, cadascun d'ells corresponent a un clúster. Dins cada component, l'aparició d'un esdeveniment en la seqüència es modela usant una distribució multinomial, i cada aparició es considera independent de la resta donada la component (assumpció de *Naive Bayes*). El model resultant pot formular-se com:

$$\begin{aligned} p(d \mid \Theta) &= \sum_{m=1}^{k} \alpha_m \cdot p(d \mid \Theta_m) \\ p(d \mid \Theta_m) &= \prod_{j=1}^{|d|} p(\eta_j \mid \Theta_m) \\ p(\eta_j \mid \Theta_m) &= \prod_{e=1}^{|E|} \vartheta_{me}^{\delta(\eta_j, \bar{\eta}_e)} \end{aligned}$$

on la funció d'igualtat $\delta(x, x')$ val 1 quan $x = x'$ i 0 altrament. Addicionalment, cal que els paràmetres acompleixin les restriccions:

$$\begin{aligned} \sum_{m=1}^{k} \alpha_m &= 1 \\ \forall m \sum_{e=1}^{|E|} \vartheta_{me} &= 1 \end{aligned}$$

és a dir, les $\{\alpha_m\}$ i cada conjunt de $\{\vartheta_{me}\}$ han de pertànyer a un n-simplex.

Tant l'estimació de Màxima Versemblança com la de Màxim a Posteriori dels paràmetres $\widehat{\Theta} = \{\alpha_1, \widehat{\Theta}_1 \dots \alpha_m, \widehat{\Theta}_M\}$ poden obtenir-se utilitzant l'algorisme d'Expectation-Maximization (Dempster et al., 1977). Donat que, com hem comentat, els paràmetres del model es troben dins un n-simplex, com a distribució a priori es pot utilitzar una distribució de Dirichlet, assignant els paràmetres del prior tots iguals a 1. En aquest cas, l'estimació de Màxim a Posteriori és equivalent a una estimació amb suavitzat de Laplace (Manning i Schütze, 1999).

Per a obtenir el clustering, es busca el component m que té la màxima probabilitat d'haver generat cada document d en aquesta estimació $\widehat{\Theta}$:

$$\begin{aligned} \Pi(d) &= \arg\max_m \left[p(m \mid d, \widehat{\Theta}) \right] \\ &= \arg\max_m \left[\alpha_m \cdot p(d \mid \widehat{\Theta}_m) \right] \end{aligned}$$

Es consideren dos tipus de característiques dels documents com a esdeveniments en el model anterior:

Mots Els mots dels documents, eliminant xifres i *stop words* i ignorant la seva caixa, com a Surdeanu et al. (2005).

Patrons Els patrons que apareixen als documents. Es tracta de patrons sintàctics instanciats a partir d'un conjunt reduït de meta-patrons (del tipus mostrat a la figura 2.1), i que inclouen Entitats amb Nom dels tipus de les conferències MUC (Persona, Lloc, Organització i Altres), així com Dates i Quantitats, com a Surdeanu et al. (2006). Un exemple del procés d'instanciació es pot veure a la figura 2.2.

En tots dos casos, les característiques que només apareixen en un document s'eliminen.

sv	:	Subjecte	-	Verb				
svo	:	Subjecte	-	Verb	-	Objecte		
svoc	:	Subjecte	-	Verb	-	Objecte	-	Complement
svc	:	Subjecte	-	Verb			-	Complement
so	:	Subjecte			-	Objecte		
soc	:	Subjecte			-	Objecte	-	Complement
sc	:	Subjecte					-	Complement
vo	:			Verb	-	Objecte		
voc	:			Verb	-	Objecte	-	Complement
oc	:					Objecte	-	Complement
vc	:			Verb			-	Complement

Figura 2.1: Meta-patrons

Francesc Veguer va anotar 20 punts a La Gleva.									
sv	(	PERSONA	,	anotar					)
svo	(	PERSONA	,	anotar	,	punts			)
svoc	(	PERSONA	,	anotar	,	punts	,	a LLOC	)
svc	(	PERSONA	,	anotar	,			a LLOC	)
so	(	PERSONA	,			punts			)
soc	(	PERSONA	,			punts	,	a LLOC	)
sc	(	PERSONA	,					a LLOC	)
vo	(			anotar	,	punts			)
voc	(			anotar	,	punts	,	a LLOC	)
oc	(					punts	,	a LLOC	)
vc	(			anotar	,			a LLOC	)

Figura 2.2: Exemple d'instanciació dels meta-patrons en patrons

2.2.1 Experiments

Per tal d'estudiar l'efectivitat de les dues representacions **Mots** i **Patrons** proposades a l'hora de capturar l'estructura semàntica de col·leccions de documents, vam aplicar el model proposat a la secció anterior sobre diversos conjunts de dades, i vam avaluar la qualitat del clustering resultant comparant-lo amb una classificació dels documents en categories considerada *real*. La inicialització de l'algorisme d'Expectation-Maximization va ser manual, per a evitar que el procés d'inicialització influís en el resultat final.

2.2.1.1 Dades

Vam utilitzar un conjunt de 4 col·leccions de documents en llengua anglesa provinents de diverses fonts. En tots els casos es tracta de dades provinents del món real.

APW El subconjunt d'Associated Press (any 1999) de la col·lecció AQUAINT. Com a categoria vam utilitzar l'etiqueta CATEGORY de les pròpies dades.

LAT El subconjunt del Los Angeles Times de la col·lecció del TREC-5. Com a categoria vam agafar el departament del diari que va generar l'article, com a Zhao i Karypis (2004).

REU El subconjunt de 10 categories més freqüents de la col·lecció Reuters-21578. De forma similar al treball de Nigam et al. (2000) i Surdeanu et al. (2005), utilitzem la partició ModApte. Més concretament, usem la partició de test directament, ja que els algorismes a estudiar són no supervisats.

Col·lecció	Documents	Categories
APW	5000	11
LAT	5000	8
REU	2545	10
SMT	5467	4

Taula 2.1: Mida de les col·leccions

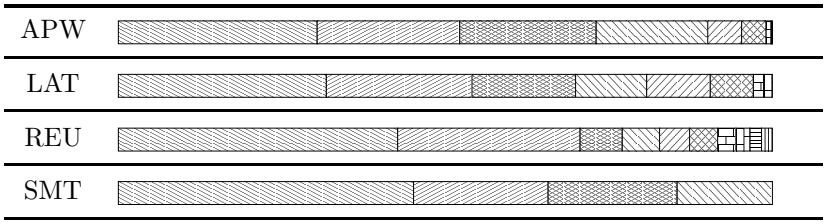


Figura 2.3: Distribució de les categories en les col·leccions

SMT Una col·lecció desenvolupada per a l'avaluació del sistema de recuperació d'informació SMART.

El nombre de documents i categories de cada col·lecció pot veure's a la taula 2.1, i la figura 2.3 conté una representació gràfica de la distribució de documents en categories dins cada col·lecció. Pot veure's com, especialment a **APW**, **LAT** i **REU**, existeix un conjunt de relativament poques categories que cobreix la gran majoria dels documents, mentre que altres categories tenen una presència força marginal.

2.2.1.2 Mètriques

Per a mesurar la qualitat del clustering resultant $\Pi = \{\pi_1 \dots \pi_k\}$, es compara el clustering amb la classificació considerada *real* dels documents en categories $\Lambda = \{\lambda_1 \dots \lambda_q\}$. Les mètriques d'avaluació utilitzades són:

Puresa (Pur) Avalua el grau en què cada clúster π_m conté documents d'un única categoria λ_c. La puresa d'un clúster és la fracció de la mida del clúster que representa la seva categoria majoritària (Zhao i Karypis, 2004). La puresa del clustering és la mitja ponderada de la puresa de cada clúster:

$$\begin{aligned} Pur(\pi_m, \Lambda) &= \frac{\max_{\lambda_c \in \Lambda} |\pi_m \cap \lambda_c|}{|\pi_m|} \\ Pur(\Pi, \Lambda) &= \frac{\sum_{m=1}^{k} |\pi_m| \cdot Pur(\pi_m, \Lambda)}{\sum_{m=1}^{k} |\pi_m|} \end{aligned}$$

Puresa Inversa (IPur) Avalua el grau en què els documents de cada categoria estan agrupats en un sol clúster. La puresa inversa d'una categoria λ_c és la fracció de la mida de la categoria que representa el clúster amb més documents d'aquesta categoria. La puresa inversa del clustering és la mitja ponderada de la puresa inversa de cada categoria:

$$\begin{aligned} IPur(\Pi, \lambda_c) &= \frac{\max_{\pi_m \in \Pi} |\pi_m \cap \lambda_c|}{|\lambda_c|} \\ IPur(\Pi, \Lambda) &= \frac{\sum_{c=1}^{q} |\lambda_c| \cdot IPur(\Pi, \lambda_c)}{\sum_{c=1}^{q} |\lambda_c|} \end{aligned}$$

	Mots			Patrons		
Col	Pur	IPur	F1	Pur	IPur	F1
APW	0.732	0.763	0.747	0.583	0.593	0.588
LAT	0.786	0.832	0.809	0.584	0.660	0.620
REU	0.804	0.833	0.818	0.690	0.705	0.697
SMT	0.934	0.934	0.934	0.613	0.617	0.615

Taula 2.2: Resultats del clustering de documents amb el model probabilístic

svo	(	PERSONA	,	anotar	,	punts		)
svo	(	jugador	,	anotar	,	punts		)
svo	(	PERSONA	,	anotar	,	punts		)
svo	(	PERSONA	,	aconseguir	,	punts		)
svo	(	PERSONA	,	anotar	,	punts		)
svoc	(	PERSONA	,	anotar	,	punts	, a LLOC	)

Figura 2.4: Exemples de parells de patrons similars

F1 És la mitja harmònica de puresa i puresa inversa.

$$F_1(\Pi, \Lambda) = \frac{2 \cdot Pur(\Pi, \Lambda) \cdot IPur(\Pi, \Lambda)}{Pur(\Pi, \Lambda) + IPur(\Pi, \Lambda)}$$

2.2.1.3 Resultats

Els resultats obtinguts es poden veure a la taula 2.2.

A la taula es pot observar clarament com, mentre que el clustering usant la representació **Mots** és capaç de capturar l'estructura de les col·leccions, amb valors de F1 per sobre de 0.80 en gairebé tots els casos, i fins i tot per sobre de 0.90 en la col·lecció **SMT**; el clustering usant la representació **Patrons** condueix a clusterings de qualitat inferior en totes les mètriques.

2.2.2 Altres Direccions

A la llum dels resultats de la secció anterior, es pot creure que una raó d'aquest comportament rau en el fet que el model proposat fa decisions polars respecte a la similitud d'esdeveniments: o són iguals o són diferents. En el cas de la representació **Mots** això no degrada significativament el rendiment, especialment en una llengua de morfologia simple com l'anglès. Els esdeveniment de la representació en **Patrons** porten molta més informació semàntica, però si el model no és capaç de detectar un cert grau de similitud entre parells de patrons com els de la figura 2.4, es produeix una gran pèrdua d'informació, i en resulta la incapacitat de detectar relacions entre documents que parlen del mateix tema.

Resulta doncs lògic de creure que millorant la funció de similitud entre patrons, aquests resultats inicials podien millorar. Amb aquesta idea vam dur a terme els diferents experiments que van tenir lloc durant l'estada que vaig fer a la New York University durant els mesos d'Abril a Juny del 2007, sota la direcció del Doctor Satoshi Sekine, i amb el suport d'una beca BE de la Generalitat de Catalunya. Tanmateix, els resultats d'aquests experiments van ser inconcloents, i la nostra línia de treball principal va continuar utilitzant paraules com a representació dels documents a clusteritzar. En qualsevol cas, referim a la memòria d'aquesta estada (Gonzàlez, 2008) per a més detalls.

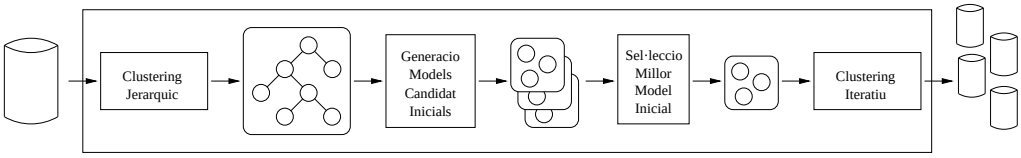


Figura 2.5: Mètode híbrid geomètric

2.3 Mètodes

2.3.1 Mètodes Individuals

Tal com també s'ha mencionat a la introducció del capítol, els biaixos particulars dels mètodes individuals de clustering, així com el tipus de representació dels documents i la mesura de similitud que aquests mètodes utilitzen, impliquen un punt de vista diferent sobre els conjunts de dades a clusteritzar. En la nostra recerca, hem treballat amb un conjunt heterogeni de mètodes individuals de clustering no supervisat, que procedirem a detallar.

El primer d'ells és el mètode híbrid geomètric de Surdeanu et al. (2005), de què s'ha demostrat que obté bons resultats en clustering no supervisat de documents de diferents col·leccions del món real. El segon és un mètode híbrid basat en teoria d'informació. Es tracta d'una versió del mètode anterior amb un biaix, una representació de documents i una mesura de similitud diferents. El tercer és un mètode clàssic basat en un algorisme jeràrquic i una funció de criteri per a determinar el millor punt de tall del dendrograma. A continuació es dóna una descripció detallada de cadascun d'ells.

2.3.1.1 Mètode Híbrid Geomètric

A la figura 2.5 es pot veure un esquema del mètode presentat a Surdeanu et al. (2005).

El procés té com a objectiu trobar un *bon* clustering inicial per a un algorisme de refinament iteratiu. Es requereix doncs trobar una *bona* estimació del nombre de clusters, així com un *bon* conjunt de documents *llavor*. Aquest clustering inicial es troba aplicant una variació dels mètodes clàssics:

1. S'utilitza un algorisme jeràrquic per a obtenir un dendrograma, o representació jeràrquica de la subsumpció dels clusters a la col·lecció (Zhao i Karypis, 2002).

2. Es genera un conjunt de clusterings candidats mitjançant un procediment consistent en ordenar els nodes del dendrograma utilitzant una determinada mesura de qualitat de cluster, i llavors agafar el major nombre de nodes millor puntuats de forma que la fracció de documents de la col·lecció que contenen no superi un determinat nivell llindar. Utilitzant diverses mesures de qualitat i nivells llindar es genera un espectre de candidats (els detalls sobre les mesures de qualitat es poden consultar a l'article original, però en tots els casos es tracta de quocients entre mesures de distància inter-cluster i intra-cluster).

3. Es tria el millor candidat utilitzant una mesura de qualitat global de clustering.

Encara que aquest esquema admet diversos algorismes i mesures, a Surdeanu et al. (2005), el mètode s'instancia utilitzant un punt de vista geomètric:

- Els documents es representen usant vectors de $tf \cdot idf$ de paraules (Spärck-Jones, 1972).
- L'algorisme jeràrquic utilitzat és el HAC.
- La funció de qualitat global és la C de Calinski i Harabasz (1974).

- La similitud entre documents usada en el clustering jeràrquic, en les mesures de qualitat de cluster i en la funció de qualitat global es basa en la distància de cosinus.
- L'algorisme de refinament iteratiu aplicat és l'EM de Nigam et al. (2000), ja mencionat a la secció 2.2.

Ens referirem d'ara en avant a aquest mètode com a **Geo**.

2.3.1.2 Mètode Híbrid Basat en Teoria de la Informació

El camp de la Teoria de la Informació es remunta al treball seminal de Shannon (1948). Recentment, hi ha hagut un interès en aplicar mesures de Teoria de la Informació a la tasca de clustering de documents (Dhillon i Guan, 2003; Slonim, 2003). Per aquesta raó, i per a obtenir una vista de les dades diferent de la de **Geo**, s'ha instanciat el mencionat mètode híbrid usant algorismes i mesures de Teoria de la Informació:

- Els documents es representen com a distribucions de probabilitat de paraules.
- L'algorisme jeràrquic utilitzat és l'Agglomerative Information Bottleneck (aIB) (Slonim i Tishby, 1999).
- La funció de qualitat global és una funció basada en longituds de missatges, descrita més avall.
- La distància entre documents usada en les mesures de qualitat de cluster es basa en la divergència de Jensen-Shannon (Lin, 1991). Hi ha altres mesures que podrien ser útils en aquest context, com ara la divergència de Kullback-Leibler (Kullback i Leibler, 1951) o la informació mútua. Tanmateix, i a diferència de la divergència de Jensen-Shannon, no són simètriques i/o requereixen continuïtat absoluta d'una distribució respecte a l'altra.
- L'algorisme de refinament iteratiu aplicat és el Divisive Information Theoretical Clustering (DITC) (Dhillon i Guan, 2003).

Ens referirem a aquest mètode com a **IT**.

2.3.1.2.1 Criteri de Longitud de Missatge Els criteris de selecció basats en Teoria de la Informació clàssics, com ara el Minimum Description Length (Rissanen, 1978) o el Minimum Message Length (Boulton i Wallace, 1969) requereixen que el model defineixi una distribució de probabilitat, propietat que no tenen els dendrogrames. Tanmateix, intentant seguir els seus principis, vam dissenyar un criteri per a seleccionar el millor clustering usant codis, missatges i longituds.

La idea és usar la informació en un clustering Π per a enviar una col·lecció de documents $\mathcal{D}$ com un missatge. Inicialment s'envia el centroide de cada cluster usant un codi basat en el meta-centroide de la col·lecció (un primer missatge de longitud $L_C(\Pi)$), i llavors s'envia la distribució de mots dins cada document usant un codi basat en el centroide del cluster a què pertany el document (un segon missatge de longitud $L_D(\Pi)$). Usant les fórmules de Teoria de la Informació, la longitud total del missatge, $L(\Pi)$, és aproximadament:

$$\begin{aligned} L(\Pi) &\approx L_C(\Pi) + L_D(\Pi) \\ L_C(\Pi) &\approx -\sum_{\substack{\pi_i \in \Pi \\ w}} p(w|c_i) \cdot \log p(w|mc) \\ L_D(\Pi) &\approx -\sum_{\substack{\pi_i \in \Pi \\ d_l \in \pi_i \\ w}} p(w|d_l) \cdot \log p(w|c_i) \end{aligned}$$

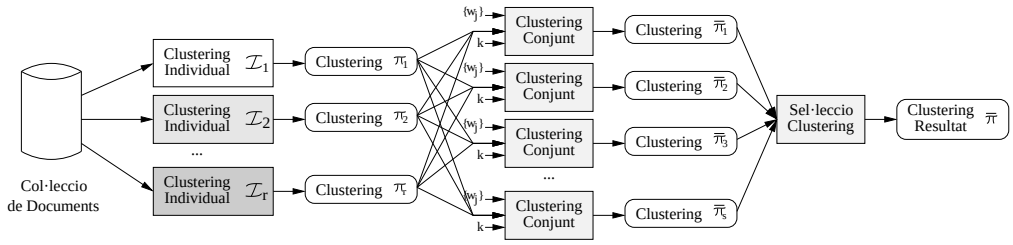


Figura 2.6: Clustering conjunt ponderat

on les w són les mots, les c_i és el centroides del cluster π_i, i mc és el meta-centroide.

S'espera que els *millors* clusterings (és a dir, més adequats a les dades) permetin una millor compressió de de les dades i, per tant, missatges més curts. Per tant, com a criteri prenem el clustering Π que tingui la $L(\Pi)$ menor, esperant que sigui el *millor*.

Aquesta fórmula va ser la que va donar millors resultats en experiments previs, comparada a una versió de la C de Calinski i Harabasz usant la divergència de Jensen-Shannon. Addicionalment, és atractiva ja que inclou una mesura implícita de la bondat del nombre de clusters (un major nombre de clusters implica una major $L_C(\Pi)$ però una menor $L_D(\Pi)$ i viceversa, de manera que es penalitzen tant les sub-estimacions com les sobre-estimacions d'aquest nombre).

2.3.1.3 Mètode Jeràrquic

El tercer mètode de clustering individual considerat és un mètode clàssic basat en un algorisme jeràrquic. S'obté un dendrograma usant l'algorisme aIB, i s'hi cerca el primer màxim local de la C de Calinski i Harabasz en direcció creixent del nombre de clusters. El punt trobat es pren com a nivell de dendrograma on s'espera trobar el millor clustering.

Ens referirem a aquest mètode com a **Hi**.

2.3.2 Clustering Conjunt Ponderat

La primera aproximació al problema del clustering conjunt que vam considerar es basa en mètodes de clustering conjunt supervisat, i en la incorporació d'un procediment de cerca heurística per a la determinació del nombre de clusters, $\bar{k}$, del clustering resultat.

Per altra banda, la majoria d'aproximacions al problema de clustering conjunt donen la mateixa rellevància a tots els clusterings individuals a combinar. Tanmateix, com ja s'ha mencionat, els diferents mètodes de clustering poden ser més o menys adequats a diverses col·leccions de documents. Pren sentit doncs considerar estratègies de clustering conjunt ponderat. Amb aquesta intenció es va incorporar ponderació a aproximacions de clustering conjunt supervisat ja existents, i es va incloure també la detecció de la ponderació òptima en l'estratègia de cerca.

L'aproximació proposada per a clustering conjunt ponderat es pot veure a la figura 2.6, i consta de tres passos:

1. Generar el conjunt de clusterings inicials, $\{\Pi_j\}$, cadascun amb un nombre de clusters k_j, aplicant diferents mètodes individuals de clustering, $\{\mathcal{I}_j\}$, a la col·lecció de documents d'entrada $\mathcal{D}$.

2. Generar clusterings conjunts ponderats, $\bar{\Pi}_\alpha$, a partir dels inicials, $\{\Pi_j\}$, usant mètode de clustering conjunt Γ amb diferents conjunts de pesos, $\{w_j\}_\alpha$, i nombres de clusters, $\bar{k}_\alpha$.

3. Escollir el millor clustering conjunt $\bar{\Pi}$ d'entre els generats al pas anterior.

Aquesta aproximació defineix una família d'esquemes de combinació ponderada. Un dels paràmetres és el mètode de clustering conjunt Γ. Les seccions 2.3.2.1 i 2.3.2.2 presenten els dos

mètodes que hem emprat ens els nostres experiments, basats en grafs i en un model probabilístic, respectivament.

Per altra banda, el conjunt de clusterings conjunts $\{\bar{\Pi}_\alpha\}$ és virtualment infinit. Seleccionar la millor combinació pot veure's com un problema de cerca.

Per a trobar la millor combinació d'entre tots els possibles clusterings conjunts ponderats, $\bar{\Pi}_\alpha$, cal determinar una funció de puntuació a ser maximitzada σ, que depèn del mètode de combinació.

Tanmateix, també és necessari de definir una estratègia per a explorar els possibles valors de $\bar{k}$ i w_j. L'aproximació seguida als nostres experiments ha estat la de cerca global restringida. Totes les combinacions amb $\bar{k}$ ¡ w_j tals que:

$$\begin{aligned} \bar{k} &\in \{\max(2, \lfloor \min k_j - \sigma_k \rfloor) \dots \min(|\mathcal{D}|, \lceil \max k_j + \sigma_k \rceil)\} \\ w_j &\in \{1 \dots G\} \end{aligned}$$

són trobades, essent k_j el nombre de clusters del clustering inicial π_j, σ_k la desviació estàndard de les k_j, i G un paràmetre fixat manualment. La combinació amb un valor de la funció η major és la que és seleccionada com a clustering conjunt resultat.

El nombre de clusters $\bar{k}$ que el clustering conjunt pot contenir, doncs, es troba limitat a valors en l'interval definit pel mínim i màxim nombre de clusters trobats pels mètodes individuals, extès per banda i banda per la desviació estàndard σ_k. Aquest nombre hauria de ser major o igual a 2, i menor o igual al nombre de documents a la col·lecció, $|\mathcal{D}|$. La justificació a aquest heurístic és que, si hi ha acord entre els diversos mètodes individuals, fet que implica que σ_k serà petit, el nombre de clusters òptim per a la $\bar{k}$ hauria de trobar-se prop de les k_j individuals, i no és necessari explorar valors llunyans. Per contra, si les k_j difereixen considerablement, fet que comporta una σ_k major, aquest nombre de clusters òptim per a $\bar{k}$ es podria trobar lluny, i té més sentit explorar una part major de l'espai de cerca.

El paràmetre G, per la seva banda, permet podar el nombre de combinacions de pesos limitant-los a valors naturals dins el rang $\{1 \dots G\}$. Quant major sigui el valor de G, més fi pot ser l'ajustament dels pesos. Tanmateix, aquesta aproximació té un cost computacional que creix exponencialment amb el nombre de clusterings a combinar.

2.3.2.1 Mètode Basat en Grafs

Strehl i Ghosh (2002) proposa diversos mètodes per a clustering conjunt basats en la resolució de problemes de partició de grafs. Addicionalment, ofereix un criteri per a seleccionar quin mètode escollir en cada cas, usant una versió normalitzada de la informació mútua.

Donats clusterings inicials $\{\Pi_j\}$, i el nombre de clusters desitjat $\bar{k}$, es resolen tres problemes de partició de grafs, per a obtenir tres combinacions possibles $\bar{\Pi}$:

CSPA (Cluster based Similarity Partitioning Algorithm) Es construeix un graf en què cada document $d_a \in \mathcal{D}$ és un vèrtex. El pes de l'aresta entre d_a i d_b és el nombre de clusterings Π_j en què els documents es troben al mateix cluster, $|\Pi_j \mid \Pi_j(d_a) = \Pi_j(d_b)|$. Es cerca la partició d'aquest graf en $\bar{k}$ parts, i el clustering conjunt és directament el que indueix aquesta partició sobre $\mathcal{D}$.

HGPA (HyperGraph Partitioning Algorithm) Es construeix un hipergraf en què cada document $d_a \in \mathcal{D}$ és un vèrtex. Cada cluster π_j^i en cada clustering inicial Π_j és una hiperaresta, i l'hipergraf es particiona en $\bar{k}$ parts, i el clustering conjunt és directament el que indueix aquesta partició sobre $\mathcal{D}$.

MCLA (Meta-CLustering Algorithm) Es construeix un graf en què cada cluster π_j^i en cada clustering inicial Π_j és un vèrtex. El pes de l'aresta entre els clusters π_j^i i $\pi_{j'}^{i'}$ és el coeficient de Jaccard dels dos conjunts: $|\pi_j^i \cap \pi_{j'}^{i'}| / |\pi_j^i \cup \pi_{j'}^{i'}|$. Aquest graf es particiona en $\bar{k}$ parts, anomenades meta-clusters γ_q. El clustering conjunt s'obté assignant cada document d_l al meta-cluster a què més contribueix, és a dir $\arg\max_{\gamma_q} |\pi_j^i \in \gamma_q \mid d_l \in \pi_j^i|$.

Per a decidir quin dels tres clustering conjunts $\bar{\Pi}$ és el més adequat, es defineix una mesura de informació mútua normalitzada (NMI) entre dos clusterings:

$$NMI(\Pi_j, \Pi_{j'}) = \frac{I(\Pi_j, \Pi_{j'})}{\sqrt{H(\Pi_j) \cdot H(\Pi_{j'})}} \tag{2.1}$$

on I i H són la informació mútua i l'entropia habituals, respectivament. Per a cada clustering conjunt, la seva informació mútua normalitzada mitja (ANMI) respecte als clusterings inicials $\{\Pi_j\}$ és:

$$ANMI(\bar{\Pi}, \{\Pi_j\}) = \frac{\sum NMI(\bar{\Pi}, \Pi_j)}{|\{\Pi_j\}|} \tag{2.2}$$

El clustering conjunt $\bar{\Pi}$ amb major ANMI se selecciona com a clustering conjunt resultat, i aquesta mateixa funció ANMI s'utilitza com a criteri σ per a determinar la millor $\bar{k}$ (i el millor vector de pesos $\{w_j\}$ en la versió ponderada).

Com Strehl i Ghosh (2002), hem usat els paquets lliurament disponibles[4] METIS i HMETIS de Karypis i Kumar (1998); Karypis et al. (1997) per a resoldre els problemes de partició de grafs i hipergrafs.

2.3.2.1.1 Versió Ponderada Per a incorporar ponderació a aquest mètode, es modifiquen els pesos de les arestes dels grafs a particionar en cada cas:

CSPA El pes de l'aresta entre els documents d_a i d_b és la suma dels pesos dels clusterings Π_j en què els documents són al mateix cluster, $\sum w_j \mid \Pi_j(d_a) = \Pi_j(d_b)$.

HGPA El pes de la hiperaresta que representa el cluster π_j^i és el pes w_j del clustering Π_j a què pertany el cluster.

MCLA El pes de l'aresta entre els clusters π_j^i i $\pi_{j'}^{i'}$ és el coeficient de Jaccard dels dos conjunts, multiplicat pel pesos dels clusterings Π_j i Π_p a què pertanyen els clusters: $|\pi_j^i \cap \pi_{j'}^{i'}|/|\pi_j^i \cup \pi_{j'}^{i'}| \cdot w_j \cdot w_{j'}$

En tots els casos, la versió no ponderada és equivalent a assignar un pes de $w_j = 1$ a tots els clusterings Π_j.

Ens referirem a les versions no ponderada i ponderada d'aquest mètode basat en grafs com a **Gr.Eq** i **Gr.P**.

2.3.2.2 Mètode Probabilístic

Topchy et al. (2005) introdueixen una visió probabilística del clustering conjunt, que es resol utilitzant EM.

Donats clusterings inicials $\{\Pi_j\}$, i el nombre de clusters desitjat $\bar{k}$, es pot definir una matriu Y amb tantes fileres com documents a la col·lecció, i tantes columnes com clusterings inicials. Cada posició y_{lj} correspon a l'etiqueta del cluster a què el document d_l pertany al clustering Π_j. Aquestes etiquetes es poden veure llavors com variables aleatòries generades per una distribució de probabilitat formada per una mescla de $\bar{k}$ components. En aquest cas, un document d_l es passa a representar mitjançant les seves etiquetes $y_l = (y_{l1} \dots y_{lr})$. Prenent l'assumpció de *Naive Bayes* d'independència entre etiquetes donada la classe, i cada etiqueta y_{lj} es considera generada per una distribució multinomial, la probabilitat de y_l és:

$$P(y_l \mid \Theta) \quad = \quad \sum_{m=1}^{\bar{k}} \alpha_m P(y_l \mid \Theta_m) \tag{2.3}$$

[4] `http://glaros.dtc.umn.edu/gkhome/views/metis/`

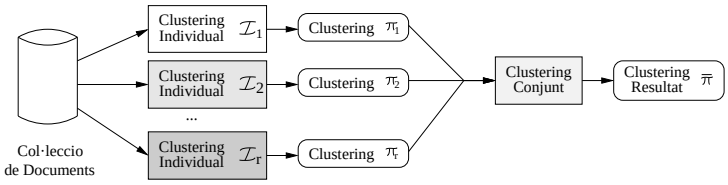


Figura 2.7: Clustering conjunt no supervisat

$$P(y_l \mid \Theta_m) = \prod_{j=1}^{r} P(y_{lj} \mid \Theta_{mj}) \tag{2.4}$$

$$P(y_{lj} \mid \Theta_{mj}) = \prod_{k=1}^{k_j} \vartheta_{mjk}{}^{\delta(y_{lj},k)} \tag{2.5}$$

Els paràmetres del model, Θ, es poden estimar usant EM, i el clustering conjunt $\bar{\Pi}$ s'obté assignant cada document al component més probable:

$$\bar{\Pi}(d_l) = \arg\max_m P(y_l \mid \Theta_m) \tag{2.6}$$

Com a funció σ per a determinar la millor k, s'utilitza la log-versemblança[5] del clustering donat el model probabilístic trobat:

$$LL(\bar{\Pi}) = \sum_{d_l} \log P(y_l \mid \Theta)$$

on P és la definida a l'equació 2.3.

2.3.2.2.1 Versió Ponderada L'extensió més natural a l'aproximació anterior es l'ús de models de *Naive Bayes* ponderats (Ferreira et al., 2001). Les equacions 2.3 i 2.5 romanen idèntiques, i els pesos s'introdueixen a l'equació 2.4 de la forma següent:

$$P(y_l \mid \Theta_m) = \prod_{j=1}^{r} P(y_{lj} \mid \theta_{mj})^{w'_j} \tag{2.7}$$

Aquests pesos w'_j han d'estar normalitzats de forma que la seva suma sigui igual al nombre de clusterings inicial r. El model modificat també s'ajusta usant l'algorisme d'EM.

Ens referirem a les versions no ponderada i ponderada d'aquest mètode probabilístic com a **Pr.Eq** i **Pr.P**.

2.3.3 Clustering Conjunt No Supervisat

El cost computacional de l'estratègia d'exploració de l'aproximació proposada de clustering conjunt ponderat, així com la recerca recent en clustering conjunt no supervisat (Gionis et al., 2005), van fer que optéssim per explorar també aquestes darreres tècniques no supervisades, en la tasca de clustering de documents.

Per als nostres experiments, en vam centrar en l'aproximació per a clustering conjunt no supervisat de Gionis et al. (2005), que incorpora mètodes per a la determinació del nombre de clusters. D'entre els mètodes proposats al treball, vam escollir l'algorisme **Agglomerative**, incorporant-hi **LocalSearch**. Aquesta combinació és la que donava millors resultats en experiments preliminars.

[5] *log-likelihood*

Totes els mètodes considerats a Gionis et al. (2005) comencen generant una matriu de correlació C_{ab} entre els elements de $\mathcal{D}$, considerant com a distància entre dos documents d_a i d_b la fracció del total de clusterings $\{\Pi_j\}$ en què es troben a clusters diferents:

$$C_{ab} = \frac{|\{\Pi_j \mid \Pi_j(d_a) \neq \Pi_j(d_b)\}|}{|\{\Pi_j\}|}$$

L'algorisme **Agglomerative** llavors aplica l'algorisme HAC per a trobar un dendrograma, que talla al punt on la distància entre els dos nodes a fusionar és major que 0.5. Posteriorment, el procés de **LocalSearch** aplica cerca local sobre el clustering resultant per a minimitzar la funció:

$$d(\bar{\Pi}) = \sum_{\substack{a,b \\ \bar{\Pi}(a)=\bar{\Pi}(b)}} C_{ab} + \sum_{\substack{a,b \\ \bar{\Pi}(a)\neq\bar{\Pi}(b)}} (1 - C_{ab})$$

Ens referirem al clustering resultant d'aplicar el mètode de clustering conjunt no supervisat sobre els tres clusterings generats amb els algorismes individuals de la secció 2.3.1 com a **Unsup**.

2.3.3.1 Combinació Massiva

Aprofitant el menor cost computacional que té l'aproximació no supervisada respecte a l'aproximació ponderada de la secció 2.3.2, i tenint en compte que hi ha hagut interès en la recerca en clustering conjunt de diverses execucions d'un únic algorisme inicialitzat a l'atzar (Topchy et al., 2005), es va voler estudiar una estratègia de generació de clusterings individuals aleatòria i massiva.

El procediment de generació és el següent:

- Manualment s'escull un nombre de clusterings r.
- Per a obtenir cadascun dels clusterings Π_j:
  - un nombre de clusters k_j s'escull a l'atzar, seguint una distribució uniforme entre 2 i una certa k_{max}.
  - S'escullen k_j documents a l'atzar de la col·lecció com a centroides inicials de k_j clusters.
  - S'aplica un algorisme de clustering iteratiu per a obtenir el clustering resultant Π_j.

En concret, com a algorisme de clustering iteratiu es va utilitzar de nou l'EM de Nigam et al. (2000), de la secció 2.2.

Pel que fa als valors de r i k_{max}, es va experimentar amb diversos valors, i es va trobar que, a les col·leccions sobre què es van dur a terme els experiments, el millor rendiment s'obtenia amb $r = 50$ i $k_{max} = 10$.

Ens referirem al clustering resultant d'aplicar el mètode de clustering conjunt no supervisat sobre els clusterings generats mitjançant aquesta estratègia massiva com a **UMass**.

2.4 Experiments

Amb la intenció d'avaluar els mètodes i estratègies per a clustering proposats a la secció anterior, es van dur a terme un conjunt d'experiments usant les mateixes dades i mètriques que a la secció 2.2.1. Addicionalment, es consideren el nombre de clusters k detectat, i el nombre de clusters *rellevants* k_r, on per *rellevant* entenem els clusters que representen més del 5% de la col·lecció. La raó per a incloure aquesta mètrica rau en el fet que alguns dels mètodes considerats tendeixen a detectar petits clusters, possiblement *outliers*, sense que aquests petits clusters hagin de veure's com un error en la detecció del nombre de clusters total.

Per altra banda, degut a que el mètode **UMass** conté una component aleatòria, els seus resultats són la mitja de 5 execucions.

2.4.1 Resultats

Els resultats dels experiments poden veure's a la taula 2.3.

Comparant els mètodes individuals entre ells, es pot veure com, d'una banda, els mètodes **Geo** i **IT** acostumen a trobar solucions amb més puresa que puresa inversa, mentre que **Hi** acostuma a proporcionar més puresa inversa que puresa. Globalment, els millors resultats s'obtenen amb **Geo** a totes les col·leccions, seguit de **Hi** i, per últim, **IT**, en alguns casos força per sota (col·leccions **APW** i **SMA**).

Pel que fa a les combinacions ponderades, així com els resultats dels mètodes basats en grafs **Gr.Eq** i **Gr.P** són propers, si no millors, als obtinguts pels mètodes individuals, els mètodes probabilístics **Pr.Eq** i **Pr.P** es comporten força pitjor, i en cap col·lecció no suposen una millora respecte a l'ús de mètodes individuals. Pel que fa a l'ús de ponderació o no, els resultats de **Gr.P** són sempre iguals als de **Gr.Eq** excepte a la col·lecció **SMA**, mentre que els resultats de **Pr.P** són fins i tot en el cas de **REU** pitjors que els de **Pr.Eq**. El fet que els resultats no siguin millors de forma concloent respecte a les versions no ponderades, a pesar de l'increment significant en el cost computacional, fan pensar que l'estratègia considerada per a clustering conjunt ponderat no és tan adient com semblava, i que calen millores per a poder-la aplicar de forma satisfactòria.

Per últim, dels mètodes de combinació no supervisada, **Unsup** obté a totes les col·leccions excepte **APW** iguals o millors resultats que els clusterings individuals, i a totes excepte a **LAT** iguals o millors resultats que la combinació basada en grafs **Gr.Eq**. Tanmateix, és el mètode massiu **UMass** el que destaca en aquesta avaluació, obtenint els millors resultats en totes les col·leccions. El problema d'aquest mètode, tanmateix, rau en la seva dependència de l'ajustament dels mencionats paràmetres t i k_{max} (veure secció 2.3.3.1).

Col·lecció	Mètode	w_j	Pur	IPur	F_1	k	k_r
APW	Geo	-	0.78	0.73	0.75	10	9
	IT	-	0.72	0.56	0.63	8	8
	Hi	-	0.63	0.88	0.74	3	3
	Gr.Eq	-	0.71	0.73	0.72	7	6
	Gr.P	2+1+3	0.72	0.72	0.72	7	6
	Pr.Eq	-	0.73	0.64	0.68	11	8
	Pr.P	1+1+4	0.63	0.88	0.74	3	3
	Unsup	-	0.74	0.70	0.72	19	7
	UMass	-	0.80	0.70	0.75	60.6	7.0
LAT	Geo	-	0.78	0.59	0.67	14	9
	IT	-	0.75	0.61	0.67	7	7
	Hi	-	0.66	0.68	0.67	6	6
	Gr.Eq	-	0.75	0.68	0.72	7	6
	Gr.P	1+3+1	0.76	0.68	0.72	8	6
	EM.Eq	-	0.78	0.53	0.63	16	13
	EM.P	1+4+1	0.75	0.61	0.67	7	7
	Unsup	-	0.79	0.53	0.67	14	7
	UMass	-	0.73	0.53	0.75	27.2	4.8
REU	Geo	-	0.84	0.92	0.88	6	6
	IT	-	0.77	0.76	0.76	6	6
	Hi	-	0.73	0.86	0.79	4	4
	Gr.Eq	-	0.81	0.84	0.83	7	7
	Gr.P	4+3+4	0.82	0.85	0.83	7	7
	EM.Eq	-	0.82	0.82	0.82	8	8
	EM.P	1+1+4	0.73	0.86	0.79	4	4
	Unsup	-	0.85	0.89	0.88	13	6
	UMass	-	0.86	0.90	0.88	18.2	5.2
SMT	Geo	-	0.92	0.80	0.85	6	5
	IT	-	0.89	0.58	0.71	9	7
	Hi	-	0.71	0.97	0.82	3	3
	Gr.Eq	-	0.91	0.91	0.91	4	4
	Gr.P	1+3+3	0.92	0.91	0.92	5	4
	EM.Eq	-	0.91	0.68	0.78	11	9
	EM.P	1+1+4	0.71	0.97	0.82	3	3
	Unsup	-	0.93	0.90	0.91	18	4
	UMass	-	0.93	0.92	0.93	20.6	4.0

Taula 2.3: Resultats per al Clustering de Documents

Capítol 3

Adquisició de Patrons

Molts sistemes d'Extracció d'Informació tenen el seu coneixement codificat a mà per experts humans. Això implica que l'adaptació d'un sistema existent a un domini, estil o idioma diferent és costós en quant a esforç humà. Fins i tot si hi ha components independents del domini, el coneixement com les jerarquies de conceptes o els patrons d'extracció depenen enormement del domini. Aquesta adaptació doncs pot suposar un procés complet de re-enginyeria, habitualment per part d'algú amb un coneixement profund del sistema.

Amb la intenció de reduir aquest cost, i estimulada per l'èxit de les aproximacions basades en corpus en altres tasques de Processament del Llenguatge Natural (Young i Bloothoft, 1997; Manning i Schütze, 1999), des dels inicis dels 90 part de la recerca en Extracció d'Informació s'ha concentrat en l'aplicació de tècniques d'Aprenentatge Automàtic a la tasca. Estudis del ventall de tècniques emprades poden trobar-se a Cardie (1997); Yangarber i Grishman (2000); Turmo et al. (2006).

Dins les diverses tècniques emprades, en el context d'aprenentatge poc supervisat s'ha demostrat que les tècniques de bootstrapping (Yarowsky, 1995; Abney, 2004) són útils per a tasques de Processament del Llenguatge Natural. En concret, s'han mostrat competitives en tasques d'adquisició de patrons per a Extracció d'Informació (Riloff i Jones, 1999; Agichtein i Gravano, 2000; Yangarber et al., 2000; Yangarber, 2003; Stevenson i Greenwood, 2005; Surdeanu et al., 2006). Aquestes aproximacions es basen en l'ús d'un conjunt inicial d'exemples o de patrons *llavor*, a partir de què es poden aprendre condicions de context. Aquestes condicions permeten hipotetitzar nous exemples positius, que a l'hora permeten aprendre noves condicions de context, i així successivament.

Degut a la baixa supervisió que requereixen els mètodes de bootstrapping, el nostre mètode d'adquisició de patrons parteix d'un d'aquests mètodes i n'elimina la supervisió de *llavors*, reemplaçant-la per informació provinent del clustering de documents.

La secció 3.1 presenta amb més detall el mètode proposat. La següent secció 3.2 presenta els experiments que es van dur a terme per a validar-lo i avaluar-lo.

3.1 Mètode

L'aproximació emprada inicialment per a l'extracció de patrons és una adaptació del mètode de bootstrapping de Surdeanu et al. (2006). Un esquema d'aquest mètode es pot veure a la figura 3.1.

Es tracta d'un mètode basat en co-training Blum i Mitchell (1998) de dos classificadors:

- Un classificador basat en el model d'EM de Nigam et al. (2000), que usa les paraules en els documents de la col·lecció per a clusteritzar-los.
- Un classificador de llista de decisió, que adquireix patrons de forma incremental, seguint el formalisme basat en meta-patrons presentat a la secció 2.2, i assignant cada patró a un cluster segons una certa funció de criteri.

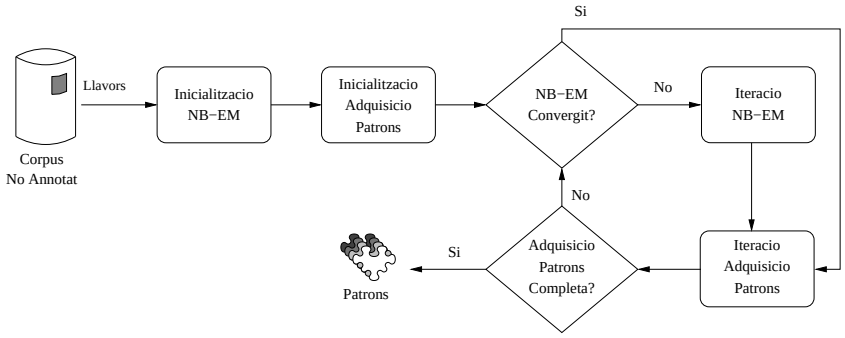


Figura 3.1: Mètode de Co-Training

El procés d'aprenentatge va alternant entre un classificador i l'altre de forma iterativa, utilitzant la sortida d'un com a entrada de l'altre i viceversa de forma iterativa. El punt de partida a l'article original de Surdeanu et al. (2006) és un conjunt de documents llavor.

Per tal d'incorporar clustering en aquest procés, utilitzem l'esquema seqüencial de la figura 1.3b. En concret, substituïm el conjunt de documents llavor per l'aplicació de tècniques de clustering. En concret, considerem els diversos clusterings obtinguts mitjançant els mètodes individuals (**Geo**, **IT** i **Hi**) i els mètodes de clustering conjunt supervisat (**Gr.Eq** i **Pr.Eq**). Addicionalment, comparem el resultat obtingut mitjançant les llavors manuals, **Manual**.

D'entre els diversos criteris d'adquisició de patrons considerats a l'article original, els nostres experiments utilitzen el de Collins i Singer (1999), que assigna com a puntuació d'un patró ξ en un cert cluster π, $\sigma(\xi, \pi)$, la seva freqüència al cluster, $freq(\xi, \pi)$, sempre i quant la fracció de documents en què apareix el patró i que pertanyen al cluster (precisió) sigui major que un cert valor llindar T:

$$\begin{aligned} \sigma(\xi,\pi) &= \begin{cases} freq(\xi,\pi) & prec(\xi,\pi) > T \\ 0 & \text{altrament} \end{cases} \\ prec(\xi,\pi) &= \frac{freq(\xi,\pi)}{freq(\xi)} \\ freq(\xi,\pi) &= |\{d \in \pi \mid \xi \in d\}| \\ freq(\xi) &= |\{d \in \mathcal{D} \mid \xi \in d\}| \end{aligned}$$

Per als nostres experiments, seguint el criteri de Surdeanu et al. (2006), hem usat com a valor llindar $T = 0.95$, i hem establert que a cada iteració el classificador basat en llistes de decisió adquireixi els 10 patrons millor puntuats per a cada cluster.

3.2 Experiments

Per tal d'avaluar els patrons obtinguts hem seguit l'estratègia d'avaluació indirecta ja usada per Yangarber (2003); Surdeanu et al. (2006). L'avaluació en aquest cas es du a terme mitjançant la tasca de classificació de documents. El classificador basat en llistes de decisió que conté el conjunt de patrons apresos s'utilitza per a classificar els documents d'una partició de test, i la qualitat de la classificació induïda s'usa com una indicació de la qualitat dels patrons.

Les dades usades són les mateixes col·leccions que als experiments de clustering de documents de les seccions 2.2.1 i 2.4.

3.2.1 Mètriques

Les mètriques usades per a l'avaluació de la tasca de classificació de documents són les de precisió i cobertura micro-promitjada [1]. Cada cluster π_i es mapeja a la categoria λ_j amb què té més solapament mitjançant una funció de mapeig ϕ:

$$\phi(\pi_i;\Lambda) = \lambda_j \longleftrightarrow \lambda_j = \arg\max_{\lambda_j \in \Lambda} |\pi_i \cap \lambda_j|$$

Els clusters mapejats a la categoria λ_j s'obtenen a través de la funció inversa ϕ^{-1}:

$$\phi^{-1}(\lambda_j;\Pi,\Lambda) = \{\pi_i \in \Pi \mid \phi(\pi_i;\Lambda) = \lambda_j\}$$

La precisió (Prec) i cobertura (Cob) micro-promitjada es defineixen llavors com:

$$Prec(\Pi,\Lambda) = \frac{\sum_{j=1}^{q} |\lambda_j \cap \phi^{-1}(\lambda_j;\Pi,\Lambda)|}{\sum_{j=1}^{q} |\phi^{-1}(\lambda_j;\Pi,\Lambda)|} \tag{3.1}$$

$$Cob(\Pi,\Lambda) = \frac{\sum_{j=1}^{q} |\lambda_j \cap \phi^{-1}(\lambda_j;\Pi,\Lambda)|}{\sum_{j=1}^{q} |\lambda_j|} \tag{3.2}$$

$$\tag{3.3}$$

3.2.2 Procediment

L'avaluació es va dur a terme usant una validació creuada en 5 parts. Cada col·lecció es va partir en 5 parts, i successivament els classificadors van ser entrenats en 4 d'aquestes parts i avaluats sobre la cinquena. Els resultats presentats són la mitja dels obtinguts en cadascuna de les 5 execucions.

Per altra banda, el rendiment del patrons s'avalua de forma incremental: les gràfiques mostren la precisió i cobertura usant els 100, 200, 300... patrons millor puntuats de cada cluster. A mida que el nombre de patrons creixi, s'espera que la cobertura s'incrementi i la precisió disminueixi.

3.2.3 Resultats

Els resultats obtinguts en cadascuna de les 4 col·leccions pels diversos mètodes poden veure's a les figures 3.2 i 3.3.

Es pot veure com, encara que les llavors **Manual** donen els millor resultats a totes les col·leccions excepte a **APW**, en molts casos el rendiment obtingut amb llavors automàtiques és comparable a l'obtingut amb **Manual**. El comportament dels mètodes de combinació **Gr.Eq** i **Pr.Eq** i de **IT** és en les quatre col·leccions només lleugerament inferior a **Manual**. Només a la col·lecció **SMA** les corbes estan més d'un 5% per sota de **Manual**, especialment en termes de precisió.

Geo proporciona bons resultats, de fet millors fins i tot que **Manual**, a **APW**. Se'n mantenen a prop a **LAT**, però a **REU** i **SMA** el seu rendiment és força més baix que **Manual** i que altres aproximacions automàtiques. Per últim, pot veure's com l'ús de les llavors proporcionades per **Hi** produeix un descens considerable en el rendiment del procés d'adquisició de patrons en totes les col·leccions.

Aquests resultats confirmen com, utilitzant una combinació seqüencial senzilla de clustering i adquisició de patrons, es poden aprendre patrons útils per a classificació de documents, i de forma competitiva amb aproximacions amb supervisió humana. Cap esperar que patrons d'aquest tipus puguin ser útils també per a Extracció d'Informació.

[1] *micro-averaged precision and recall*

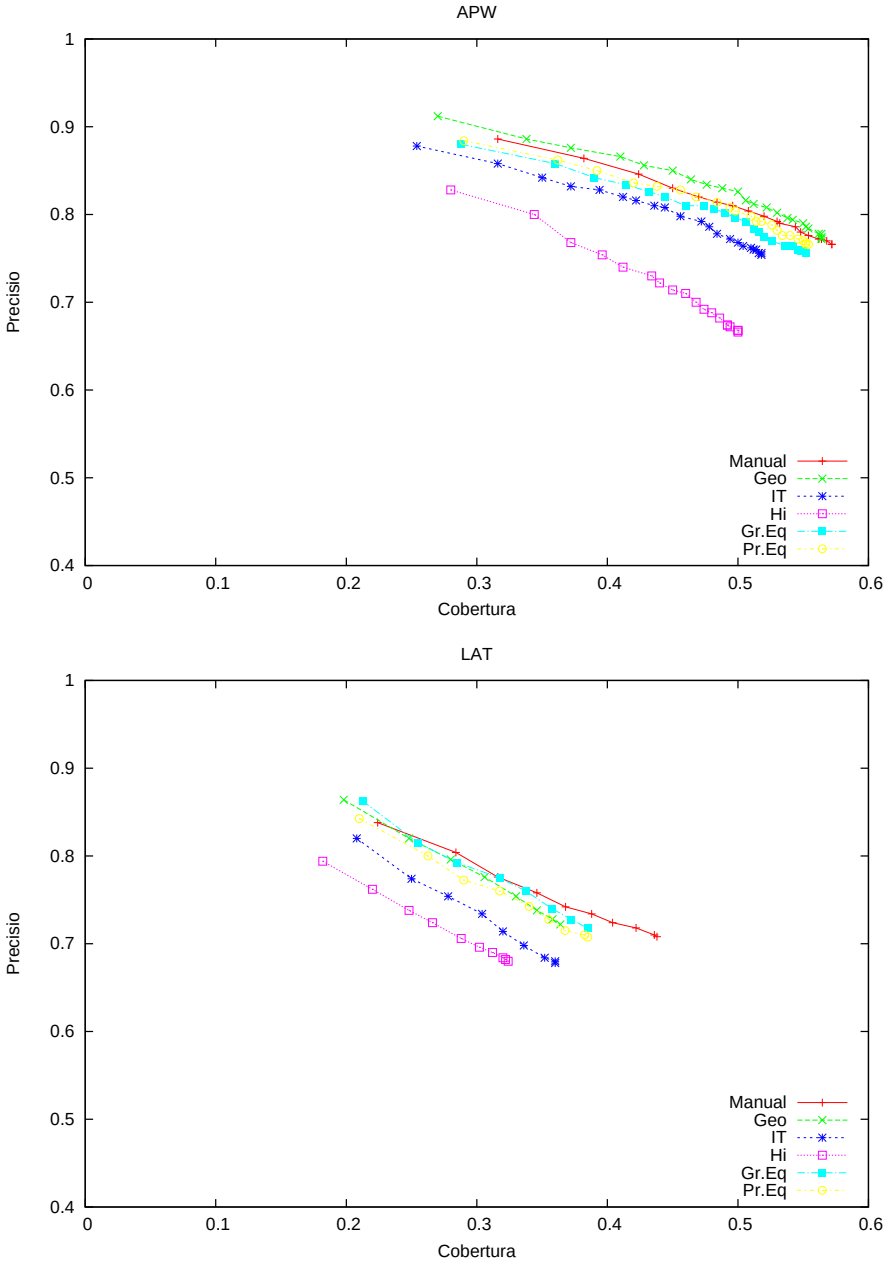


Figura 3.2: Resultats per a l'Adquisició de Patrons (I)

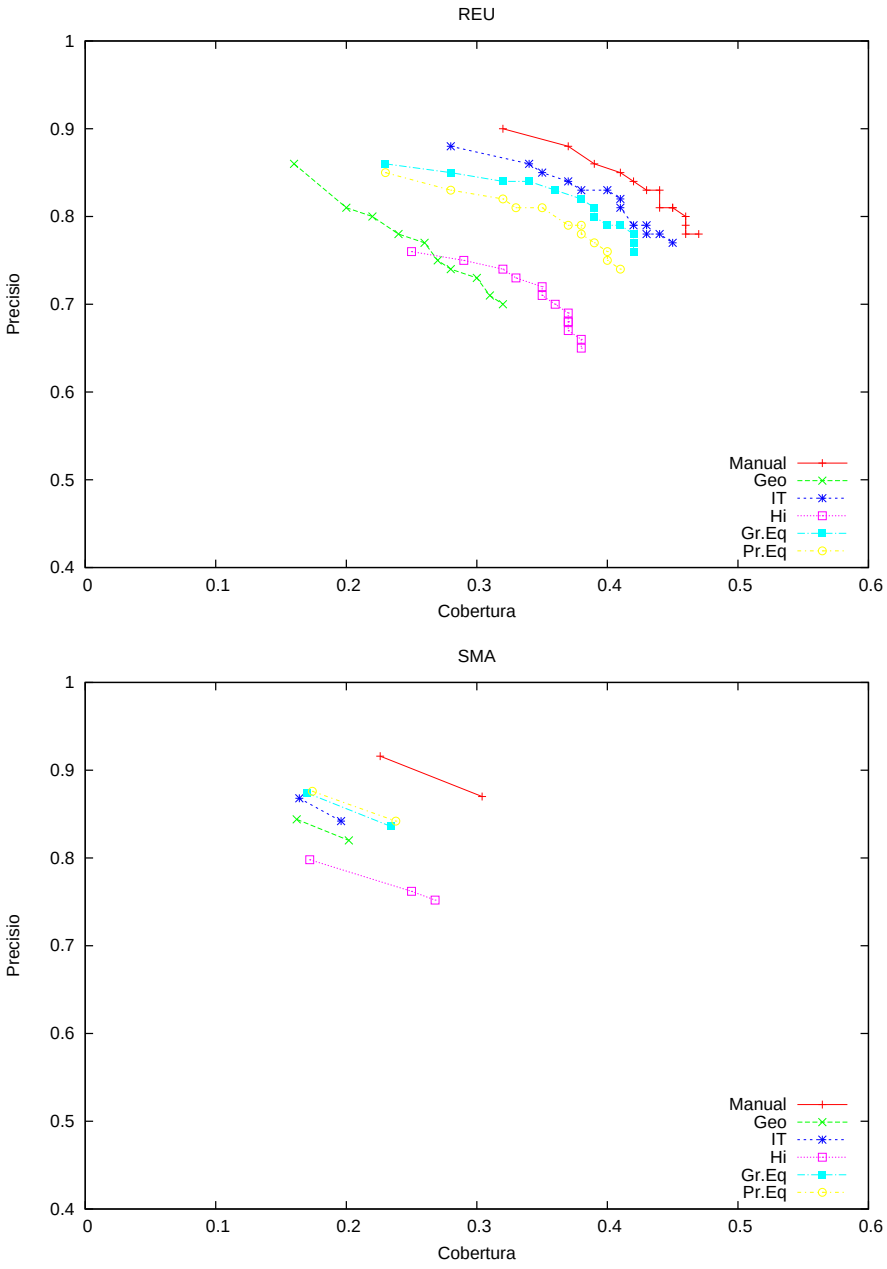


Figura 3.3: Resultats per a l'Adquisició de Patrons (i II)

Capítol 4

Conclusions

Hem realitzat experiments per a reduir la supervisió en processos d'obtenció de patrons per a Extracció d'Informació mitjançant tècniques de clustering.

Hem començat per estudiar la representació de documents més adequada per a la tasca de clustering. Hem considerat diversos mètodes individuals i no supervisat de clustering. Tanmateix, per tal d'evitar els seus biaixos, hem considerat mètodes de clustering conjunt. D'entre ells, hem explorat diversos mètodes de combinació supervisada, i hi hem afegit estratègies automàtiques per a determinar el nombre de clusters de la combinació. També hem considerat mecanismes per a obtenir clusterings conjunts ponderats, així com estratègies de combinació no supervisada. Finalment, hem utilitzat els resultats del clustering en un sistema d'adquisició de patrons per a substituir els elements de supervisió humana.

Hem avaluat totes aquestes estratègies i mètodes en tasques de clustering de documents i adquisició de patrons usant dades reals. I hem pogut comprovar que els mots com representació de documents superen altres models per a la tasca de clustering, així com que el clustering conjunt supera les limitacions dels clusterings individuals, i que les estratègies no supervisades d'adquisició de patrons obtenen resultats competitius respecte a les estratègies supervisades.

Tota aquesta recerca i els resultats que n'hem pogut extreure inviten a continuar la recerca en aquesta direcció per tal de concretar-la en una tesi doctoral en breu.

Agraïments

Gràcies al Grup de Processament del Llenguatge Natural del Departament de Llenguatges i Sistemes Informàtics de la Universitat Politècnica de Catalunya per haver-me donat la oportunitat de dur a terme la recerca de la meva tesi. En particular, gràcies al meu director Jordi Turmo i Borràs per la seva supervisió durant tots aquests anys[1].

Aquest treball ha estat realitzat amb el suport del Comissionat per a Universitats i Recerca del Departament d'Innovació, Universitats i Empresa de la Generalitat de Catalunya i del Fons Social Europeu.

[1] i els que queden

Bibliografia

S. Abney. Understanding the yarowsky algorithm. *Computational Linguistics*, 30(3), 2004.

E. Agichtein i L. Gravano. Snowball: Extracting relations from large plain-text collections. In *Proceedings of the 5th ACM Conference on Digital Libraries (DL)*, 2000.

A. Blum i T. Mitchell. Combining labeled and unlabeled data with co-training. In *Proceedings of the 11th Annual Conference on Computational Learning Theory (COLT)*, 1998.

D.M. Boulton i C.S. Wallace. The information content of a multistate distribution. *Journal of Theoretical Biology*, 23, 1969.

T. Calinski i J. Harabasz. A dendrite method for cluster analysis. *Communications in Statistics*, 3, 1974.

C. Cardie. Empirical methods in information extraction. *AI Magazine*, 18(4), 1997.

M. Collins i Y. Singer. Unsupervised methods for named entity classification. In *Proceedings of EMNLP/VLC-99*, 1999.

G.F. DeJong. Prediction and substantiation: A new approach to natural language processing. *Cognitive Science*, 3, 1979.

A.P. Dempster, N.M. Laird, i D.B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Royal Statistical Society, Series B*, 39(1), 1977.

I.S. Dhillon i Y. Guan. Information theoretic clustering of sparse co-occurrence data. In *Proceedings of IEEE International Conference on Data Mining*, 2003.

E. Dimitriadou. *Exploratory Data Analysis and Applications.* PhD thesis, Technische Universität Wien, 2003.

J.T.A.S. Ferreira, D.G.T. Denison, i D.J. Hand. Data mining with products of trees. In *Proceedings of the 4th International Conference on Advances in Intelligent Data Analysis (IDA)*, 2001.

C. Fraley i A.E. Raftery. How many clusters? which clustering method? answers via model-based cluster analysis. *The Computer Journal*, 41(8), 1998.

A. Gionis, H. Mannila, i P. Tsaparas. Clustering aggregation. In *Proc. of ICDE*, 2005.

E. Gonzàlez. Una aproximació d'aprenentatge automàtic per a extracció d'informació adaptativa. Master's thesis, Departament de Llenguatges i Sistèmes Informàtics, Universitat Politècnica de Catalunya, 2007.

E. Gonzàlez. Kernels semàntics per a clustering de patrons. Memòria Beca BE, 2008.

E. Gonzàlez i J. Turmo. Unsupervised clustering of spontaneous speech documents. In *Proceedings of the 9th European Conference on Speech Communication and Technology (EuroSpeech/InterSpeech)*, 2005.

E. Gonzàlez i J. Turmo. Unsupervised document clustering by weighted combination. Technical report, Departament de Llenguatges i Sistèmes Informàtics, Universitat Politècnica de Catalunya, 2006.

E. Gonzàlez i J. Turmo. Non-parametric document clustering by ensemble methods. *Procesamiento del Lenguaje Natural*, 40, 2008a.

E. Gonzàlez i J. Turmo. Comparing non-parametric ensemble methods for document clustering. In *Proceedings of the 13th International Conference on Applications of Natural Language to Information Systems (NLDB)*, 2008b.

D. Greene i P. Cunningham. Efficient ensemble methods for document clustering. Technical report, Department of Computer Science, Trinity College Dublin, 2006.

J.A. Hartigan. *Clustering Algorithms*. John Wiley & Sons, 1975.

J. Hobbs. The generic information extraction system. In *Proceedings of the 5th Message Understanding Conference (MUC)*, 1993.

A.K. Jain, M.N. Murty, i P.J. Flynn. Data clustering: A review. *ACM Computing Surveys*, 31(3), 1999.

G. Karypis i V. Kumar. A fast and high quality multilevel scheme for partitioning irregular graphs. *SIAM Journal on Scientific Computing*, 20(1), 1998.

G. Karypis, R. Aggarwal, V. Kumar, i S. Shekhar. Multilevel hypergraph partitioning: Applications in vlsi domain. In *Proceedings of the Design and Automation Conference*, 1997.

S. Kullback i R. A. Leibler. On information and sufficiency. *Annals of Mathematical Statistics*, 22 (1), 1951.

T. Li, S. Ma, i M. Ogihara. Document clustering via adaptive subspace iteration. In *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2004.

J. Lin. Divergence measures based on the shannon entropy. *IEEE Transactions on Information Theory*, 37(1), 1991.

C. Manning i H. Schütze. *Foundations of Statistical Natural Language Processing*. MIT Press, 1999.

J. McQueen. Some methods for classification and anlysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, 1967.

G.W. Milligan i M.C. Cooper. An examination of procedures for determining the number of clusters in a data set. *Psychometrica*, 50, 1985.

M.N. Murty i G. Krishna. A computationally efficient technique for data clustering. *Pattern Recognition*, 12, 1980.

K. Nigam, A. McCallum, S. Thrun, i T. Mitchell. Text classification from labeled and unlabeled documents using EM. *Machine Learning*, 39(2–3), 2000.

E. Riloff. Automatically constructing a dictionary for information extraction tasks. In *Proceedings of the 11th National Conference on Artificial Intelligence (AAAI)*, 1993.

E. Riloff i R. Jones. Learning dictionaries for information extraction by multi-level bootstrapping. In *Proceedings of the 16th National Conference on Artificial Intelligence (AAAI)*, 1999.

J.J. Rissanen. Modelling by shortest data description. *Automatica*, 14, 1978.

X. Sevillano, G. Cobo, F. Alías, i J.C. Socoró. Robust document clustering by exploiting feature diversity in cluster ensembles. *Procesamiento del Lenguaje Natural*, 37, 2006.

C.E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27, 1948.

N. Slonim. *The Information Bottleneck: Theory and Applications*. PhD thesis, The Hebrew University, 2003.

N. Slonim i N. Tishby. Agglomerative information bottleneck. In *Proceedings of Advances in Neural Information Processing Systems (NIPS-12)*, 1999.

K. Spärck-Jones. A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28, 1972.

M. Stevenson i M.A. Greenwood. A semantic approach to ie pattern induction. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL)*, 2005.

A. Strehl i J. Ghosh. Cluster ensembles - a knowledge reuse framework for combining multiple partitions. *Journal of Machine Learning Research*, 3, 2002.

M. Surdeanu, J. Turmo, i A. Ageno. A hybrid unsupervised approach for document clustering. In *Proceedings of the 11th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2005.

M. Surdeanu, J. Turmo, i A. Ageno. A hybrid approach for the acquisition of information extraction patterns. In *Proceedings of the EACL 2006 Workshop on Adaptive Text Extraction and Mining (ATEM)*, 2006.

D.K. Tasoulis i M.N. Vrahatis. Unsupervised distributed clustering. In M.H. Hamza, editor, *Proceedings of the IASTED International Conference on Parallel and Distributed Computing and Networks (PDCN)*, 2004.

R. Tibshirani, G. Walther, i T. Hastie. Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society, Series B*, 63(2), 2001.

A. Topchy, A.K. Jain, i W. Punch. Clustering ensembles: Models of consensus and weak partitions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(12), 2005.

J. Turmo, A. Ageno, i N. Català. Adaptive information extraction. *ACM Computing Surveys*, 38, 2006.

R. Xu i D. Wunsch, II. Survey of clustering algorithms. *IEEE Transactions on Neural Networks*, 16(3), 2005.

R. Yangarber. Counter-training in discovery of semantic patterns. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics (ACL)*, 2003.

R. Yangarber i R. Grishman. Issues in corpus-trained information extraction. In *Proceedings of the International Symposium on Spontaneous Speech: Toward the Realization of Spontaneous Speech Engineering*, 2000.

R. Yangarber, R. Grishman, P. Tapanainen, i S. Huttunen. Automatic acquisition of domain knowledge for information extraction. In *Proceedings of the Conference on Computational Linguistics (COLING)*, 2000.

D. Yarowsky. Unsupervised word sense disambiguation rivaling unsupervised methods. In *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics (ACL)*, 1995.

S. Young i B. Bloothoft, editors. *Corpus-Based Methods in Language and Speech Processing*. Kluwer Academic Press, 1997.

Y. Zhao i G. Karypis. Evaluation of hierarchical clustering algorithms for document datasets. In *Proceedings of ACM International Conference on Information and Knowledge Management*, 2002.

Y. Zhao i G. Karypis. Empirical and theoretical comparisons of selected criterion functions for document clustering. *Machine Learning*, 55(3), 2004.