Projecte Inspira'ns
Detecció d'Idees Relacionades

TALP Research Center
Universitat Politècnica de Catalunya

Índex

1 Introducció i objectius

El projecte Inspira'ns compta amb un sistema online per a la recollida de suggeriments i comentaris (*Idea*) d'usuaris. Aquests comentaris es classifiquen per categories (Aplicacions mòbils, Banca mòbil, Banca per Internet, Caixers i Targetes), i pels temes que tracten (el tema és lliure, generalment es tracta de peticions de millora d'un servei o problemes detectats amb algun producte o servei, entre d'altres). Sovint els usuaris es refereixen a una mateixa *Idea*, és a dir, es reben suggeriments i comentaris repetits. Actualment, la classificació d'aquestes idees es duu a terme de forma manual. L'objectiu d'aquest projecte és desenvolupar una metodologia que permeti detectar que una nova *Idea* és similar a una *Idea* anterior rebuda. És a dir, donat un conjunt D d'idees, determinar si una nova *Idea* d està relacionada (és similar) amb alguna de les idees del conjunt D.

Per dur a terme aquesta recerca, es disposa de dades anotades manualment. El conjunt de dades consta de 729 idees, classificades en 127 grups d'idees que estan relacionades entre elles. Cadascuna de les idees pertany a un únic grup d'idees, és a dir, estan relacionades amb només un grup d'idees. A més, cada grup s'identifica per una *Idea* de les que pertany al grup. El nombre d'elements de cada grup d'idees varia força. En concret, hi ha 65 grups que consten de només 2 idees, 24 grups que consten de 3 idees, 9 grups que consten de 4 idees i 7 grups de 5 idees. Respecte els grups més nombrosos, hi ha 8 grups amb un total d'idees entre 6 i 10; 5 grups consten d'entre 11 i 20 idees; 5 grups consten d'entre 21 i 30 idees, 3 grups consten d'entre 31 i 40 idees, i el grup més nombrós consta d'un total de 83 idees.

Aquest informe descriu el procediment que s'ha dut a terme per determinar si és possible identificar automàticament si dues idees estan relacionades. L'objectiu és establir una metodologia eficaç que porti a identificar aquestes idees. En resum, el procediment consisteix en (*i*) crear un corpus d'entrenament que contingui parelles d'idees i un conjunt d'atributs que representi la similitud entre dos elements d'una parella; (*ii*) entrenar un classificador binari que aprengui a detectar si dues idees estan relacionades o no ho estan; i (*iii*) analitzar els resultats.

La següent secció 2 descriu el procediment per crear un corpus de dades adequat per aquest anàlisi. La secció 3 i 4 descriuen, respectivament, el procés d'extracció d'atributs a partir del corpus generat i el procés de selecció d'atributs que representen les dades de forma més eficient (un conjunt d'atributs més petit). La secció 5 explica el procés d'aprenentatge automàtic d'un classificador binari que detecta parelles d'idees que estan relacionades i analitza els resultats obtinguts amb aquest classificador. Finalment, la secció 6 descriu una metodologia per obtenir un rànquing d'idees relacionades d'un conjunt donat i seleccionar el subconjunt d'aquelles que tenen més probabilitats de ser idees relacionades i analitza els resultats obtinguts.

2 Creació del corpus d'entrenament

L'objectiu d'aquesta fase és crear un corpus de parelles d'idees, tant si estan relacionades com si no ho estan, a partir de les dades proporcionades. Per a cadascuna de les parelles seleccionades, es calcula un conjunt d'atributs (Secció 3) i es construeix un corpus d'entrenament pel detector de parelles d'idees relacionades (Secció 5).

El nombre total de parelles que es poden obtenir a partir de les dades originals és de 531.441 (de les quals 16.084 són parelles relacionades). Aquest és un nombre molt elevat, tant pel temps que requereix calcular els atributs que representen a cada parella, com pel

temps que requereix entrenar el classificador. Per aquest motiu s'opta per seleccionar un nombre de parelles que representi aproximadament un 10% del total. El criteri de selecció de parelles es basa en trobar aquelles parelles d'idees no relacionades que siguin més semblants a les parelles relacionades, de forma que el detector de parelles relacionades sigui el més robust possible a l'hora d'aprendre a diferenciar-les. El procediment per obtenir aquestes parelles consisteix en crear un rànquing de similituds (basats en índexs de paraules i caràcters per accelerar el temps de càlcul) entre totes les parelles d'idees i seleccionar (*i*) totes les parelles relacionades d'acord a les dades originals anotades manualment, i (*ii*) un conjunt de parelles no relacionades però que tenen una similitud alta. És a dir, les parelles d'idees no relacionades que estan més a prop en el rànquing a les parelles relacionades.

El procediment detallat és el següent:

1. Es creen quatre índexs k, independents entre ells, que guarden les 729 idees del corpus d'acord a quatre caracteritzacions diferents:

   bag-of-words (BOW) : Cada *Idea* I_i es representa per un vector $v_{I_i}^{bow}$ amb el nombre d'ocurrències de cadascuna de les paraules.

   pseudo–cognats (COG) : Una paraula és un *p*seudo–cognat si (*i*) conté com a mínim un dígit, (*ii*) conté només lletres i la seva longitud és més gran o igual a 4, o (*iii*) és un signe de puntuació (el concepte *cognateness* es va proposar originalment a [SFI92]). Cada *Idea* I_i es representa amb un vector de cognats $v_{I_i}^{cog}$ que conté aquelles paraules de la *Idea* que són *p*seudo–cognats.

   3-grames de paraules (WNG) : Anomenem *n-grama* a una subseqüència de n elements consecutius d'una llista d'elements. En aquest cas, generem totes les subseqüències de 3 paraules consecutives del text de cada *Idea*. El vector de *3-grames* de paraules $v_{I_i}^{wng}$ representa a la *Idea* I_i.

   3-grames de caràcters (CNG) : Igual que en el cas anterior, generem totes les subseqüències de 3 caràcters presents en el text de cada *Idea*. El vector d'aquestes subseqüències $v_{I_i}^{cng}$ representa una *Idea* I_i.

2. A continuació, calculem la similitud de cada parella d'idees $s_{i,j}^k$ com el *cosinus* dels vectors que representen les idees dins l'index k: $s_{i,j}^k = cos(v_{I_i}^k, v_{I_j}^k)$. L'estimació final de la similitud entre dues idees s'obté com la mitjana de les similituds obtingudes amb cada índex: $s_{i,j} = \frac{s_{i,j}^{bow} + s_{i,j}^{cog} + s_{i,j}^{wng} + s_{i,j}^{cng}}{4}$ A partir d'aquestes similituds, es genera un rànquing[1] R_i d'idees I_{ij} de més a menys similars, per a cadascuna de les idees I_i.

3. Finalment, el corpus d'entrenament per al detector d'idees relacionades es composa de totes les parelles d'idees relacionades segons les dades originals (un total de 16.084 parelles) i aquelles parelles d'idees no relacionades que tenen una similitud per sobre d'un determinat *llindar* segons els rànquings (en total 42.087 parelles). Aquest llindar l s'ha calculat com:

$$l = \frac{\text{mitjana de la similitud de les parelles relacionades}}{2}$$

[1] Aquest rànquing es troba al fitxer `grup_idees.top2bot.txt`.

Index	Ratio_R'	Precisió_R'	Precisió_10	Cobertura_10
WNG	68.02%	0.10	0.76	0.29
COG	49.63%	0.17	0.74	0.40
BOW	60.66%	0.16	0.77	0.38
CNG	72.69%	0.05	0.79	0.22

Taula 1: Anàlisi dels rànquings generats a partir dels índexs WNG, COG, BOW i CNG.

És a dir, s'han seleccionat aquelles parelles no relacionades que tenen una similitud superior o igual a la meitat de la mitjana de la similituds de les parelles que estan relacionades. En total són 58.171 parelles d'idees no relacionades.

La taula 1 mostra un resum de l'anàlisi del contingut dels rànquings, on la primera columna indica el nom de l'índex.[2] Per a cada *Idea* I_i hem buscat dins el rànquing R_i totes les idees que pertanyen al mateix grup (és a dir, hi estan relacionades), i anotarem la posició p que ocupa l'última idea relacionada dins el rànquing. Aquest subconjunt d'idees $[I_{i1}, I_{ip}]$ del rànquing l'anomenem R'_i. Calculem el percentatge d'idees que representa el subconjunt R'_i respecte el nombre total d'idees (Ratio_R'), i calculem la precisió d'idees relacionades respecte al subconjunt d'idees R'_i. Després analitzem el contingut d'un subconjunt R'_i d'una mida *topN* determinada (enlloc de buscar la posició de l'última idea relacionada). La taula 1 indica els resultats obtinguts de precisió i cobertura per a *topN*=10 primeres idees del rànquing. La conclusió és que, tot i que algunes de les idees relacionades les podem trobar en posicions molt baixes del rànquing (tercera columna), la majoria de les idees relacionades les trobem a les posicions altes (quarta columna) i per tant els índexs són vàlids per ajudar-nos a fer la selecció d'idees que es necessita per construir el corpus d'entrenament del detector descrit a la secció 5.

3 Extracció d'atributs

El corpus de parelles d'idees es completa afegint un conjunt heterogeni d'atributs obtinguts a partir de diferents mesures de similitud. Per a obtenir aquests atributs hem usat l'eina Asiya [GM10]. Asiya és un programa amb llicència LGPL. Inicialment es va dissenyar per a l'avaluació de la traducció automàtica, és a dir, la comparació de dos textos. És per això que els seus mecanismes de càlcul de mètriques ens permeten usar aquest programa per a altres propòsits que requereixin la comparació de dos textos.

Asiya conté un nombre molt elevat de mesures (més de 500 per al castellà i català) basades en diferents principis de similitud (precisió, recall, overlap, etc.) i que actuen a diferents nivells lingüístics (lèxic, sintàctic i semàntic). Tot i que aquestes mesures són apropiades en la traducció automàtica, no totes elles són adequades pel nostre propòsit. Seleccionem un subconjunt de 46 mesures que es poden usar per comparar idees:

- **WER [NOLN00]**: Aquesta mesura es basa en la distància de Levenshtein [Lev66]. Calcula el nombre mínim de substitucions, eliminacions i insercions que s'han d'efectuar per convertir un text en un altre.

[2] Un anàlisi més detallat del rànquing de cada *Idea* es troba al fitxer `rank_analysis_statistics.txt`.

- **PER [TVN$^+$97]**: Una variant de la mesura WER que compara les paraules dels dos texts sense tenir en compte l'ordre de les paraules.
- **TERbase, TER [SMDS09], TERp, TERp-A**: TER mesura les post-edicions que s'han d'efectuar per convertir un text en un altre. A diferència de WER i PER, TER també compta els canvis de posició a dins de la seqüència de paraules. Es calculen 4 variants: TERbase només té en compte paraules, TER usa stemming i cerca de sinònims, TERp usa a més paràfrasi i TERpA és una mesura adaptada per mesurar la coincidència semàntica.
- **BLEU [PRWZ02]**: Mesura lèxica basada en el comptatge de *n-grames* de paraules per a $n \in [1, 4]$.
- **NIST [Dod02]**: Una variant millorada de BLEU per a *n-grames* $\in [1, 5]$.
- **GTM-3 [MGT03]**: Mesura GTM per a valor del paràmetre $e = 3$ (e és el paràmetre que controla el guany per a coincidències de seqüències llargues).
- **ROUGE-SU* [LO04]**: Una variant de la mesura ROUGE que ignora els *n-grames* sense max-gap-length, incloent els unigrames.
- **Pl, Rl, Fl, Ol**: Precisió, cobertura, F-measure i solapament del conjunt de paraules d'un text respecte un altre text.
- **ESA-es, ESA-lkxa-6, ESA-lkxa-7, ESA-lkxa-8, ESA-lkxa-9, ESA-lkxa-10, ESA-lkxa-11, ESA-lkxa-12, ESA-lkxa-13, ESA-lkxa-14, ESA-lkxa-15**: Explicit Semantic Analysis (ESA) compara dos texts a partir d'un vector de similituds entre cada text i un conjunt de documents. En el nostre cas, la col·lecció de documents consisteix en extractes de tots els articles de la Viquipèdia en castellà. Les variants *lkxa-X* es corresponen a subconjunts d'articles que tenen l'etiqueta "finances" i tots els articles relacionats a una distància menor o igual a X.
- **METEOR-ex, METEOR-pa, METEOR-st [DL14]**: 3 variants de la mesura METEOR, una que considera només la coincidència exacta de paraules, una altre que té en compte paràfrasi, i la última que té en compte l'arrel de les paraules.
- **NGRAM-cosChar2ngrams, NGRAM-cosChar3ngrams, NGRAM-cosChar4ngrams, NGRAM-cosChar5ngrams, NGRAM-cosTok2ngrams, NGRAM-cosTok3ngrams, NGRAM-cosTok4ngrams, NGRAM-cosTok5ngrams, NGRAM-jacChar2ngrams, NGRAM-jacChar3ngrams, NGRAM-jacChar4ngrams, NGRAM-jacChar5ngrams, NGRAM-jacTok2ngrams, NGRAM-jacTok3ngrams, NGRAM-jacTok4ngrams, NGRAM-jacTok5ngrams**: Mesures de similitud basades en el cosinus i el coeficient de Jaccard per a n-grames de tokens i caràcters ($n \in [2, 5]$).
- **NGRAM-jacCognates**: Mesura basada en *pseudo-cognats*.
- **NGRAM-lenratio**: Mesura basada en la proporció de la longitud dels texts.

En total tenim 50 atributs que representen cada parella d'idees: les 46 mesures calculades per Asiya i les 4 mesures proporcionades pels índexs creats en la secció anterior: BOW, COG,

WNG i CNG. Com a resultat després d'obtenir aquestes mesures, cadascuna de les parelles d'idees del corpus que es va construir a la secció 2 es representa per un vector de 50 atributs.

El procés d'obtenir tots aquests atributs és computacionalment costós, i a més no tenim la certesa que la combinació de tots ells sigui avantatjosa. El següent pas en aquest estudi és entrenar i obtenir un classificador que aprengui a diferenciar les idees relacionades de les que no ho estan. Per una banda, algun dels 50 atributs que hem calculat podria empitjorar el rendiment del nostre detector de parelles d'idees relacionades. A més, per altra banda, els temps necessari per entrenar el classificador augmenta també amb el nombre d'atributs que ha de processar. Per aquests motius, tal com s'explica a la següent secció, es realitza un estudi de selecció d'atributs. L'objectiu de trobar un subconjunt més petit que doni igual o fins i tot millor rendiment.

4 Selecció d'atributs

La fase de selecció d'atributs té per objectiu determinar quina combinació atributs és la que millor pot representar les dades del nostre corpus. La selecció d'atributs es fa en dues fases: primer es crea un rànquing d'atributs, i després s'avaluen iterativament diferents conjunts d'atributs.

En la primera fase s'avalua cadascun dels atributs per separat per estimar la seva habilitat predint la similitud entre dos texts i el grau de redundància entre els diferents atributs (informació mútua). A partir d'aquesta avaluació es genera un rànquing d'atributs.

En la segona fase, un algoritme de selecció d'atributs afegeix/elimina atributs en ordre, segons el rànquing, fins que troba el subconjunt que millor s'ajusta a les dades del corpus, i per tant les representa millor. L'algoritme dóna preferència a aquells subconjunts que presentin una correlació alta amb les dades i, a la vegada, tinguin un baixa inter–correlació. És a dir, que no siguin redundants entre ells [Hal98]. L'algoritme de cerca emprat és *Greedy Stepwise*[3], que realitza una cerca "forward insertion" i "backward elimination" a través de l'espai dels diferents subconjunts d'atributs. És a dir, pot començar amb cap atribut o tots a la vegada, segons si s'escull el mode "forward" o "backward", respectivament. A cada iteració s'afegeix o es treu un nou atribut i es torna a avaluar el subconjunt. Si el nou model resulta ser pitjor que el model anterior, es torna a provar canviant l'últim atribut afegit/eliminat per un altre atribut. La cerca s'atura quan afegir o eliminar qualsevol de la resta dels atributs empitjora la capacitat predictiva del subconjunt d'atributs. La llista ordenada d'atributs que es va generar en la primera fase s'usa per determinar l'ordre en què s'explora l'espai d'atributs.

Del conjunt inicial de 50 atributs, el procés de selecció d'atributs va seleccionar els següents 31: WNG, CNG, PER, TER, TERbase, TERp, TERp-A, WER, BLEU, NIST, GTM-3, ROUGE-SU*, Ol, Pl, Rl, Fl, ESA-es, ESA-lkxa-6, ESA-lkxa-7, ESA-lkxa-8, ESA-lkxa-9, ESA-lkxa-10, ESA-lkxa-11, ESA-lkxa-12, ESA-lkxa-13, ESA-lkxa-14, ESA-lkxa-15, METEOR-ex, METEOR-pa, METEOR-st, NGRAM-lenratio.

És a dir, va eliminar dos atributs relacionats amb els índexs de la Secció 2, i la col·lecció d'atributs NGRAM. Com a resultat de la selecció, és interessant observar que els dos mètodes ("backward" i "forward") van obtenir el mateix subconjunt d'atributs, el que suggereix la robustesa del subconjunt d'atributs seleccionats.

Per altra banda, a l'anàlisi del rànquing d'atributs de la primera fase es pot apreciar que tots els atributs de la col·lecció ESA tenen una capacitat predictiva similar (veure la taula 2).

[3] `http://weka.sourceforge.net/doc.dev/weka/attributeSelection/GreedyStepwise.html`

Aquests atributs són especialment costosos de calcular computacionalment, donat que usen col·leccions de documents molt grans. Amb l'objectiu de determinar si es pot mantenir el rendiment final del sistema millorant la eficiència, es va decidir crear un altre subconjunt més petit d'atributs que només inclou l'atribut ESA-lkxa-6 (el que conté la col·lecció més petita de documents).

5 Detector automàtic de parelles d'idees relacionades

En aquesta fase de l'estudi entrenem un classificador binari que aprengui a diferenciar (detectar) les parelles d'idees que estan relacionades (són similars) de les que no ho estan. El programa emprat és SVM^{light}, una implementació de Support Vector Machines [Joa02].[4] S'han realitzat els experiments amb kernels polinòmics de grau 1, 2 i 3, amb diferents valors de c (valor de trade-off entre l'error d'entrenament i l'hiperplà que separa les dades). En particular, el valor de la constant c s'ha optimitzat samplejant-la de forma logarítmica entre $1e-6$ i $1e+3$. Les figures dels experiments mostren algunes de les corves de precisió que s'aconsegueixen en aquest procés. Donat que per algunes configuracions els classificadors són molt sensibles al valor de c, es podria fer una exploració molt més detallada del marge i potser s'aconseguirien precisions lleugerament superiors.

L'experiment s'ha dut a terme amb els tres conjunts d'atributs descrits en les seccions anteriors:

- `all_features`: 50 atributs seleccionats de la secció 3.
- `selected_features`: 31 atributs seleccionats de la secció 4.
- `selected_features_min`: 20 atributs del conjunt `selected_features` que requereixen menys recursos i no són redundants.

L'objectiu és veure si amb un subconjunt més petits d'atributs que requereixin menys recursos i per tant siguin més ràpids d'obtenir i processar, aconseguim obtenir un model tan precís com el que conté tots els atributs.

Les dades d'entrenament final són 57.442 parelles amb els seus vectors d'atributs. Aquest nombre és lleugerament inferior que el corpus que es va construir a la secció 2 perquè s'han eliminat les parelles on es comparava una *Idea* amb ella mateixa. Per garantir que els resultats de l'experiment són independents de la partició de les dades d'entrenament i test, s'aplica a més una validació creuada de 10 iteracions.

Per altra banda, s'han construït uns corpus especialitzats per a cada categoria. L'objectiu d'aquest experiment és determinar si un classificador especialitzat per a cada categoria és més eficaç que un classificador general. Per aquest motiu, les proves de validació es realitzen sobre el conjunt de dades de test obtingut de la partició corresponent, i sobre el el subconjunt de parelles de test on el primer element de la parella (I_i) pertany a una categoria determinada.

En resum, totes les variables de l'experiment són:

D : El grau del polinomi del kernel de la SVM: 1, 2 i 3.

C : El valor del paràmetre c del classificador.

Categoria : Aplicacions mòbils, Banca mòbil, Banca per Internet, Caixers i Targetes.

[4] `http://svmlight.joachims.org/`

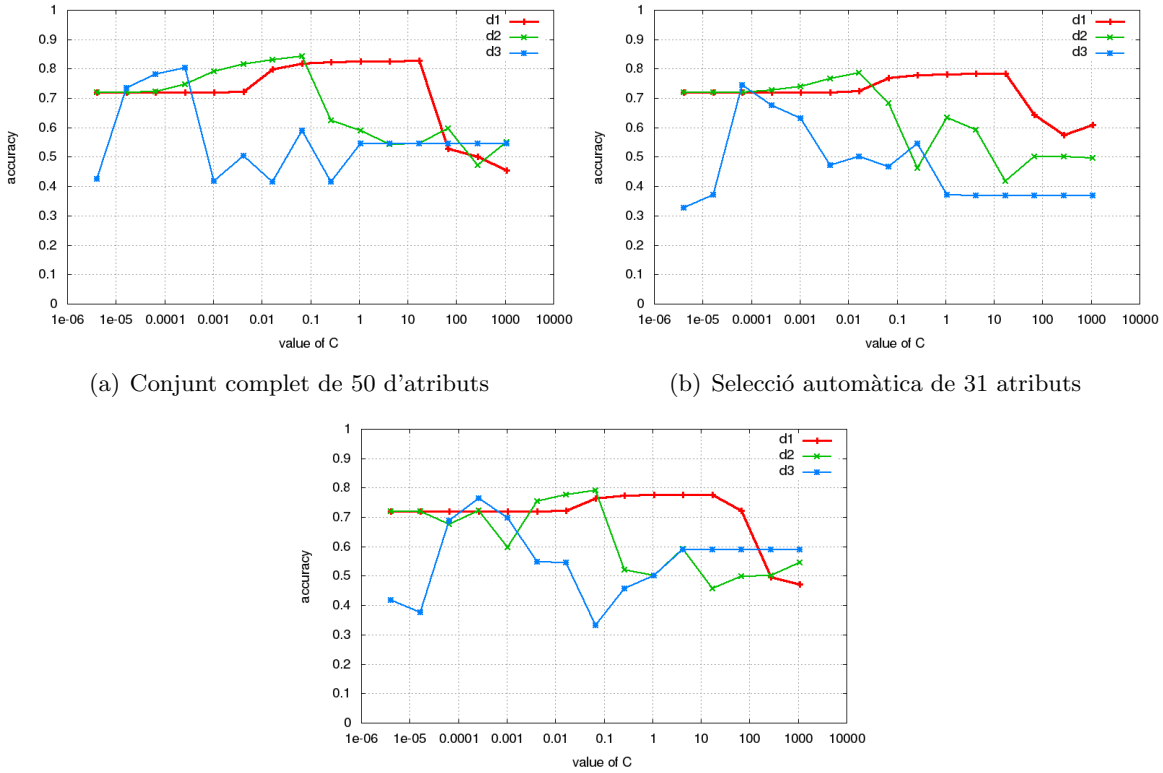
forward insertion		Backward elimination	
rank	*atribut*	*rank*	*atribut*
1.0	CNG	1.0	CNG
1.0	-PER	1.0	-PER
1.0	-TER	1.0	-TER
1.0	-TERbase	1.0	-TERbase
1.0	-TERp	1.0	-TERp
1.0	-TERp-A	1.0	-TERp-A
1.0	BLEU	1.0	BLEU
1.0	ESA-es	1.0	ESA-es
1.0	ESA-lkxa-6	1.0	ESA-lkxa-6
1.0	ESA-lkxa-7	1.0	ESA-lkxa-7
1.0	ESA-lkxa-8	1.0	ESA-lkxa-8
1.0	ESA-lkxa-9	1.0	ESA-lkxa-9
1.0	ESA-lkxa-10	1.0	ESA-lkxa-10
1.0	ESA-lkxa-11	1.0	ESA-lkxa-11
1.0	ESA-lkxa-12	1.0	ESA-lkxa-12
1.0	ESA-lkxa-13	1.0	ESA-lkxa-13
1.0	ESA-lkxa-14	1.0	ESA-lkxa-14
1.0	ESA-lkxa-15	1.0	ESA-lkxa-15
1.0	Fl	1.0	Fl
1.0	GTM-3	1.0	GTM-3
1.0	METEOR-ex	1.0	METEOR-ex
1.0	METEOR-pa	1.0	METEOR-pa
1.0	METEOR-st	1.0	METEOR-st
1.0	NGRAM-cosChar2ngrams	1.0	NGRAM-cosChar2ngrams
1.0	NIST	1.0	NIST
1.0	Ol	1.0	Ol
1.0	Pl	1.0	Pl
1.0	ROUGE-SU*	1.0	ROUGE-SU*
1.0	Rl	1.0	Rl
5.6E-12	BOW	1.0	BOW
2.8E-12	COG	5.6E-12	COG
1.8E-12	WNG	2.8E-12	NGRAM-cosChar3ngrams
1.4E-12	NGRAM-cosChar3ngrams	1.9E-12	-WER
1.1E-12	NGRAM-cosChar4ngrams	1.4E-12	NGRAM-cosChar4ngrams
9.3E-13	NGRAM-cosChar5ngrams	1.1E-12	NGRAM-cosChar5ngrams
8.0E-13	NGRAM-cosTok2ngrams	9.3E-13	NGRAM-cosTok2ngrams
7.0E-13	NGRAM-cosTok3ngrams	8.0E-13	NGRAM-cosTok3ngrams
6.2E-13	NGRAM-cosTok4ngrams	7.0E-13	NGRAM-cosTok4ngrams
5.6E-13	NGRAM-cosTok5ngrams	6.2E-13	NGRAM-cosTok5ngrams
5.1E-13	NGRAM-jacChar2ngrams	5.6E-13	NGRAM-jacChar2ngrams
4.7E-13	-WER	5.1E-13	NGRAM-jacChar3ngrams
4.3E-13	NGRAM-jacChar3ngrams	4.7E-13	NGRAM-jacChar4ngrams
4.0E-13	NGRAM-jacChar4ngrams	4.3E-13	NGRAM-jacChar5ngrams
3.7E-13	NGRAM-jacChar5ngrams	4.0E-13	NGRAM-jacCognates
3.5E-13	NGRAM-jacCognates	3.7E-13	NGRAM-jacTok2ngrams
3.3E-13	NGRAM-jacTok2ngrams	3.5E-13	NGRAM-jacTok3ngrams
3.1E-13	NGRAM-jacTok3ngrams	3.3E-13	NGRAM-jacTok4ngrams
2.9E-13	NGRAM-jacTok4ngrams	3.1E-13	WNG
2.8E-13	NGRAM-jacTok5ngrams	2.9E-13	NGRAM-jacTok5ngrams
2.7E-13	NGRAM-lenratio	2.8E-13	NGRAM-lenratio

Taula 2: Rànquing d'atributs segons els dos mètodes de selecció d'atributs.

Atributs : El conjunt d'atributs emprats.

Les gràfiques de la figura 1, juntament amb les que es mostren a l'apèndix A, són els resultats obtinguts amb els experiments descrits. La figura 1 mostra els resultats dels classificadors entrenats amb totes les dades d'entrenament (model general) i separats segons els conjunt d'atributs emprats. L'apèndix A mostra, a més, els resultats obtinguts amb les classificadors especialitzats per a cada categoria.

Les corbes de cada gràfica indiquen la precisió mitjana obtinguda de cada partició de l'avaluació creuada, i un valor de c determinat. Cadascuna de les corbes es correspon un grau del polinomi del kernel entrenat amb un conjunt concret d'atributs i dades.



(a) Conjunt complet de 50 d'atributs

(b) Selecció automàtica de 31 atributs

(c) Selecció mínima de 20 atributs lights

Figura 1: Resultats del detector de parelles d'idees relacionades obtinguts per als diferents conjunts d'atributs

En general, s'observa que el model que conté tots els atributs obté precisions més altes (84%) que els models que contenen subconjunts reduïts d'atributs (80% i 79% de precisió). El kernel polinòmic de grau 2 es comporta millor que la resta per un rang del paràmetre c entre $0,05$ i $0,1$. El kernel lineal (polinòmic de grau 1) obté un rendiment molt similar (83%, 80% i 79%, respectivament per a cada subconjunt d'atributs) per a rangs de c idèntics. S'ha de tenir en compte que aquest kernel és més ràpid d'entrenar i per tant pot ser més apropiat en determinats entorns d'integració.

Per categories, les idees que pertanyen al grup d'Aplicacions mòbils (83%) i Targetes (85%) es classifiquen millor que els altres tres tipus de categoria (entre el 71% i el 75% de precisió). Aquests valors són similars als obtinguts pels diferents conjunts d'atributs.

Finalment, els models entrenats per a cada categoria no han obtingut millors resultats que el model general, i és interessant notar que els resultats són encara pitjors quan s'usen aquests models per classificar les idees per categories. Per aquest motiu es desaconsella l'ús de classificadors especialitzats per a cada categoria.

La taula 3 mostra un resum dels valors màxims de precisió obtinguts amb les diferents variants de detector de parelles d'idees relacionades (segons el conjunt d'atributs i el grau del kernel). Es mostren només els resultats dels models entrenats amb totes les dades (no especialitzats per categories). Cada fila correspon a un conjunt de test on només se seleccionen les parelles per categories, i l'última fila correspon al resultat per al conjunt de test sencer. S'ha de tenir en compte que la última fila no es la mitjana de les files anteriors donat que no hi ha el mateix nombre d'elements per a cada categoria.

Conjunt d'atributs →	**Complet** (50 **atrs.**)			**Automàtic** (31 **atrs.**)			**Mínim** (20 **atrs.**)		
Grau del kernel →	*d1*	*d2*	*d3*	*d1*	*d2*	*d3*	*d1*	*d2*	*d3*
Aplicacions Mòbils	0.830	0.830	0.766	0.830	0.830	0.698	0.830	0.830	0.766
Banca Mòbil	0.713	0.713	0.672	0.713	0.713	0.626	0.713	0.713	0.672
Banca per Internet	0.708	0.708	0.666	0.708	0.708	0.625	0.708	0.708	0.667
Caixers	0.742	0.742	0.694	0.742	0.742	0.645	0.742	0.742	0.694
Targetes	0.844	0.844	0.774	0.844	0.844	0.705	0.844	0.844	0.776
Totes les dades	0.827	0.844	0.804	0.784	0.787	0.745	0.777	0.790	0.764

Taula 3: Valors màxims de precisió obtinguts amb les diferents variants de detector d'idees relacionades entrenats per a totes les dades. Cada fila correspon a un conjunt de test: per a cada categoria, i per a totes les dades juntes.

5.1 Detector d'idees relacionades combinat amb el classificador de categories

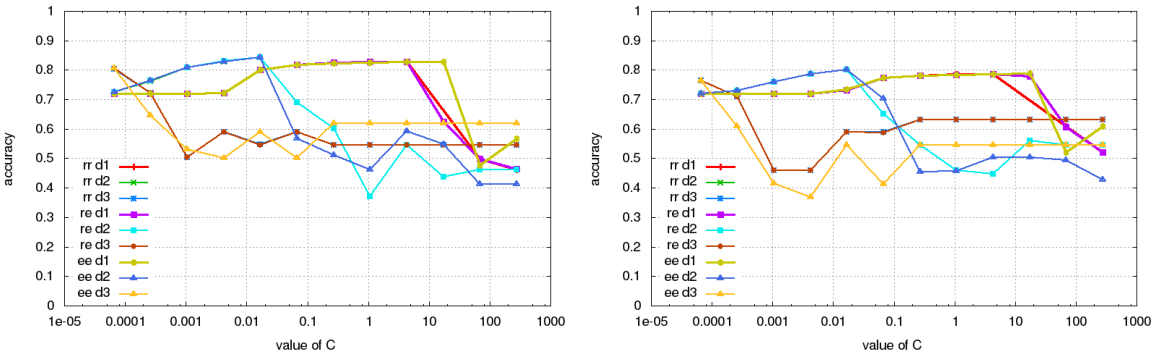
Com hem vist, entrenar models especialitzats per a cada categoria no millora el rendiment del detector de parelles d'idees relacionades, i fins i tot l'empitjora en alguns casos. Aquests últims casos es pot atribuir sobretot al fet que per determinades categories hi ha menys dades disponibles, i per tant el model generat és menys robust.

Tot i això, hem dissenyat un últim experiment on hem inclòs l'etiqueta de la categoria com un atribut més de la col·lecció per entrenar el classificador. L'objectiu que perseguim és determinar si el *classificador de categories* pot ajudar a millorar el detector d'idees relacionades.

Repetim els experiments anteriors afegint un atribut nou: la categoria. En un cas l'etiqueta l'obtindrem de les dades originals, així que el detector de parelles d'idees obté l'etiqueta correcta. En un altre experiment l'etiqueta de la categoria serà el que ha estimat el *classificador de categories*. Veurem si és un atribut rellevant pel classificador i com afecta el fet que l'etiqueta sigui *estimada* i no real.

Les gràfiques de la figura 2, juntament amb les figures de l'apèndix B mostren els resultats obtinguts per kernels de grau 1, 2 i 3, i per a la combinació d'etiquetes: *rr*: quan tant les etiquetes d'entrenament com les de test són reals; *re* quan les etiquetes d'entrenament són reals i les de test són estimades; i *ee* quan tant les etiquetes d'entrenament com les de test són estimades. La figura 2 mostra els resultats dels detectors de parelles d'idees entrenats amb totes les dades d'entrenament (model general) i separats segons els conjunt d'atributs emprats.

L'apèndix B mostra, a més, els resultats obtinguts amb les detectors d'idees especialitzats per a cada categoria.



(a) Conjunt complet de 50 d'atributs

(b) Selecció automàtica de 31 atributs

Figura 2: Resultats del detector de parelles relacionades combinat amb el classificador de categories obtinguts per a dos conjunts d'atributs diferents

Com es pot apreciar a les figures, els nous models no mostren cap guany significatiu respecte els models anteriors. Els millors models s'obtenen amb kernels polinòmics de grau 2. Amb els models generals (totes les dades) s'assoleixen valors de precisió lleugerament superiors (1 punt en el cas del conjunt d'atributs seleccionats). Mentre que amb els models especialitzats per categories, no n'hi ha cap que mostri una diferència significativa.

Si comparem els models que usen les etiquetes reals i els que usen les estimades, és interessant veure que tots obtenen els mateixos valors màxims de precisió. Aquesta dada ens indica que probable les etiquetes de categories no s'estan usant en els models amb precisió més alta.

6 Recuperador automàtic d'idees relacionades

Aquest últim experiment es va dissenyar per provar l'eficàcia del detector de parelles d'idees relacionades en una hipotètica aplicació on l'objectiu és *donada una Idea I_i trobar un conjunt d'idees I' que hi estiguin relacionades* i veure amb quina precisió el sistema troba idees relacionades dins aquest subconjunt I'. El model que hem dissenyat a la secció 5 és un classificador binari que només permet comparar un parell donat d'idees. Per poder simular un sistema de recuperació d'idees relacionades, busquem al nostre corpus totes les parelles on aparegui la idea I_i, les comparem i creem un rànquing de més similars a menys. D'aquest rànquing ens quedem amb les N millors (topN) i seguidament calculem l'exactitud, precisió i cobertura d'aquest I'_{topN}. Hem avaluat diferents mides del *topN*: *0*, 1, 3 i 5, on el valor 0 indica una mida variable que selecciona totes les parelles que tinguin una similitud positiva (és a dir, per sobre de 0).

Per dur a terme aquests experiments hem establert les següents variables:

D : El grau del polinomi del kernel de la SVM: 1, 2 i 3.

C : El valor del paràmetre c del classificador.

Categoria : Aplicacions mòbils, Banca mòbil, Banca per Internet, Caixers i Targetes.

Atributs : El conjunt d'atributs emprats: tots els atributs, la selecció d'atributs i la selecció d'atributs sense redundància.

N : El nombre d'elements en el conjunt I'_{topN}: *0*, 1, 3, i 5.

Per a cada experiment, calculem l'exactitud, la precisió i el cobertura de la següent forma:

- E^N és la mitjana de l'exactitud $E^N_{I_i}$ per a totes les idees I_i de les dades de test. $E^N_{I_i}$ és l'exactitud del conjunt I'_{topN} respecte la idea I_i.

$$E^N_{I_i} = \begin{cases} 1 & \text{if } \exists I'_j \backslash I_i, I'_j \text{ estan relacionat} \\ 0 & \text{altrament} \end{cases}$$

- P^N és la mitjana de la precisió $P^N_{I_i}$ per a totes les idees I_i de les dades de test. $P^N_{I_i}$ és la precisió del conjunt I'_{topN} respecte la idea I_i. $P^N_{I_i} = \frac{\text{tp}}{\text{tp+fp}}$, on *tp* és el nombre d'idees que s'han identificat correctament com idees relacionades, i *fp* és el nombre d'idees que s'han etiquetat errròniament com a relacionades.

- R^N és la mitjana del cobertura $R^N_{I_i}$ per a totes les idees I_i de les dades de test. $R^N_{I_i}$ és el cobertura del conjunt I'_{topN} respecte la idea I_i. $R^N_{I_i} = \frac{\text{tp}}{\min\{M,N\}}$, on *tp* és el nombre d'idees que s'han identificat correctament com idees relacionades, M és el nombre d'idees relacionades que hi ha al conjunt I', i N és la mida del subconjunt I'_{topN}.

Els resultats obtinguts amb els experiments es poden trobar al directori de dades que s'adjunta a aquest document. En general s'observa que els millors resultats s'obtenen amb el kernel polinòmic de grau 2. Respecte als conjunts d'atributs, els millors resultats també s'obtenen usant el conjunt complet de 50 atributs, tot i que els resultats amb el subconjunt seleccionat de forma automàtica (secció 4) són molt similars. En canvi, els models entrenats amb el subconjunt més petit d'atributs obtenen uns resultats notablement més baixos. Finalment, comparem els models generats amb totes les dades disponibles amb els models entrenats per a cada categoria. A diferència del que es va observar als experiments amb el classificador binari (secció 5), en aquests experiments els models entrenats per a cada categoria si obtenen millors resultats que el model general. En concret, si ens fixem en el valor de l'exactitud, el model general aconsegueix valors al voltant del 70% en els millors casos, mentre que els models per categories assoleixen valors d'entre el 70% i el 80%. Aquests valors fins i tot es superen en el cas de *Banca per Internet* (80%–90%). L'excepció són alguns casos en què s'usa només el conjunt petit d'atributs i la mesura d'exactitud baixa fins al 50%–60%.

Com que el nombre total de figures amb els resultats estadístics que és molt alt, l'apèndix C mostra només les figures que es corresponen a la millor configuració: kernel polinòmic de grau 3 ($D = 3$) i models generals (és a dir, sense especialitzar per categoria). Cada figura mostra el valor de E^N, P^N o R^N obtingut en mitjana durant l'avaluació creuada. Cadascuna de les figures es correspon al model entrenat amb un conjunt d'atributs, i cadascuna de les corbes representa els valor obtinguts per a un N determinat (és a dir, la mida del conjunt I') i un valor de c.

	Conjunt d'atributs →	**Complet** (50 **atrs.)**			**Automàtic** (31 **atrs.)**			**Mínim** (20 **atrs.)**		
	Grau del kernel →	*d1*	*d2*	*d3*	*d1*	*d2*	*d3*	*d1*	*d2*	*d3*
Exactitud	$N = 0$	0.663	0.704	**0.760**	0.580	0.527	**0.741**	0.512	0.533	**0.689**
	$N = 1$	0.654	0.702	**0.742**	0.570	0.525	**0.661**	0.507	0.532	**0.614**
	$N = 3$	0.673	0.710	**0.775**	0.604	0.595	**0.813**	0.528	0.543	**0.764**
	$N = 5$	0.674	0.710	**0.781**	0.608	0.659	**0.856**	0.572	0.561	**0.820**
Precisió	$N = 0$	0.632	0.690	**0.746**	0.556	0.520	**0.741**	0.492	0.526	**0.689**
	$N = 1$	0.654	0.702	**0.742**	0.570	0.525	**0.661**	0.507	0.532	**0.614**
	$N = 3$	0.572	0.588	**0.707**	0.508	0.587	**0.813**	0.498	0.483	**0.753**
	$N = 5$	0.550	0.613	**0.721**	0.490	0.647	**0.856**	0.545	0.537	**0.812**
Cobertura	$N = 0$	0.604	0.684	**0.711**	0.537	0.516	**0.654**	0.475	0.522	**0.611**
	$N = 1$	0.654	0.702	**0.742**	0.570	0.525	**0.661**	0.507	0.532	**0.614**
	$N = 3$	0.503	0.537	**0.627**	0.422	0.459	**0.668**	0.368	0.373	**0.616**
	$N = 5$	0.463	0.518	**0.649**	0.414	0.538	**0.745**	0.438	0.445	**0.685**

Taula 4: Valors màxims d'exactitud, precisió i cobertura obtinguts amb les diferents variants de recuperador d'idees entrenats per a totes les dades. Cada fila correspon a un valor d'N (mida de la finestra) diferent.

References

[DL14] Michael Denkowski and Alon Lavie. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the EACL 2014 Workshop on Statistical Machine Translation*, 2014.

[Dod02] George Doddington. Automatic Evaluation of Machine Translation Quality Using N-gram Co-Occurrence Statistics. In *Proceedings of the 2nd International Conference on Human Language Technology*, pages 138–145, San Francisco, CA, USA, 2002. Morgan Kaufmann Publishers Inc.

[GM10] Jesús Giménez and Lluís Màrquez. Asiya: An Open Toolkit for Automatic Machine Translation (Meta-)Evaluation. *The Prague Bulletin of Mathematical Linguistics*, (94):77–86, 2010.

[Hal98] M. A. Hall. *Correlation-based Feature Subset Selection for Machine Learning*. PhD thesis, University of Waikato, Hamilton, New Zealand, 1998.

[Joa02] Thorsten Joachims. *Learning to Classify Text Using Support Vector Machines: Methods, Theory and Algorithms*. Kluwer Academic Publishers, Norwell, MA, USA, 2002.

[Lev66] Vladimir Iosifovich Levenshtein. Binary Codes Capable of Correcting Deletions, Insertions and Reversals. *Soviet Physics Doklady*, 8(10):707–710, 1966.

[LO04] Chin-Yew Lin and Franz Josef Och. Automatic Evaluation of Machine Translation Quality Using Longest Common Subsequence and Skip-Bigram Statics. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL)*, 2004.

[MGT03] I. Dan Melamed, Ryan Green, and Joseph P. Turian. Precision and Recall of Machine Translation. In *Proceedings of the Joint Conference on Human Language*

Technology and the North American Chapter of the Association for Computational Linguistics (HLT-NAACL), pages 61–63, Stroudsburg, PA, USA, 2003. Association for Computational Linguistics.

[NOLN00] Sonja Nießen, Franz Josef Och, Gregor Leusch, and Hermann Ney. An Evaluation Tool for Machine Translation: Fast Evaluation for MT Research. In *Proceedings of the 2nd International Conference on Language Resources and Evaluation (LREC)*, 2000.

[PRWZ02] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: A Method for Automatic Evaluation of Machine Translation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, 2002.

[SFI92] Michel Simard, George F. Foster, and Pierre Isabelle. Using Cognates to Align Sentences in Bilingual Corpora. In *Proceedings of the Fourth International Conference on Theoretical and Methodological Issues in Machine Translation*, 1992.

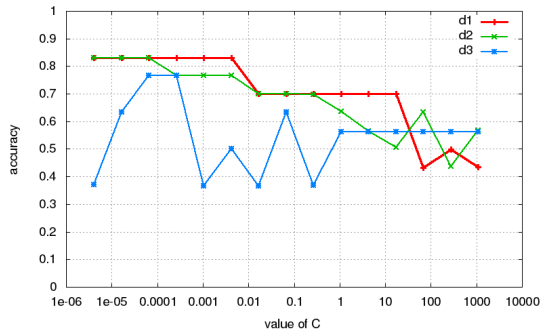
[SMDS09] Matthew Snover, Nitin Madnani, Bonnie Dorr, and Richard Schwartz. Fluency, adequacy, or HTER? Exploring different human judgments with a tunable MT metric. In *Proceedings of the 4th Workshop on Statistical Machine Translation*, pages 259–268, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics.

[TVN+97] Christoph Tillmann, Stefan Vogel, Hermann Ney, A. Zubiaga, and H. Sawaf. Accelerated DP based Search for Statistical Translation. In *Proceedings of European Conference on Speech Communication and Technology*, pages 2667–2670, 1997.

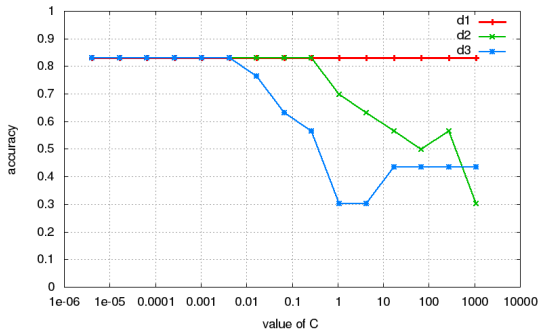
A Resultats de l'avaluació creuada del detector de parelles d'idees relacionades

Els següents resultats es corresponen a l'avaluació dels classificadors d'idees relacionades. Cada subsecció es correspon a l'avaluació per a un tipus d'idea. Per a cada tipus d'idea s'avaluen 6 models: (*i*) un model per a cadascun dels 3 conjunts d'atributs (50, 31, 20), i (*ii*) cadascun d'aquests entrenat amb totes les dades disponibles (model general) o entrenat només amb les dades particulars de la categoria (model especialitzat).

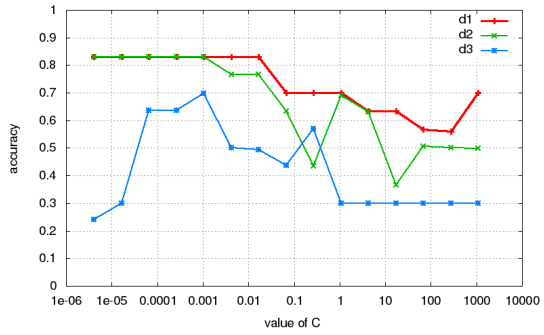
A.1 Aplicacions mòbils



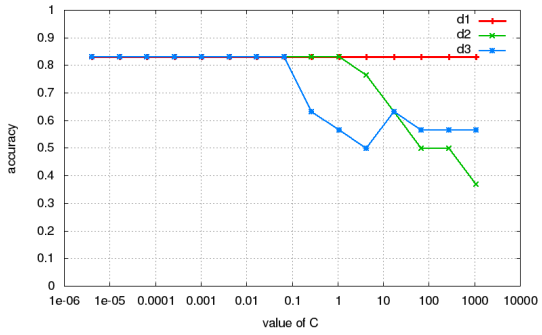
(a) 50 Atributs; Model General



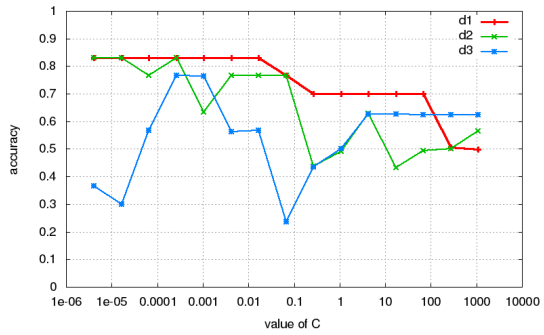
(b) 50 atributs; Model Especialitzat



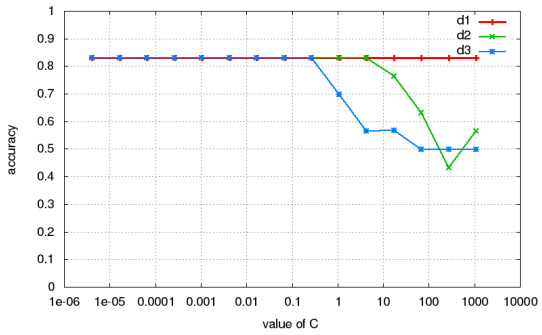
(c) 31 Atributs; Model General



(d) 31 atributs; Model Especialitzat



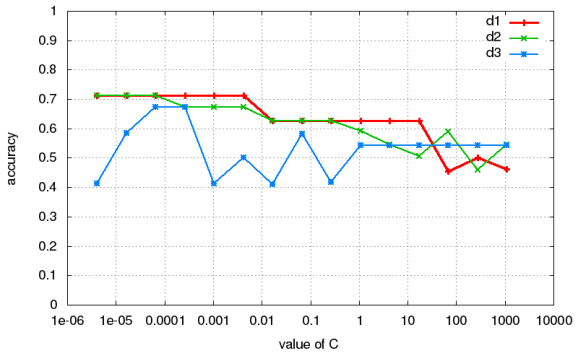
(e) 20 Atributs; Model General



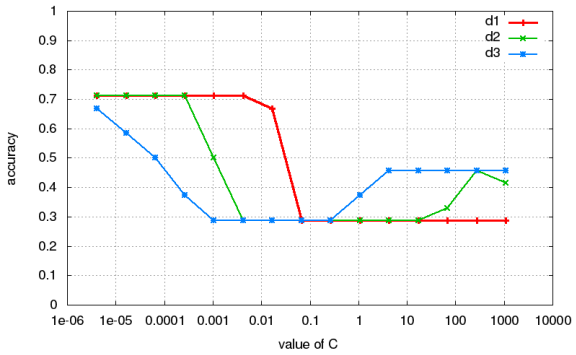
(f) 20 atributs; Model Especialitzat

Figura 3: Resultats obtinguts per a Aplicacions Mòbils

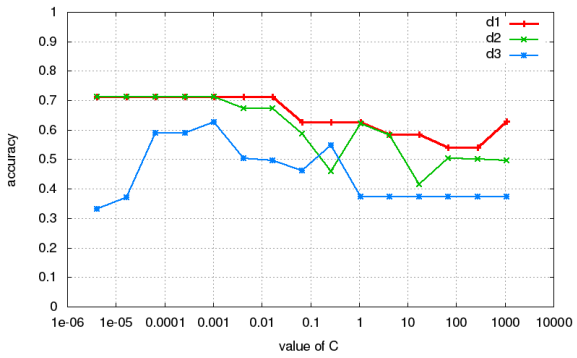
A.2 Banca mòbil



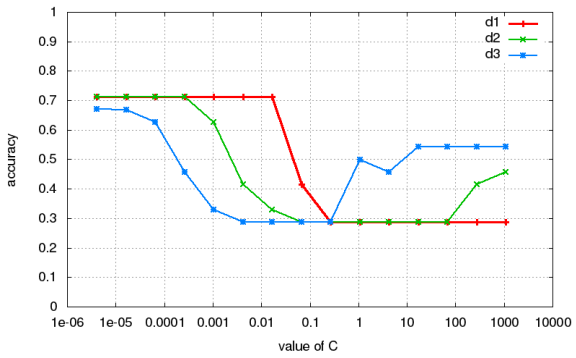
(a) 50 Atributs; Model General



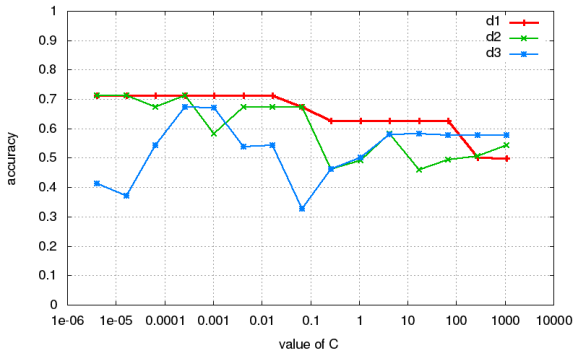
(b) 50 atributs; Model Especialitzat



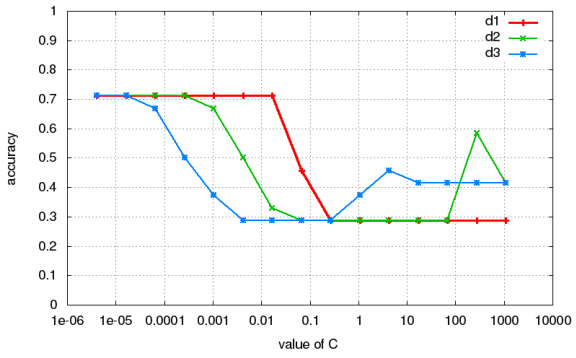
(c) 31 Atributs; Model General



(d) 31 atributs; Model Especialitzat



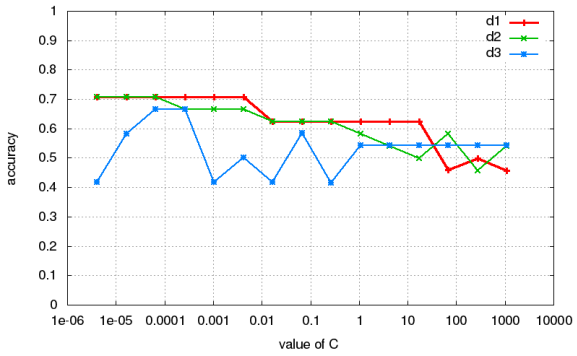
(e) 20 Atributs; Model General



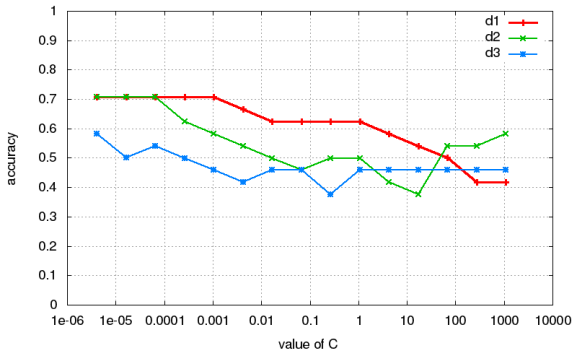
(f) 20 atributs; Model Especialitzat

Figura 4: Resultats obtingunts per a Banca Mòbil

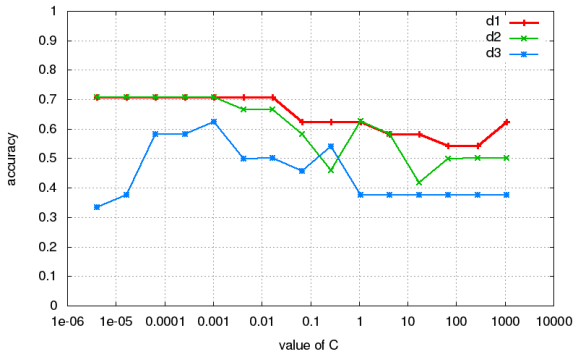
A.3 Banca per Internet



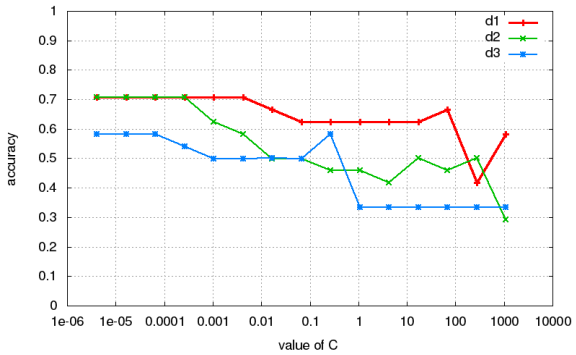
(a) 50 Atributs; Model General



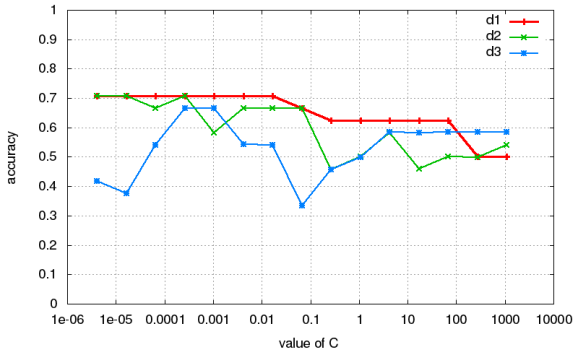
(b) 50 atributs; Model Especialitzat



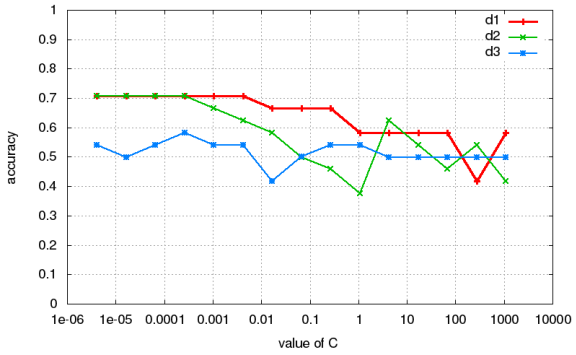
(c) 31 Atributs; Model General



(d) 31 atributs; Model Especialitzat



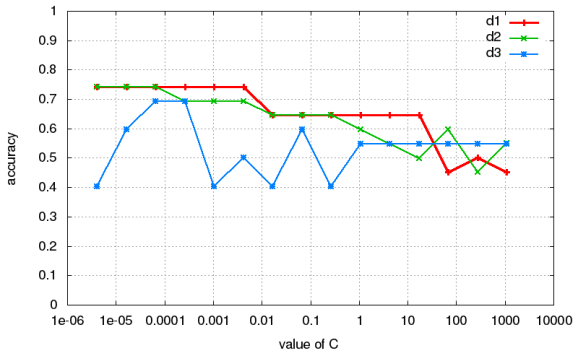
(e) 20 Atributs; Model General



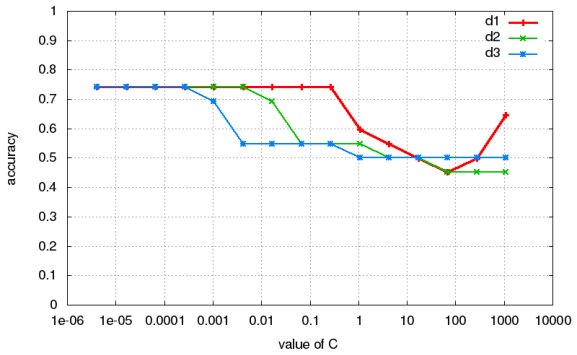
(f) 20 atributs; Model Especialitzat

Figura 5: Resultats obtinguts per a Banca per Internet

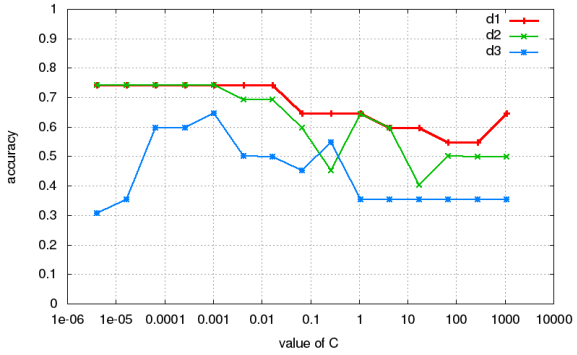
A.4 Caixers



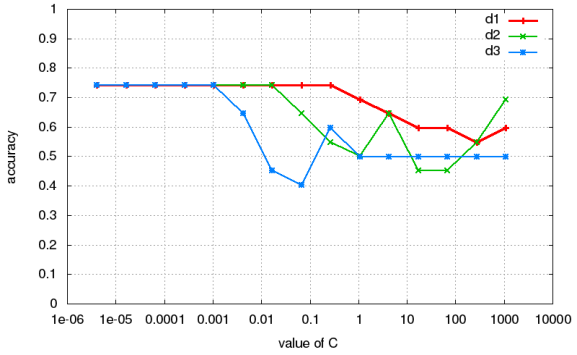
(a) 50 Atributs; Model General



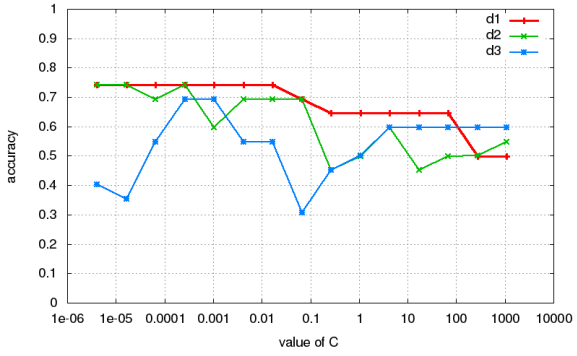
(b) 50 atributs; Model Especialitzat



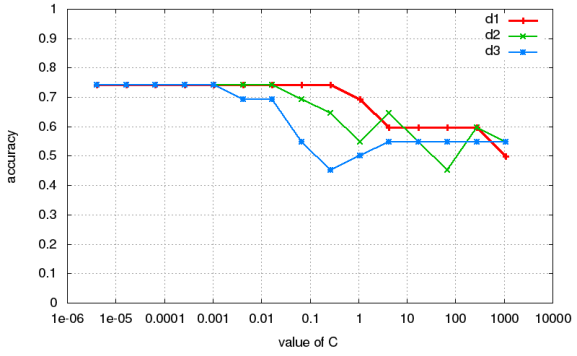
(c) 31 Atributs; Model General



(d) 31 atributs; Model Especialitzat



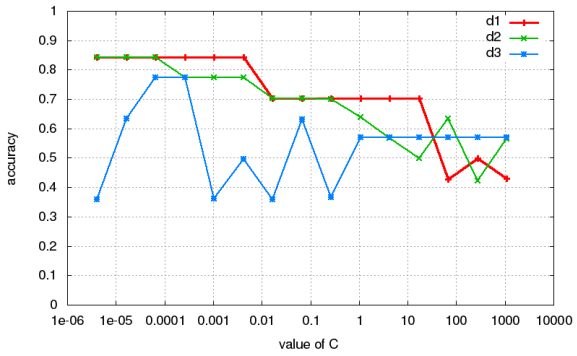
(e) 20 Atributs; Model General



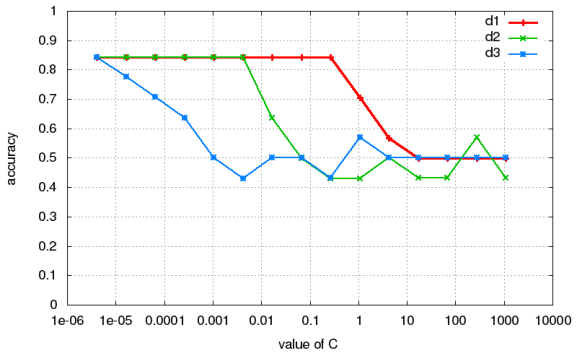
(f) 20 atributs; Model Especialitzat

Figura 6: Resultats obtinguts per a Caixers

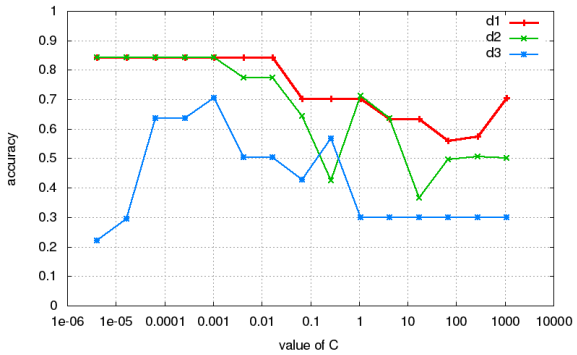
A.5 Targetes



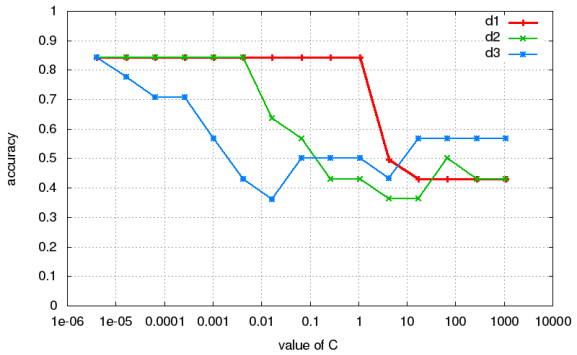
(a) 50 Atributs; Model General



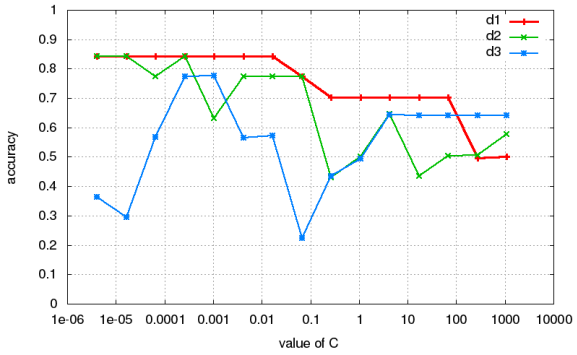
(b) 50 atributs; Model Especialitzat



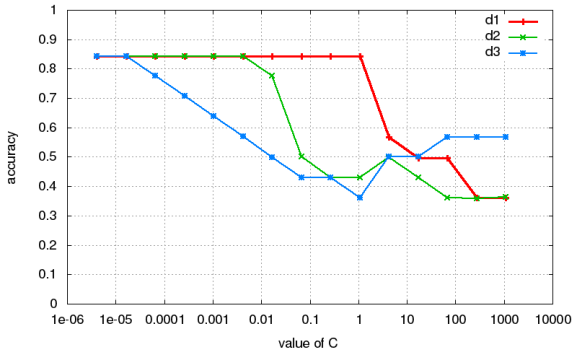
(c) 31 Atributs; Model General



(d) 31 atributs; Model Especialitzat



(e) 20 Atributs; Model General



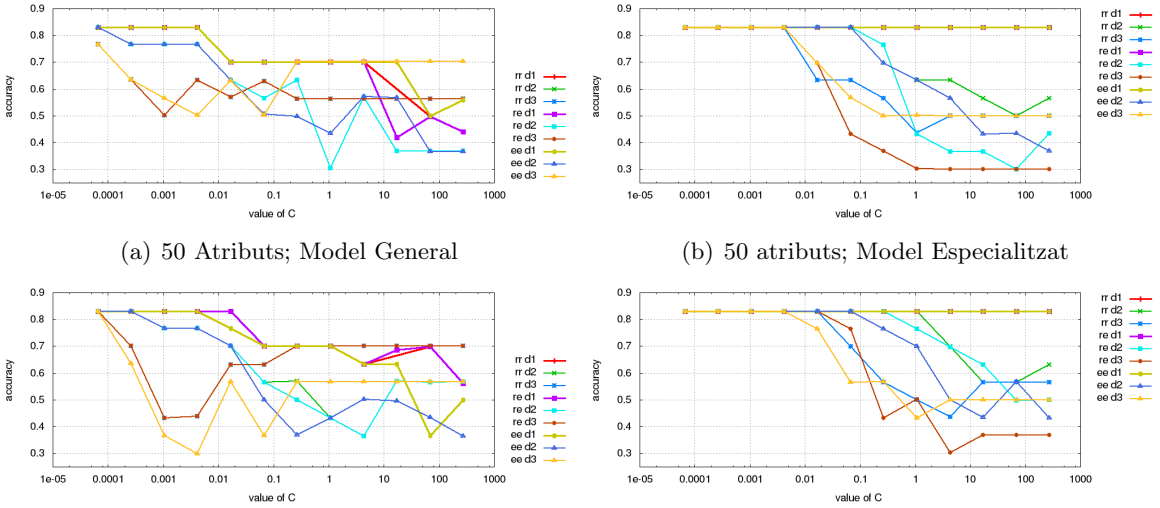
(f) 20 atributs; Model Especialitzat

Figura 7: Resultats obtinguts per a Targetes

B Resultats de l'avaluació creuada del detector de parelles d'idees relacionades amb l'etiquetador de categories

Els següents resultats es corresponen a l'avaluació dels identificadors d'idees relacionades combinat amb el **classificador de categories**. Cada subsecció es correspon a l'avaluació per a un tipus d'idea. Per a cada tipus d'idea s'avaluen 4 models: (*i*) un model per a cadascun dels 3 conjunts d'atributs (50, 31), i (*ii*) cadascun d'aquests entrenat amb totes les dades disponibles (model general) o entrenat només amb les dades particulars de la categoria (model especialitzat). Les gràfiques mostren els resultats obtinguts amb etiquetes de categories reals i estimats: *rr* són els models entrenats i testejats amb etiquetes reals; *re* són models entrenats amb etiquetes reals i testejats amb etiquetes estimades; i *ee* són els models entrenats i testejats amb etiquetes estimades.

B.1 Aplicacions mòbils



(a) 50 Atributs; Model General

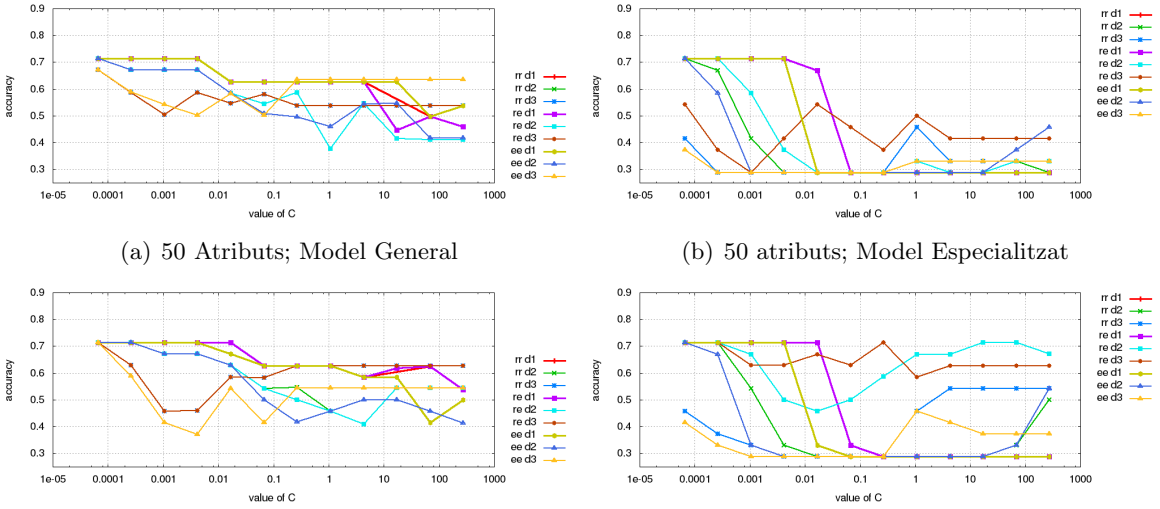
(b) 50 atributs; Model Especialitzat

(c) 31 Atributs; Model General

(d) 31 atributs; Model Especialitzat

Figura 8: Resultats obtinguts per a Aplicacions Mòbils

B.2 Banca mòbil



(a) 50 Atributs; Model General

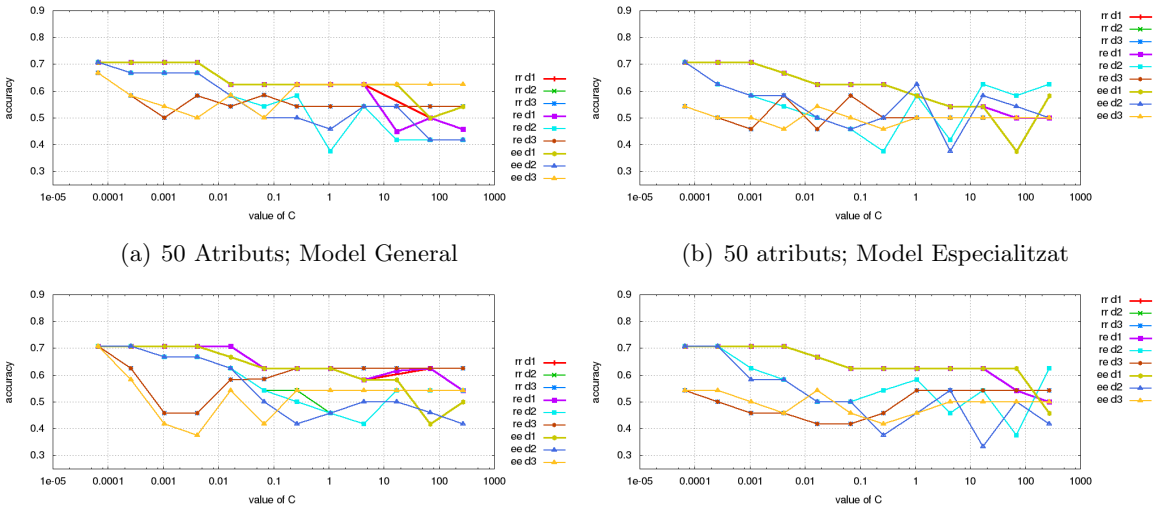
(b) 50 atributs; Model Especialitzat

(c) 31 Atributs; Model General

(d) 31 atributs; Model Especialitzat

Figura 9: Resultats obtingunts per a Banca Mòbil

B.3 Banca per Internet



(a) 50 Atributs; Model General

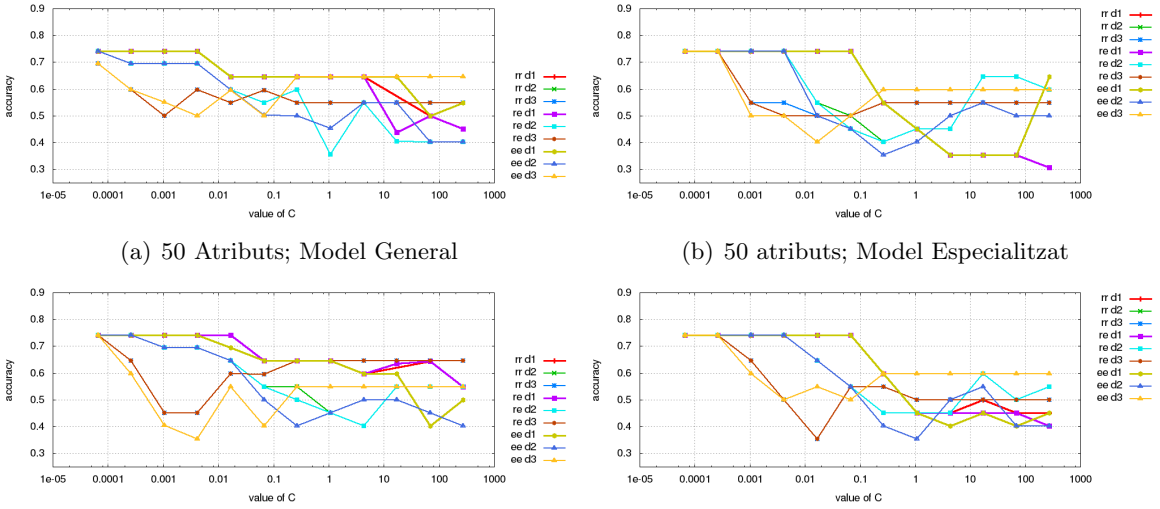
(b) 50 atributs; Model Especialitzat

(c) 31 Atributs; Model General

(d) 31 atributs; Model Especialitzat

Figura 10: Resultats obtinguts per a Banca per Internet

B.4 Caixers



(a) 50 Atributs; Model General

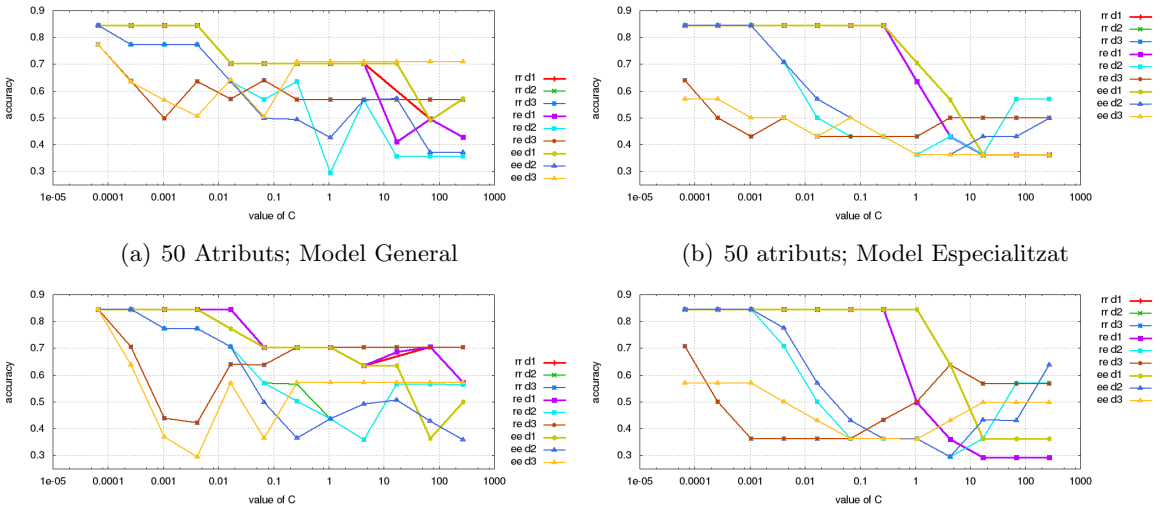
(b) 50 atributs; Model Especialitzat

(c) 31 Atributs; Model General

(d) 31 atributs; Model Especialitzat

Figura 11: Resultats obtinguts per a Caixers

B.5 Targetes



(a) 50 Atributs; Model General

(b) 50 atributs; Model Especialitzat

(c) 31 Atributs; Model General

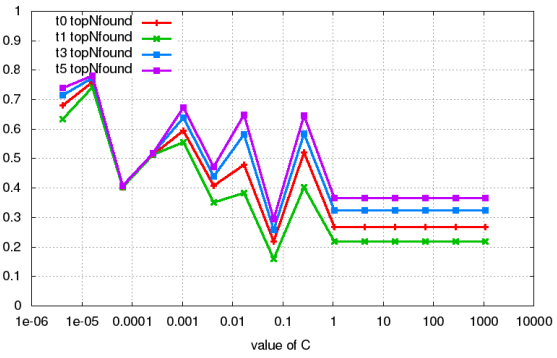
(d) 31 atributs; Model Especialitzat

Figura 12: Resultats obtinguts per a Targetes

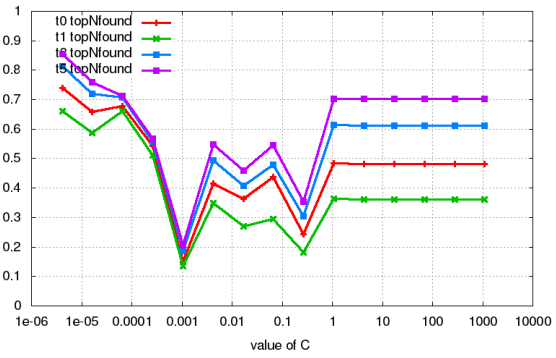
C Resultats de l'avaluació creuada del recuperador d'idees

Els següents resultats es corresponen a l'avaluació dels recuperadors d'idees relacionades per a tres mesures (exactitud, precisió i cobertura) per a cadascun dels 3 conjunts d'atributs (50, 31, 20). Els millors resultats que es van obtenir amb els classificadors de grau 2, entrenats amb totes les dades disponibles (sense especialització per categories). Cada corba es correspon a una mida diferent del conjunt d'idees relacionades que s'avalua, on *t0* es correspon a una mida variable que abasta totes les parelles amb una similitud positiva.

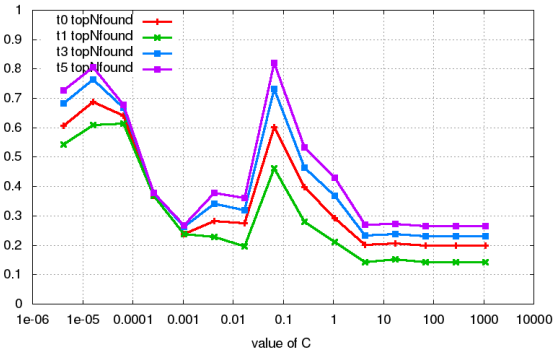
C.1 Exactitud



(a) 50 Atributs



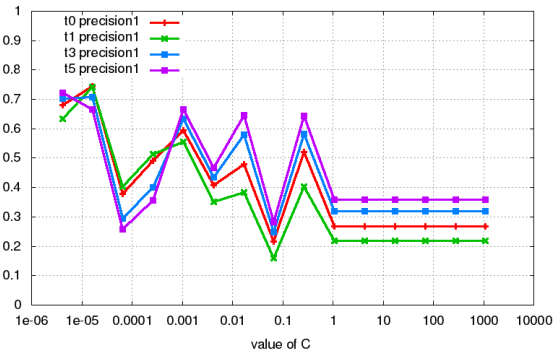
(b) 31 Atributs



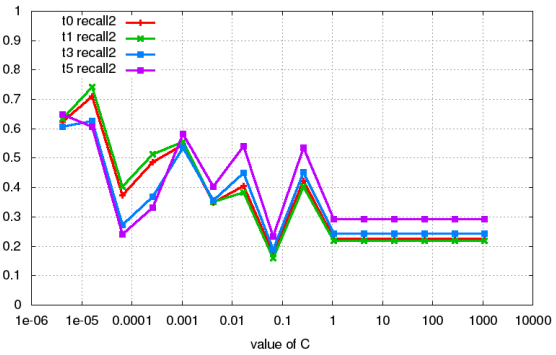
(c) 20 Atributs

Figura 13: Resultats d'exactitud per a diferents conjunts d'atributs

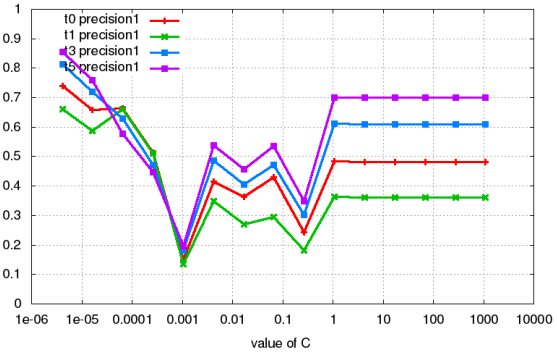
C.2 Precisió i cobertura



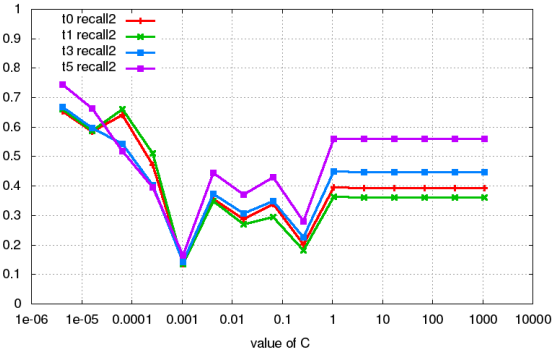
(a) 50 Atributs; Precisió



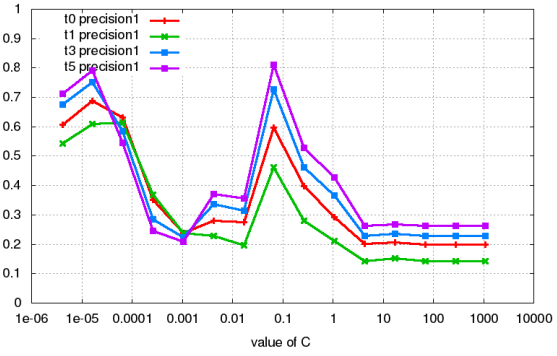
(b) 50 Atributs; Cobertura



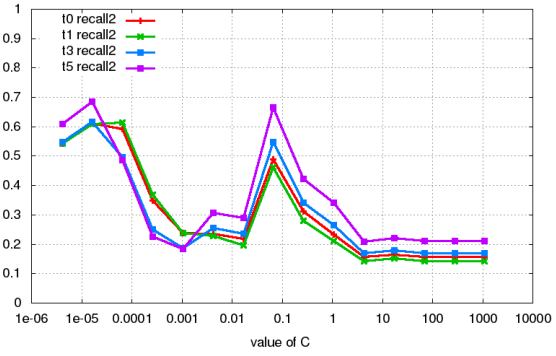
(c) 31 Atributs; Precisió



(d) 31 Atributs; Cobertura



(e) 20 Atributs; Precisió



(f) 20 Atributs; Cobertura

Figura 14: Resultats de precisió i cobertura per a diferents conjunts d'atributs