MKIDS: identificación de roles, grupos de inclusión y proyectos

Tom Murray & Ed Hovy
Information Sciences Institute
University of Southern California
{yohzik, hovy}@isi.edu

Víctor Peinado
Grupo de PLN y Recuperación de Información de la UNED
victor@lsi.uned.es

NLP Group

14 de enero de 2005

Índice

- Proyecto MKIDS: objetivo y aproximación desde el PLN
- Trabajo previo sobre corpus ISI
  - Identificando temas (*clustering*)
  - *CBC clustering*
  - *Signatures*
  - Actos de habla
- Trabajo realizado sobre corpus Enron
  - Indetificando temas (*clustering*)
  - Identificando temas tratados.
  - Identificando grupos de proyecto (cliques)
- Ideas para el artículo

MKIDS

OBJETIVO: Descubrir y modelar procesos y cambios tal y como se observan en los participantes de entornos ricos en flujos de información.

APROXIMACIÓN: Estudiamos el lenguaje (correo-e) como evidencia del entorno.

· ¿Cuáles son los intereses de la gente y los temas que tratan?
· ¿Cuál es la topología de la red social?
· ¿Qué tipos de interacción y roles encontramos?

Trabajo realizado sobre el corpus ISI

591 mensajes (2600 párrafos) de cinco miembros del ISI a lo largo de un mes.

TAREAS:

· Procesado de los datos.
· *Clustering* de los temas tratados (*topic clustering*).
· Análisis y visualización de la red.
· Clasificación de actos de habla.

Trabajo realizado sobre el corpus ISI

CLUSTERING DE TEMAS:

Identificamos temas de discusión en el corpus, quién participa y a quién se dirige.

· *Clusters* de mensajes y párrafos a nivel de palabra.

· Creamos una firma para cada *cluster* (*topic signature*).

· Creamos una firma para cada usuario (*person signature*).

· Emparejamos las firmas de cada usuario con la firma del tema más similar.

Clustering

· K-means: rápido aunque hay que especificar un número de *clusters* (25, 40, 80).
· C-Link: *clustering* jerárquico aglutinante. Resultados sin éxito.
· CBC *Clustering by Comitee (*Pantel&Lin): no hace falta especificar *a priori* un número de *clusters* pero es lento.

SOLUCIÓN: CBC sobre una muestra aleatoria de mensajes y K-means sobre la colección entera.

Clustering by Committee

Otros algoritmos representan un cluster por el centroide de sus miembros o un elemento representativo.

· El centroide puede verse influenciado por los elementos marginales del cluster.

· Un solo representante puede ser problemático, dado que cada elemento tiene sus propias rasgos idiosincrásicos.

CBC construye el centroide del cluster promediando los vectores de rasgos de un subcojunto de los miembros del cluster. Eligiendo cuidadosamente los miembros de un comité, las rasgos del centroide serán cada vez más característicos con la clase en cuestion.

Clustering by Committee

Clusters de nombre de estados por contexto de aparición:

la corte de __	la capital de __	el senador por __
el gobernador de __		ilegal en __

Nueva York y Washington son nombres de ciudades, por lo que el centroide contendrá rasgos como:

el alcalde de __	el metro de __
el aeropuerto de __	el ayuntamiento de __

CBC formará un comité que contenga los miembros ideales de la clase:

Nebraska, Illinois, Michigan, Iowa...

Signatures

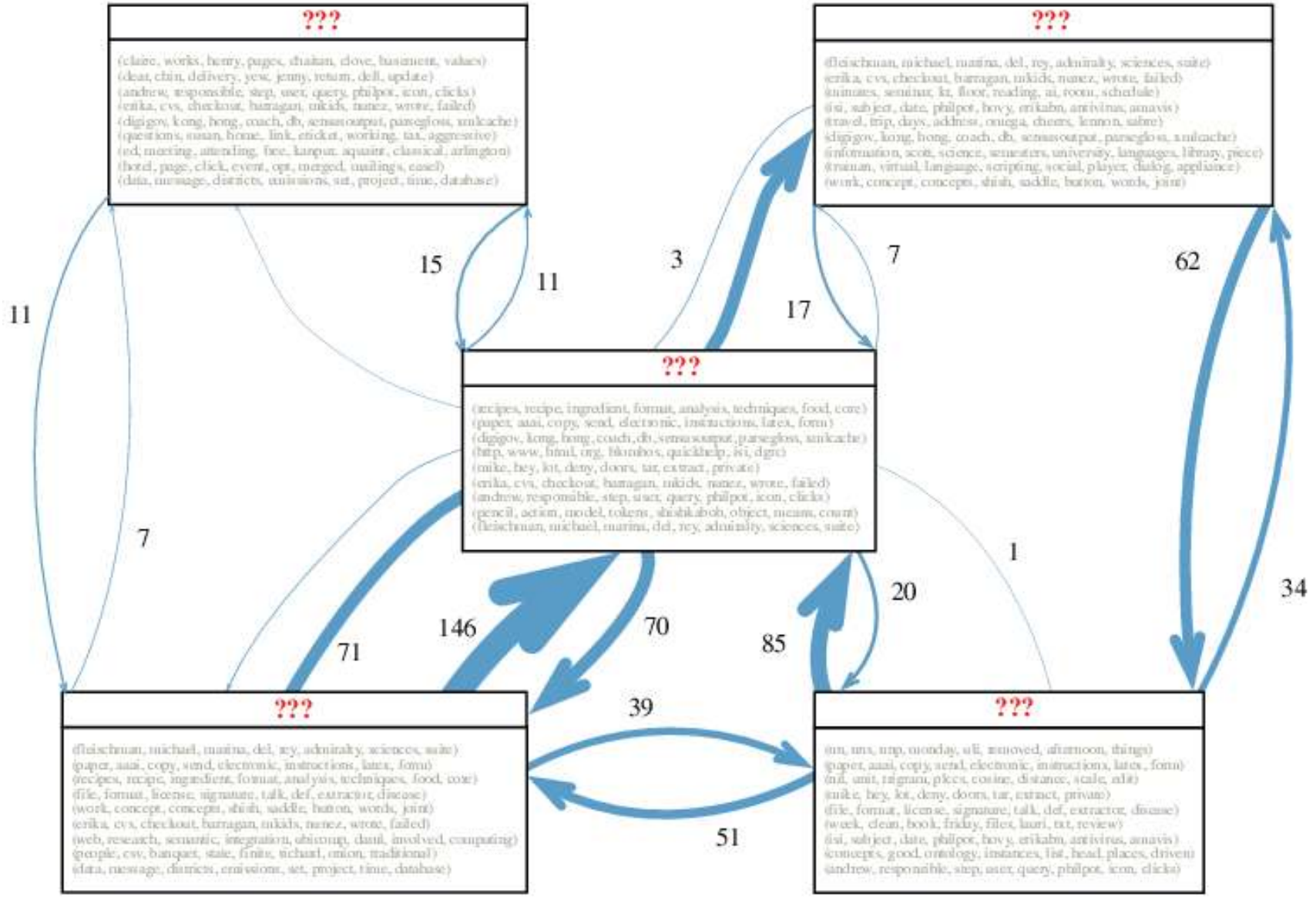
Conjunto de palabras que son relevantes para una determinada colección de documentos, es decir, que aparecen en dicho conjunto más a menudo que en otros. (Lin&Hovy).

Consideramos únicamente aquellos términos que superan un umbral de relevancia utilizando el test estadístico χ^2 (Manning&Schüzte, pp. 172-73).

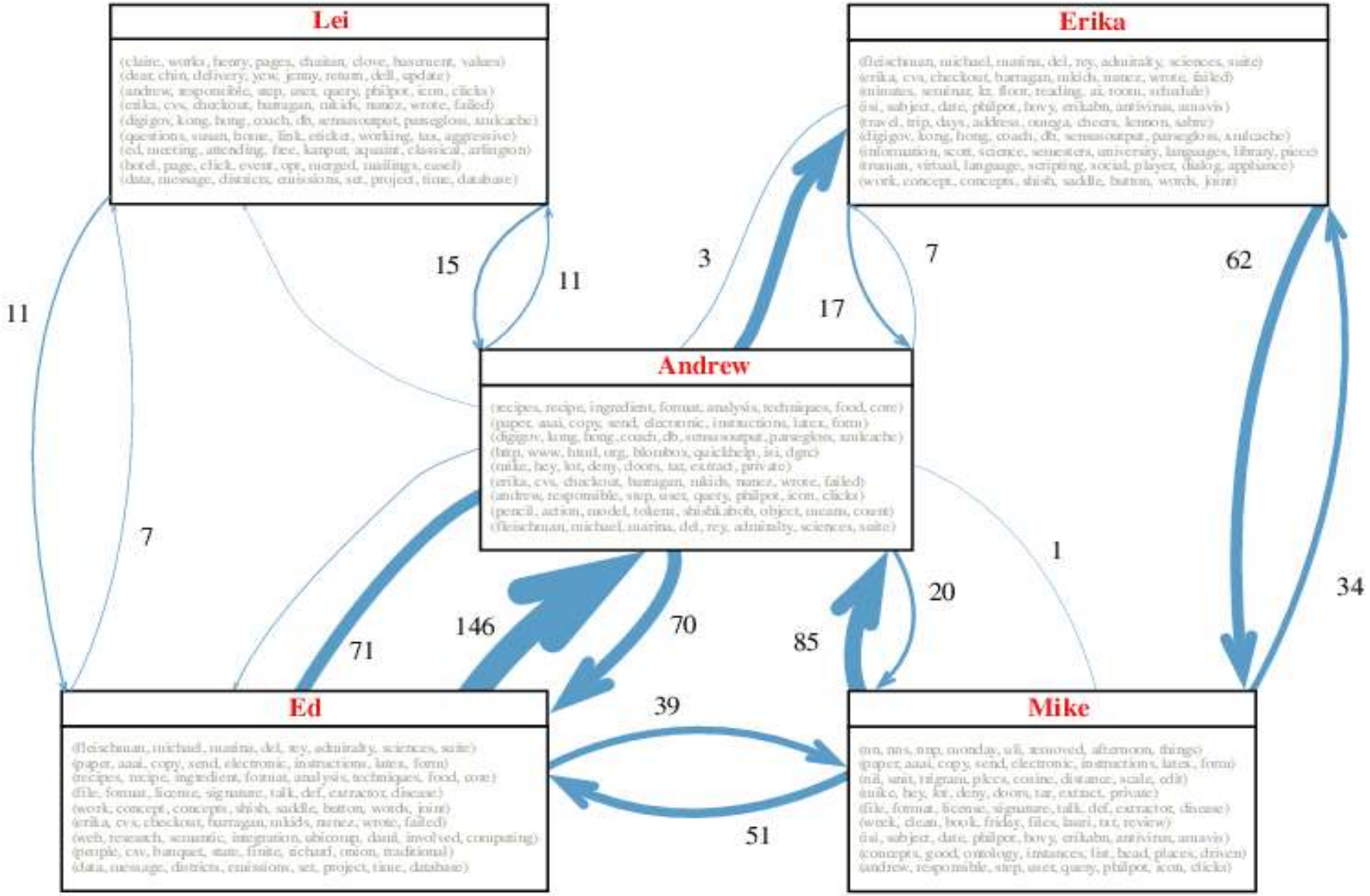
Ejemplos de *topic signatures*

596 { rigzon gif http www spacer image expedia rfrr chtah postdirect }
26 { deal enter ce sale term chang image volum bookout creat }
167 { capac transwestern releas shipper dty post rate pipelin transport recours }
1 { d www http html obido amazon cust hol mk subscrib }
58 { pipelin paso el mmcf ga expansion project capac natur northern }
302 { curv valid basy file gd price gdy map omicron vol }
146 { hey guy ndb sunday town weekend talk hope work mom }
357 { ferc order market commiss file iso refund rehear california issu }
173 { game play ticket dvd watch wizard playstat season saturday blazer }
90 { sara shackleton isda cheryl tanya agreement laurel carolin yair ena }

Corpus ISI


???
(claire, works, henry, pages, chatan, clove, basement, values)
(dear, chin, delivery, yew, jenny, return, dell, update)
(andrew, responsible, step, user, query, philpot, icon, clicks)
(erika, cvs, checkout, barragan, nikids, nunez, wrote, failed)
(digigov, kong, hong, coach, db, sensusoutput, paraeglous, xmlcache)
(questions, susan, home, link, cricket, working, tax, aggressive)
(ed, meeting, attending, free, kanpur, aquaint, classical, arlington)
(hotel, page, click, event, opt, merged, mailings, easel)
(data, message, districts, emissions, set, project, time, database)
???
(fleischman, michael, marina, del, rey, admiralty, sciences, suite)
(erika, cvs, checkout, barragan, nikids, nunez, wrote, failed)
(inmates, seminar, kt, floor, reading, ai, room, schedule)
(isi, subject, date, philpot, hovy, erikabn, antivirus, amavis)
(travel, trip, days, address, omega, cheers, lennon, sabre)
(digigov, kong, hong, coach, db, sensusoutput, paraeglous, xmlcache)
(information, acoli, science, semesters, university, languages, library, piece)
(iranian, virtual, language, scripting, social, player, dialog, appliance)
(work, concept, concepts, shirh, saddle, button, words, joint)
???
(recipes, recipe, ingredient, format, analysis, techniques, food, core)
(paper, aaai, copy, send, electronic, instructions, latex, form)
(digigov, kong, hong, coach, db, sensusoutput, paraeglous, xmlcache)
(http, www, html, org, thumbnos, quickhelp, isi, dgtc)
(mike, hey, let, deny, doors, tar, extract, private)
(erika, cvs, checkout, barragan, nikids, nunez, wrote, failed)
(andrew, responsible, step, user, query, philpot, icon, clicks)
(pencil, action, model, tokens, shishkabob, object, means, count)
(fleischman, michael, marina, del, rey, admiralty, sciences, suite)
???
(fleischman, michael, marina, del, rey, admiralty, sciences, suite)
(paper, aaai, copy, send, electronic, instructions, latex, form)
(recipes, recipe, ingredient, format, analysis, techniques, food, core)
(file, format, license, signature, talk, def, extractor, disease)
(work, concept, concepts, shirh, saddle, button, words, joint)
(erika, cvs, checkout, barragan, nikids, nunez, wrote, failed)
(web, research, semantic, integration, ubisoup, dasl, involved, computing)
(people, csw, banquet, state, finite, richard, onion, traditional)
(data, message, districts, emissions, set, project, time, database)
???
(im, ims, imp, monday, ufl, removed, afternoon, things)
(paper, aaai, copy, send, electronic, instructions, latex, form)
(mil, unit, telegram, place, cosine, distance, scale, nfit)
(mike, hey, let, deny, doors, tar, extract, private)
(file, format, license, signature, talk, def, extractor, disease)
(week, clean, book, friday, files, lauri, txt, review)
(isi, subject, date, philpot, hovy, erikabn, antivirus, amavis)
(concepts, good, ontology, instances, list, head, places, driven)
(andrew, responsible, step, user, query, philpot, icon, clicks)
11
15
11
3
17
7
62
11
7
1
20
34
71
146
70
85
39
51

Corpus ISI


Lei
(claire, works, henry, pages, chanian, clove, basement, values)
(dear, chin, delivery, yew, jenny, return, dell, update)
(andrew, responsible, step, user, query, philpot, icon, clicks)
(erika, cvs, checkout, barragan, nikids, nunez, wrote, failed)
(digigov, kong, hong, coach, db, semioutput, parseglos, xmlcache)
(questions, vasan, bone, link, cricket, working, tax, aggressive)
(ed, meeting, attending, flee, kamput, aquaint, classical, arlington)
(hotel, page, click, event, opt, merged, mailings, easel)
(data, message, districts, emissions, set, project, time, database)
Erika
(fleischman, michael, marina, del, rey, admiralty, sciences, suite)
(erika, cvs, checkout, barragan, nikids, nunez, wrote, failed)
(minutes, seminar, lit, floor, reading, ai, room, schedule)
(isi, subject, date, philpot, bovy, erikabn, antivirus, umavis)
(travel, trip, days, address, omega, cheers, lennon, sabre)
(digigov, kong, hong, coach, db, semioutput, parseglos, xmlcache)
(information, scott, science, semesters, university, languages, library, piece)
(truman, virtual, language, scripting, social, player, dialog, appliance)
(work, concept, concepts, shish, saddle, button, words, joint)
Andrew
(recipes, recipe, ingredient, format, analysis, techniques, food, corn)
(paper, aaai, copy, send, electronic, instructions, latex, form)
(digigov, kong, hong, coach, db, semioutput, parseglos, xmlcache)
(http, www, html, org, bloonbox, quickhelp, isi, dgrc)
(mike, hey, lot, deny, doors, tar, extract, private)
(erika, cvs, checkout, barragan, nikids, nunez, wrote, failed)
(andrew, responsible, step, user, query, philpot, icon, clicks)
(pencil, action, model, tokens, shishkabob, object, means, count)
(fleischman, michael, marina, del, rey, admiralty, sciences, suite)
Ed
(fleischman, michael, marina, del, rey, admiralty, sciences, suite)
(paper, aaai, copy, send, electronic, instructions, latex, form)
(recipes, recipe, ingredient, format, analysis, techniques, food, corn)
(file, format, license, signature, talk, def, extractor, disease)
(work, concept, concepts, shish, saddle, button, words, joint)
(erika, cvs, checkout, barragan, nikids, nunez, wrote, failed)
(web, research, semantic, integration, ubicomp, david, involved, computing)
(people, csv, banquet, state, finite, richard, onion, traditional)
(data, message, districts, emissions, set, project, time, database)
Mike
(mn, mrs, mrp, monday, ufl, removed, afternoon, things)
(paper, aaai, copy, send, electronic, instructions, latex, form)
(nil, sent, telegram, piece, cosine, distance, scale, edit)
(mike, hey, lot, deny, doors, tar, extract, private)
(file, format, license, signature, talk, def, extractor, disease)
(week, clean, book, friday, files, lauri, txt, review)
(isi, subject, date, philpot, bovy, erikabn, antivirus, umavis)
(concepts, good, ontology, instances, list, head, places, driven)
(andrew, responsible, step, user, query, philpot, icon, clicks)
11
15
11
3
17
7
62
11
7
1
20
34
71
146
70
85
39
51

Actos de habla: «decir es hacer»

ACTO DE HABLA: acción llevada a cabo a través del lenguaje, como una descripción («Hace sol»), pregunta («¿Hace sol?»), petición u orden («¿Puedes pasarme la sal», «¡Arriba las manos!»), o promesa («Te prometo fidelidad absoluta»):

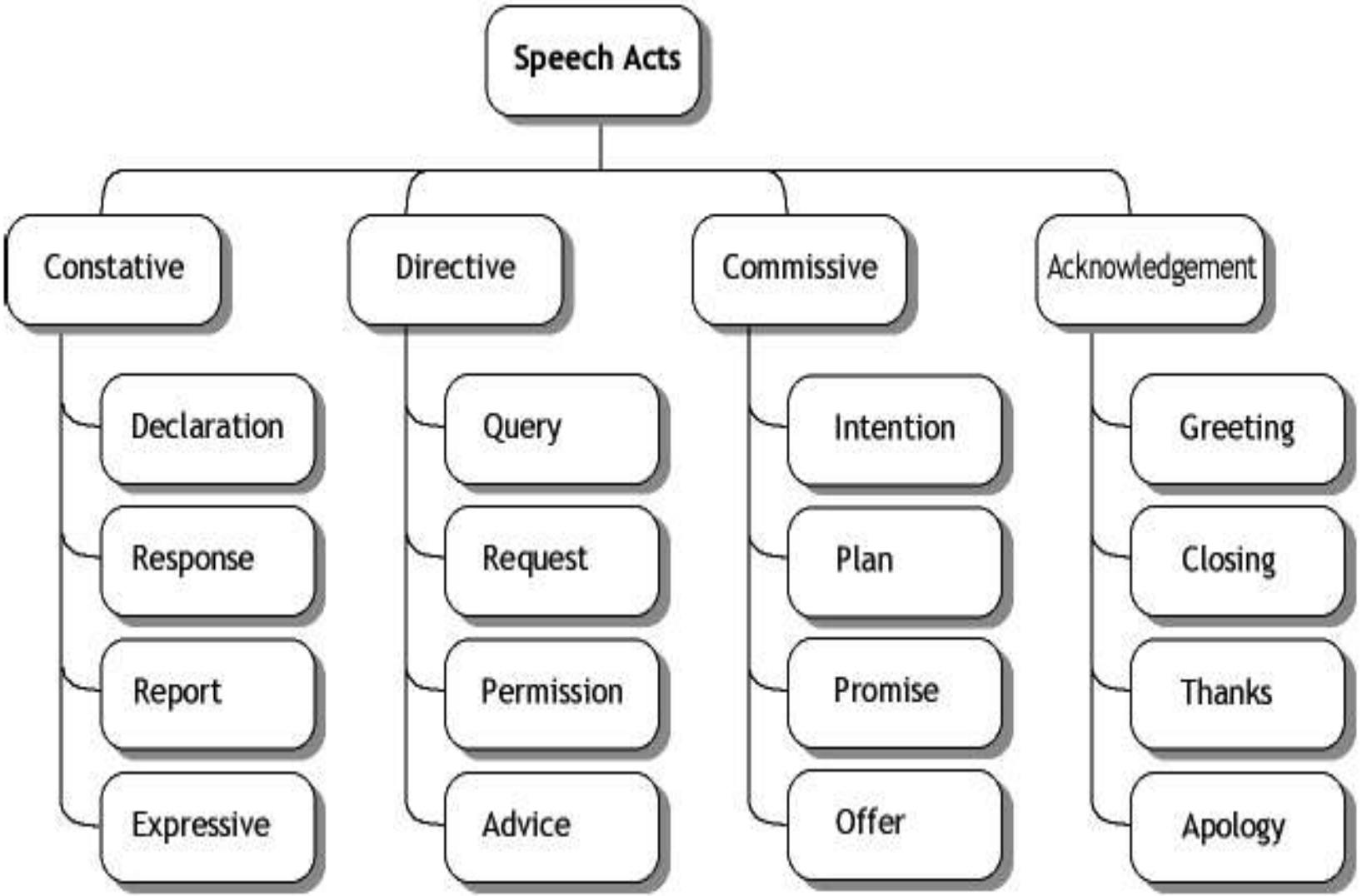
Al decir algo no solo enunciamos (locución) sino que también ejecutamos acciones como afirmar, mandar, rogar, prometer (ilocutivo). Pero al decir algo, las palabras producen resultados extralingüísticos, pueden desanimar, convencer o llevarnos a ejecutar cualquier acción (perlocución).

Actos de habla: experimentos

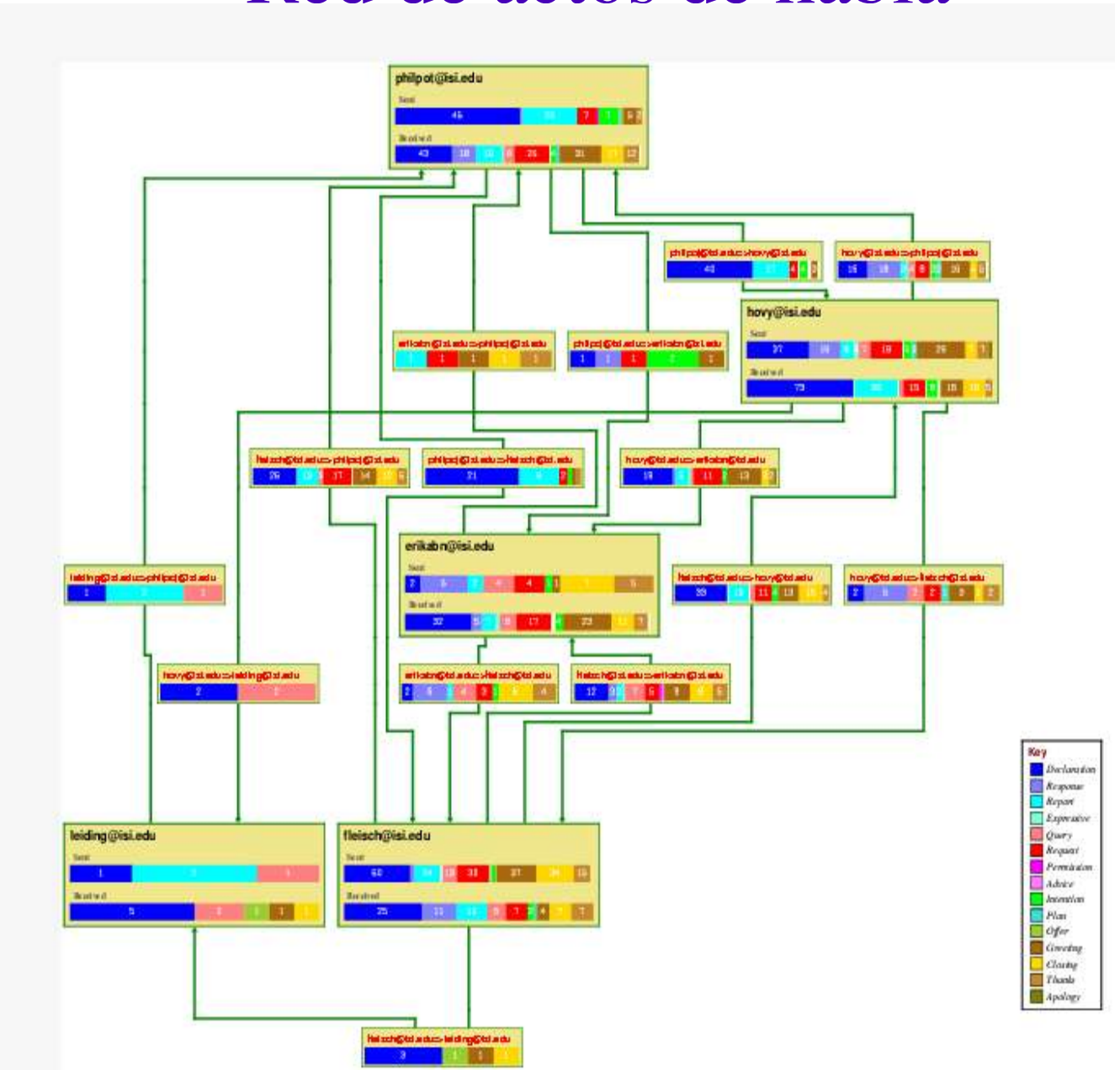
OBJETIVO: identificar cómo se comunica la gente para entender sus relaciones y roles

· Identificamos los actos de habla (mensaje, párrafo, oración).
· Clasificamos los textos automáticamente (métodos Máxima Entropía).
· Construímos redes sociales etiquetadas para comprender los roles de la gente a partir de los actos de habla.

Clasificación de actos de habla (Bach&Harnish simplificado)



Red de actos de habla



Corpus de Enron

· Disponible en `http://www-2.cs.cmu.edu/~enron`
· Corpus de 517.431 mensajes, distribuidos en 3600 carpetas de 151 empleados, entre 1998 y 2002.
· Hecho público por la comisión de investigación, adquirido por el MIT, organizado por Melinda Gervasio (SRI) y cedido a la comunidad investigadora por William Cohen (CMU).
· Son datos reales. Algunos mensajes y cabeceras han sido eliminados. Hay muchas inconsistencias.
· Información contenida: remitente, destinatarios, asunto, cuerpo del texto, fecha y hora de envío y otras cabeceras.

Base de datos de Enron

· Jaffar Adibi ha solucionado algunos problemas de integridad de los mensajes y eliminado inconsistencias (direcciones de correo-e duplicados, mensajes corruptos y repetidos, carpetas no válidas).
· Ha creado una base de datos MySQL y ha analizado estadísticamente los datos, haciendo hincapié en la distribución de mensajes por empleados, a lo largo del tiempo, etc.
· Estudio de la red social de los 151 empleados teniendo en cuenta la posición dentro de la empresa (Shetty&Adibi).

Emparejando temas, personas y mensajes

· Generamos las *signatures* para cada *cluster* (temas).
· Generamos las *users' signatures* a partir del conjunto de mensajes de cada persona.
· Generamos la *message signature* para cada mensaje individual.

Representando cada *signature* como un vector de términos y calculando la similitud entre vectores, podemos buscar los temas tratados por cada persona, agrupar personas por conversaciones, mensajes, etc.

Temas tratados

jeff.skilling@enron.com (director general)
topic:302 sim:0.69 curv valid basy file gd price gdy map omicron vol gdp region phy enron trader
topic:219 sim:0.50 staf temporary project satisfy team qual servic utiliz objectiv altern evalu user implement determin research
topic:290 sim:0.47 msn explorer download free intl http messeng web asp hey mountain guz hope y'all attach
topic:133 sim:0.33 docu ljm investig litig defend attach feder court partnership relat officer observat copy fby inception
topic:165 sim:0.32 talk ljm regenc hyatt relat party transact subject earn hetty time lot today quarter prevy
topic:813 sim:0.31 cute pictur sweety blazer finney divorc enigmat whataburg girlfriend nite verdict partak tractor mania didn
topic:50 sim:0.26 august qbr th july meet midstream stakehold remind tonn dressag held dave month septemb invoic
topic:221 sim:0.25 interview candid student evalu saturday facilit super intern decis resum recruit vinc morn summer dear
topic:334 sim:0.23 ercot frontera mw jmf qse enpower oom hour oomc north protocol zone offpeak dont balanc
topic:809 sim:0.23 knew nevermind told pappa mavrix yergin sedy karaok hagler about mr kitten stephen thrill stormy

Temas tratados

james.derrick@enron.com (abogado)
topic:73 sim:0.96 children museum volunt championship enron jumper family kid uscaa walk weav celebr bear antioch art
topic:355 sim:0.26 linda robertson guinn sheinkman joshua dc edt wyden ask introductory barton adrian aggressiv mit nosk
topic:264 sim:0.14 review process perform doorstep satisfactory prc feedback mid form pep final attach year end action
topic:448 sim:0.13 rest score week wolf tiebreak arizona joey scare lenhart charit win office novesak shmuel accuracy
topic:611 sim:0.10 photo camera photowork pictory pate birth album scrapbook spillway preload photograph roster pictur easiest emonster
topic:261 sim:0.08 migrat critic informat february ubsw variety circumst environ tonight hardwar applicat tuesday commun due employe
topic:618 sim:0.08 folk newsom work ee deregulatory enron lobbyist puc cftc case commiss bev southsid samerican brooksley
topic:830 sim:0.07 kill deal zero qu liquid emw tweety o'kane kock housecat y flack primarly enter missiv
topic:790 sim:0.05 sacramento mcloughlin spec soto pownal sb ab bryce doh tax iepa corbett energy baxter confer
topic:758 sim:0.05 mtm exposur mseb reconcily accrual posit fma merchant dpc book efy asset exopsur hedg acctg

Temas tratados

benjamin.rogers@enron.com (--)
topic:244 sim:0.98 comment review attach draft propos incorpor rule red reflect file orjan cantrel bila final icc
topic:531 sim:0.23 kathy newslett michel grabstald vitrella remind person frevert mehrer weekly electron sheppard assur provid enside
topic:16 sim:0.22 guy bright friend good talk attach enron nofx kennebec willamett df pretty hartman stay last
topic:250 sim:0.13 richard sander pdf ryy litig costigan assembly leader shapiro staff ssharma republican iiy arbitrat ann
topic:736 sim:0.10 brian hoskin mobil broadband coastal net office fax perron exxon lightrad jpg ce redmond riley
topic:74 sim:0.09 trade product legal document financy platform line physic trader merril envera power fitzpatrick applicabl condit
topic:79 sim:0.09 kaminsky vinc mr dear dr professor wincenty ecth crenshaw ludmila kohly fujita kudin francesca both
topic:662 sim:0.06 yesterday nader version connector mention today uptower semmoto sachio gothorugh enronrefund datacentr curmley crossaint found
topic:370 sim:0.06 enron sa na home check busy eye feast deduct public java lobby innovat plaza deposit
topic:510 sim:0.06 pastoria michigan kalamazoo mm lv freeman calger antelop writeoff bulletproof frame chamber primarily propsal pauld

De temas a proyectos

MODELADO DE TEMAS

· Crear *clusters* de temas tratados (*expertise clusters*)

· Identificar anomalías e inconsistencias

· Crear el modelo de temas tratados por la organización

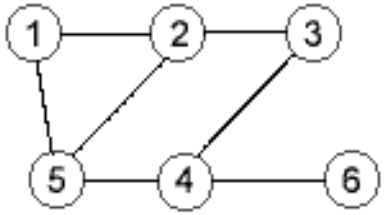
MODELADO DE PROYECTOS Y TAREAS

· Analizar los diálogos dentro de cada tema

· Modelar la evolución de las tareas y los proyectos a lo largo del tiempo

Cliques

En teoría de grafos, una clique en un grafo no dirigido G, es un conjunto de vértices V tal que para cada dos vértices cualesquiera, existe una liga que los conecta.



OBJETIVO: encontrar los cliques máximos dentro de la red asumiendo que son «comunidades» y estudiar el contenido de su correspondencia.

PROBLEMA: Encontrar cliques máximos es un problema NP-Completo.

Cliques identificados

Clique de 2 nodos (50 mensajes):

albert.meyers@enron.com (empleado), bill.williams@enron.com (--)

Clique de 9 nodos (1404 mensajes):

greg.whalley@enron.com (presidente), a..shankman@enron.com (presidente),
andy.zipper@enron.com (vicepresidente Enron OnLine),
jeff.skilling@enron.com (CEO), john.arnold@enron.com (vicepresidente),
kenneth.lay@enron.com (CEO), liz.taylor@enron.com (--),
louise.kitchen@enron.com (presidente Enron OnLine), sally.beck@enron.com
(empleada, Oficina de Operaciones)

Métricas aplicadas a cliques

1. *Size*: $S =$ número de nodos
2. *Clique Strength*: $CS =$ número de mensajes total para los nodos
3. *Average Clique Strength*: fuerza media para cada nodo
   $ACS = \frac{CS}{N}$
4. *Link Strength*: $LS(i,j) =$ número de mensajes intercambiados entre dos nodos i y j
5. *Average Link Strength*: LS media para cada nodo i
   $ALS(i) = \frac{\sum_i^{N-1} LS(i,j)}{N-1}$
6. *Centrality*: diferencia entre la ALS de un determinado nodo y la ACS
   $C(i) = ALS(i) - ACS$

Ideas para el artículo

OBJETIVO: podemos deducir temas tratados por individuales y grupos a partir del contenido del correo-e

- Objetivo*
- Background*
- Trabajo relaciona
- Temas tratados por individuales
  - *Signatures* para correo entrante y saliente. Diferencias.
- Grupos
  - *Clustering*
  - *Clustering* de las *signatures*
- Dinámica de grupos: cliques
- Resultados*

- Idea
- Trabajo realizado
- Evaluación ??
- Interpretación

Referencias

· Austin, J. L. *How to do things with words*. 1962.

· Bach & Harnish. Linguistic Communication and Speechs Acts. MIT Press. 1979.

· Klimt, B. & Yang, Y. "Introducing the Enron Corpus", CEAS-04, http://www.ceas.cc/papers-2004/168.pdf

· Lin, C-Y. & Hovy, E. "The automated adquisition of topic signatures for text summarization", COLING-00

· Pantel, P. & Lin, D. "Document Clustering with Committees", SIGIR-00.

· Shetty, J. & Adibi, J. "The Enron Email Dataset. Database schema and brief statistical report", http://www.isi.edu/~adibi/Enron/Enron_Dataset_Report.pdf

MKIDS: identificación de roles, grupos de inclusión y proyectos

Tom Murray & Ed Hovy
Information Sciences Institute
University of Southern California
{yohzik, hovy}@isi.edu

Víctor Peinado
Grupo de PLN y Recuperación de Información de la UNED
victor@lsi.uned.es

NLP Group

14 de enero de 2005