Concepts, Semantics and Syntax in E-Discovery

David Eichmann and Si-Chi Chin
Institute for Clinical and Translational Science
The University of Iowa
Iowa City, Iowa
{david-eichmann,si-chi-chin}@uiowa.edu

Abstract

The Legal Track in the annual TREC evaluation has involved a number of information retrieval researchers in the problem of E-discovery. Significant challenges are present in this domain that are not common to other domains that these researchers address. In particular, noisy metadata and document collections comprised solely of OCR require new techniques. We relate some of our recent work in exploring the Legal Track document collection.

1 – Introduction

In a rapidly-evolving sea of information, much of which is increasingly informal and ephemeral, traditional approaches to discovery begin to founder. The basic notion of a 'document' is open to interpretation as correspondence evolves from the traditional forms of letters and memoranda to wikis, blogs and RSS feeds. Search and retrieval in this context must address a broad diversity of language style, ranging from formal academic writing to abbreviated instant messages. A letter can contain scientific argument or mundane narratives. Labeling documents by format (e.g. emails, memo, news, etc.), a common practice to classify content, over-simplifies the complexity of modern correspondence.

Our approach to this problem is three-fold. We first are analyzing the human-generated metadata available for document collections with the intent in using it as ground-truth for the organizational and individual interactions captured therein. We are then exploring the syntactic and semantic nature of document content with the goal of automatic generation of metadata, allowing for reduction in the human effort required to increase the utility of such document collections. Finally, we are exploring the mapping of document vocabulary to concepts with the goal of creating clusters of documents that correspond to the current practice of boolean predicate specification of discovery criteria. This paper presents a snapshot of our current status, primarily in the area of metadata analysis.

2 – Analysis of the CDIP Metadata

The Illinois Institute of Technology Complex Document Information Processing Test Collection (IIT CDIP) version 1.0 comprises 6,910,192 documents released as part of a tobacco master settlement agreement [1, 5]. Table 1 shows the various metadata entities and their frequencies. (Note that our load of the data resulted in only 6,787,322 document records. The statistics we show from here are based upon this document count, and not the IIT document count.) Available metadata include

- attendees at meetings (with no annotation distinction between organizations and persons),
- organizations authoring, mentioned in and receiving documents
- persons authoring, mentioned in and receiving documents,
- brands of cigarettes involve, and
- evidentiary annotations (e.g., type, category, Bates numbers and production boxes).

The XML content comprises two distinct releases of the metadata, with substantial, but not total, overlap - referred to by their containing tags, '<A>' and '<LTDLWOCR>' [6]. Typical document collections used in system evaluations possess metadata that while minimal in nature (e.g, docno and title), is reasonably authoritative and error-free. This proves not to be the case for the CDIP data, as we show below. Unlike the data for various evaluations such as the Text Retrieval Conference (TREC), these metadata were manually generated by seven distinct organizations over a very large and extremely noisy base collection of physical documents. This leads to the need to address the CDIP metadata as a cor-

Table 1: Metadata Entity Frequencies

Metadata Entity	Occurrences			
	Total	Distinct	Avg./ Entity	Avg./Doc.
attendee	65,691,473	49,375	1,330.46	9.68
bates	9,476,794	8,054,075	1.18	1.40
brand	26,498,001	155,350	170.57	3.90
category	13,594,494	74	183,709.38	2.00
copied	8,775,307	322,294	27.22	1.29
doctype	18,359,644	2,501	7,340.92	2.70
document	6,787,322	n/a	n/a	n/a
organization author	8,742,976	149,641	58.43	1.29
organization mentioned	31,406,753	883,285	35.56	4.63
organization receiving	8,262,496	63,625	129.86	1.22
person author	11,128,029	875,292	12.71	1.64
person mentioned	34,683,289	1,938,310	17.89	5.11
person receiving	23,427,415	455,404	51.44	3.45
prodbox	6,830,993	6,306	1,083.25	1.01

Table 2: Top 20 Document Types with 'P' Tag

Document Type	Occurrences
MEMO, MEMORANDUM	492,983
LETT, LETTER	383,074
EMAI, E-MAIL	309,637
FORM, FORM	279,880
REPT, REPORT, OTHER	266,428
Memo	214,912
LETTER	162,993
NOTE, NOTE	127,767
TELE, TELEX	118,607
SCRT, REPORT, SCIENTIFIC	104,661
SPEC, SPECIFICATION	102,991
Letter	93,134
ZFLB, FILE FOLDER BEGIN	83,274
CHAR, CHART, GRAPH, TABLE, MAPS	82,572
ADVE, ADVERTISEMENT	78,837
Smoke/Tobacco Analysis	68,521
REPT, OTHER REPORT	65,186
COMP, COMPUTER PRINTOUT	63,795
SCIENTIFIC ARTICLE	54,544
Report	51,867

Table 3: Top 20 Document Categories

Category	Count
SMOKING BY-PRODUCTS	3,355,453
YOUNG ADULT SMOKING	2,102,562
IN-HOUSE RESEARCH ON SMOKING & HEALTH	746,920
SPORTING AND ENTERTAINMENT	541,356
SAFER CIGARETTE	471,960
NICOTINE AND ADDICTION	463,713
CTR/TIRC/TI	438,959
TRADE ASSOCIATIONS OTHER THAN CTR/TI	394,933
SMOKING AND CANCER	385,826
ADVERTISEMENTS	358,706
ETHNIC MARKET	358,269
SMOKING AND LUNG CANCER	327,286
NICOTINE CONTROL	255,748
ELECTRONIC MEDIA	249,820
MARKETING STRATEGY	235,310
BIOLOGICAL ACTIVITY OF CIGARETTES	219,460
SMOKING AND HEART DISEASE	204,503
ETS IN-HOUSE RESEARCH	195,164
SPONSORED RESEARCH ON SMOKING & HEALTH	163,051
MARKETING RESEARCH	151,469

pus in and of itself. Given our interest in exploring the connections among individuals and organizations revealed by these data, our first step was to assess the 'cleanliness' of the various categories and how we might unify potential variations in entity reference. The structure of the XML content itself is challenging to authoritatively process. Many of the fields contain semicolon delimited lists of values, rather than subelements. There are also substantial number of elements where there are no delimiters between values - the person author metadata contains 200+ entries that are longer than 200 characters each.

As shown in Table 2, the lack of a clear document taxonomy used when tagging the documents results in a need to either account for the variation when running queries, or a need to post-process the data to merge equivalent type labels into a single class. We have opted for the latter. One interesting point for this metadata category is that there is an average of 2.7 types per document. Even allowing for non-overlap of A and LTDLWOCR type labels, it is relatively common for a document to have more than one type.

Our interest extends particularly to the boundary region between the scientific literature and the corporate usage or reaction to that literature. The CDIP collection contains roughly 160,000 scientific documents, making it a substantial collection in its own right which is unfortunately fogged with the haze of optical character recognition (OCR) noise.

Table 3 list the most frequent document categories. Unlike document type, it appears to be a reasonably uniform multi-valued metadata category (and has no extraordinarily long entries). Table 4 lists the most frequently occurring brands. (While we currently have no plans to employ brand information, we include it here for those who might be

Table 4: Top 20 Brands Mentioned

Brand	Count
MARLBORO	1,489,291
CAMEL	1,176,219
WINSTON	1,093,368
SALEM	767,268
DORAL	685,204
NEWPORT	594,688
MERIT	588,789
VIRGINIA SLIMS	569,660
BENSON & HEDGES	551,798
KENT	501,915
KOOL	501,096
VANTAGE	493,844
TRUE	414,000
PARLIAMENT	346,520
NOW	336,157
VICEROY	304,354
CAMBRIDGE	289,722
MORE	273,574
NON-RJR BRANDS	273,354
OLD GOLD	268,489

Table 5 lists the more frequently listed meeting attendees. As is easily seen, person and organizational types have been conflated in this category, as is also the case for those copied on documents. Table 6 shows the entities most frequently copied on documents. We plan on separating individuals and organizations in these two categories by using the organizational author, mentioned and receiving data to lift organization entities out of the copied and attendees table, and then manually inspect at least the high frequency remaining values for residual organizations not occurring in the three source tables. We are basing this approach on the observation that the three organizational categories are

Table 5: Top 20 Attendees

Attendee	Count
PM, PHILIP MORRIS	744,207
RJR, R.J.REYNOLDS	477,257
LEO, LEO BURNETT AGENCY	361,117
ZZERO	270,947
BAT, BRITISH AMERICAN TOBACCO	268,877
LOR, LORILLARD	263,308
ZZWRE	262,980
BW, BROWN & WILLIAMSON	256,586
ZZUHA	217,769
PHILIP MORRIS	215,157
REININGHAUS,W	189,380
WALK,RA	188,153
ZZGRA	171,487
VANHARN,J	163,635
HUMMEL,GH	159,979
GANTEAUME,H	153,545
GALLAHER	152,373
ZZRWA	149,426
ROEMER,E	145,943

Table 6: Top 20 Copied

Copied	Count
RJR	79,018
B&W	65,994
UNK	57,283
RSS	30,651
RDC	26,499
SELIGMAN,RB	26,159
JFA	25,668
PRC	24,474
PHL	21,006
RESNIK,FE	19,788
HAYES AW	19,444
DRB	18,433
DIMARCO GR	17,907
WAKEHAM,H	17,754
HARDIN BV	17,158
RODGMAN A	17,115
Riehl-T	17,057
OSDENE,TS	16,782
SPEARS,AW	16,762
WRT	16,083

substantially more pure than the corresponding person categories.

Table 7 lists the most frequent authoring organizations. Table 8 lists the organizations most frequently mentioned in documents. Table 9 lists the organizations most frequently receiving generated documents. We are particularly interested that three of the top ten most frequent organizational authors and organizational receivers are research groups, rather than corporate entities.

Table 10 lists the most frequently authoring persons, Table 11 the most frequently mentioned persons, and Table 12 the persons most frequently receiving generated

Table 7: Top 20 Organizational Authors

Organizational Author	Count
PM, PHILIP MORRIS	1,247,507
LOR, LORILLARD	748,586
INBIFO, INSTITUT FUR BIOLOGISCHE FORSCHUNG	491,102
PMUSA, PHILIP MORRIS USA	243,484
LEO, LEO BURNETT AGENCY	186,912
TI, TOBACCO INST	112,619
PMMC, PHILIP MORRIS MANAGEMENT CORP	103,389
PHILIP MORRIS	99,179
CONTRACT RESEARCH CENTER	88,625
RJR, R.J.REYNOLDS	64,092
SHB, SHOOK,HARDY & BACON	61,149
HAZLETON, HAZLETON LABS	37,537
PME, PHILIP MORRIS, EUROPE	36,863
FTR, FABRIQUES DE TABAC REUNIES S.A.	36,840
BORR, BORRISTON LABS	36,282
PMI, PHILIP MORRIS INTERNATIONAL	33,063
BW, BROWN & WILLIAMSON	32,752
MICRO, MICROBIOLOGICAL ASSOCIATES	32,682
COVINGTON BURLING	30,941
ASSOCIATED PRESS	29,700

Table 8: Top 20 Organizations Mentioned

Organization Mentioned	Count
FTC, FEDERAL TRADE COMMISSION	572,777
RJR, R.J.REYNOLDS	510,079
BW, BROWN & WILLIAMSON	349,898
FDA, FOOD AND DRUG ADMINISTRATION	313,992
PHILIP MORRIS	292,739
PM, PHILIP MORRIS	276,230
PMUSA, PHILIP MORRIS USA	267,660
CONGRESS	253,994
EPA, ENVIRONMENTAL PROTECTION AGENCY	241,294
TI, TOBACCO INST	212,649
AMER, AMERICAN TOBACCO	197,508
INBIFO, INSTITUT FUR BIOLOGISCHE FORSCHUNG	184,492
PM	174,758
LOR, LORILLARD	151,272
OSHA, OCCUPATIONAL SAFETY & HEALTH ADMINISTRATION	145,198
LEO, LEO BURNETT AGENCY	135,061
LIG, LIGGETT	131,513
LM, LIGGETT & MYERS	126,571
HHS, DEPT OF HEALTH AND HUMAN SERVICES	106,763
CTR, COUNCIL FOR TOBACCO RESEARCH	106,519

documents. As noted above, these three categories are substantially polluted with organizational entities, and require clean-up to properly analyze individual-to-individual information interchanges.

3 – Assessing the Involvement of an Individual

Consider one of the high frequency entities, Wolf Reininghaus, General Manager for Contract Research, Institut fur Biologische Forschung (INBIFO), Cologne, Germany (Philip Morris' offshore research lab) according to tobaccodocuments.org [7]. Reininghaus is the top individual

Table 9: Top 20 Receiving Organizations

Receiving Organization	Count
PM, PHILIP MORRIS	1,854,109
LOR, LORILLARD	1,120,713
PMUSA, PHILIP MORRIS USA	374,540
INBIFO, INSTITUT FUR BIOLOGISCHE FORSCHUNG	161,398
PMMC, PHILIP MORRIS MANAGEMENT CORP	158,918
FTR, FABRIQUES DE TABAC REUNIES S.A.	140,535
PHILIP MORRIS	105,941
LEO, LEO BURNETT AGENCY	88,289
SHB, SHOOK,HARDY & BACON	84,152
TI, TOBACCO INST	71,047
CRC	66,407
PME, PHILIP MORRIS, EUROPE	65,217
PHILIP MORRIS COMPANIES INC	51,639
CONTRACT RESEARCH CENTER	49,616
RJR, R.J.REYNOLDS	49,327
PMI, PHILIP MORRIS INTERNATIONAL	48,854
APO, ARNOLD & PORTER	45,828
COVINGTON BURLING	36,481
BW, BROWN & WILLIAMSON	34,715
FISH NEAVE	34,139

Table 10: Top 20 Person Authors

Person Author	Count
RJR	462,280
B&W	147,606
CTR	58,581
UNK	470,25
WILLIAM ESTY	30,136
TI	26,670
STEVENS,AJ	25,617
REININGHAUS,W	23,880
WALK,RA	20,072
BW	19,930
CHAIKIN,K	19,602
CRICHTON,JS	17,959
MARKETING DEVELOPMENT DEPT	17,813
HOYT WT	17,736
SMITH,KV	16,939
MERLO,E	16,642
SELIGMAN,RB	15,715
RJR INTL	15,597
WAKEHAM,H	14,711
RJR NABISCO	14,519

listed as a meeting attendee (11th overall after 10 organizations), third most frequent author (8th overall after 5 organizations), and 23rd most frequent recipient of documents. Searching the metadata using his surname as a prefix yields the data shown in Table 13.

Clearly, the possibility exists that many of these person name occurrences relate to other people. However, even accounting for that, there are multiple variations of his name in the metadata, and some form of unification is required to achieve a high level of recall of documents involving this – and any – individual. We are currently considering how to

Table 11: Top 20 Mentioned Persons

Person Mentioned	Count
RJR	897,308
SURGEON GENERAL	315,604
PHILIP MORRIS	186,594
FTC	166,758
X	90,530
NATL FAMILY OPINION	85,191
TI	82,407
B&W	76,958
LORILLARD	76,588
FDA	66,051
RJR INTL	65,255
CTR	57,976
LIGGETT	54,998
EPA	51,031
CLINTON	47,835
ACS	44,618
CORESTA	44,612
AMES	41,390
SAB	31,945
ECUSTA	30,848

Table 12: Top 25 Receiving Persons

Receiving Person	Count
RJR	351,694
B&W	171,999
UNK	147,267
MERLO,E	87,816
OSDENE,TS	63,595
DIMARCO GR	53,215
HAN,V	48,578
SPEARS,AW	46,484
LAUFER,D	41,414
PARRISH,S	39,735
SELIGMAN,RB	39,032
WOODWARD,E	38,885
DARAGAN,K	38,842
CTR	38,229
CHAIKIN,K	38,124
MCCORMICK,B	37,200
ATCO	35,305
KEANE,D	34,755
LEVY,C	34,430
SPECTOR,J	34,385
DESEL,P	33,450
BERLIND,M	32,979
REININGHAUS,W	32,764
CARRARO,T	31,728
BW	31,493

avoid manual inspection of referent documents to resolve person identity for low-frequency candidates, but currently have no feasible options.

4 – Inferring Organizational Networks

The organizations and persons mentioned metadata are interesting categories in that they are not directly part of the

Table 13: Occurrences of Persons with Names Beginning with "REININGHAUS"

Person	# of Occurrences as			
	Attendee	Author	Receiver	Mentioned
REININGHAUS	7,337	200	1,974	2,837
REININGHAUS W		6		2
REININGHAUS,				1
REININGHAUS, WOLF			6	
REININGHAUS,A		2		
REININGHAUS,AV		1		
REININGHAUS,B	196			2
REININGHAUS,D				1
REININGHAUS,E	530	2	9	10
REININGHAUS,F		1	1	
REININGHAUS,G		1	2	
REININGHAUS,GJ			1	
REININGHAUS,H	2	2	15	1
REININGHAUS,I			10	1
REININGHAUS,M		1		
REININGHAUS,MW			5	
REININGHAUS,OW			1	
REININGHAUS,R		17	144	12
REININGHAUS,RA		3		
REININGHAUS,RNW		1		
REININGHAUS,RW			4	
REININGHAUS,S			1	2
REININGHAUS,T		3	2	
REININGHAUS,U			4	
REININGHAUS,VW		3		
REININGHAUS,W	189,380	23,880	32,764	16,152
REININGHAUS,WA		5		
REININGHAUS,WD			1	5
REININGHAUS,WE		17		5
REININGHAUS,WG		2		
REININGHAUS,WK		1		
REININGHAUS,WP			5	1
REININGHAUS,WR		2		6
REININGHAUS<REININGHAUS,W>		4		
REININGHAUSE				4
REININGHAUSEN				9
REININGHAUSER				4
REININGHAUSS				1
REININGHAUS_WOLF			6	
REININGHAUS_WOLF_1			1	
Reininghaus, Wolf			115	

document lifecycle, but are rather semantic-driven channels in the content itself. They are the only two categories that, at least for high frequency entries, involve other entities than those involved in the litigation.

Continuing our use of Wolf Reininghaus as an example, consider Table 14, which shows the highest co-mentions of persons with Reininghaus. For each of the top four individuals co-mentioned with Reininghaus, we list each of their top ten co-mentions. Interrogating [7] we find the affiliations shown in Table 15. Within this strongly interconnected co-mention cluster, we have eight employees of INBIFO, two employees of Philip Morris, and seven unknowns. Interrogating the person author/receiving data, we find 2001 instances of communication from Reininghaus to Walk and 1602 from Walk to Reininghaus, fewer than the number of their co-mentions in the document collection.

5 – Semantics and Structure

Our previous work in regenerating the logical structure of a document from PDF or HTML for scientific papers [2, 4] leads us to believe that we can achieve useful performance over the CDIP collection as well. Our approach is highly dependent upon token location on the page, which means that we may go back to the scanned images to acquire coordinate data. We see this as a substantial, if computationally

Table 14: Co-mention Connections Among Reininghaus' Co-mentions

Reininghaus		Walk		Roemer		Haussmann		Tewes	
Co-mentions	Count	Co-mentions	Count	Co-mentions	Count	Co-mentions	Count	Co-mentions	Count
WALK,RA	3,871	REININGHAUS,W	3,871	REININGHAUS,W	3,716	REININGHAUS,W	3,293	REININGHAUS,W	2784
ROEMER,E	3,716	ROEMER,E	2,883	WALK,RA	2,883	WALK,RA	2,799	TEREDESAI,A	2746
HAUSSMANN,HJ	3,293	HAUSSMANN,HJ	2,799	HACKENBERG,U	2,623	ROEMER,E	2,573	VONCKEN,P	2483
TEWES,F	2,784	HACKENBERG,U	2,360	HAUSSMANN,HJ	2,573	VONCKEN,P	2,323	OEY,J	2476
STINN,W	2,697	STINN,W	2,149	TEWES,F	2,387	STINN,W	2,257	MEISGEN,T	2424
HACKENBERG,U	2,635	TEREDESAI,A	1,935	STINN,W	1,970	TEREDESAI,A	2,248	ROEMER,E	2387
VONCKEN,P	2,584	TEWES,F	1,931	TEREDESAI,A	1,926	TEWES,F	2,064	STINN,W	2384
TEREDESAI,A	2,425	VONCKEN,P	1,788	VONCKEN,P	1,810	SCHEPERS,G	1,910	RUSTEMEIER,K	2251
RUSTEMEIER,K	2,278	KUHN,D	1,728	HOLT,KV	1,756	RUSTEMEIER,K	1,569	KINDT,R	2246
SCHEPERS,G	2,197	SCHAFFERNICHT,H	1,493	SCHEPERS,G	1,725	HOLT,KV	1,546	HAUSSMANN,HJ	2064

Table 15: Reininghaus' Co-mention Affiliations

Person	Affiliation
Reininghaus, Wolf	(Gen. Mgr., Contract Research, INBIFO) Institut fur Biologische Forschung, Cologne, Germany - (Philip Morris' offshore research lab).
Walk, Rudiger-Alexander	(Director of Human Studies, PM Richmond c. 2000) Joined PM in April 1978. Was Dir. of Scientific Affairs, PM USA
Roemer, Ewald	(INBIFO)
Haussmann, Hans-Jurgan	(Associate Prinicpal Scientist, PM c. 1997)
Tewes, F.	(Biologist at INBIFO c. 1987)
Stinn, W	?
Hackenberg, Ulrich	(INBIFO)
Voncken, P.	(INBIFO Chemist (Diplomchemiker))
Teredesai, A.	(Pathologist, INBIFO, c. 1987)
Rustemeier, K	?
Schepers, G	?
Kuhn, D.	(Study Director (mouse skin painting) INBIFO c. 1987)
Schaffernicht, Helmut	(INBIFO Head of Engineering)
Holt, K V	?
Oey, J	?
Meisgen, T	?
Kindt, R	?

intensive, advantage in the automatic generation of metadata, as much of the collection is of similar structure, or even more constrained in structure (e.g., letters and memoranda).

The variable quality of the OCR, on the other hand, presents significant difficulties in are detection of organization and person mentions in document content. Our existing tools for this rely heavily on part-of-speech tagging models that were trained on error-free newswire material. The high level of noise in the OCR output leads to numerous false positives on proper nouns (the default for an out-of-vocabulary word).

This also proves to be the case for our information extraction framework, which uses the syntactic structure of sentence parse trees to identify entities and concepts and the relationships between them. The usually straight-forward preprocessing step of sentence boundary detection struggles with the problem of easily missed punctuation in the OCR scan. Figure 1 contains a scatter plot of the number of nodes

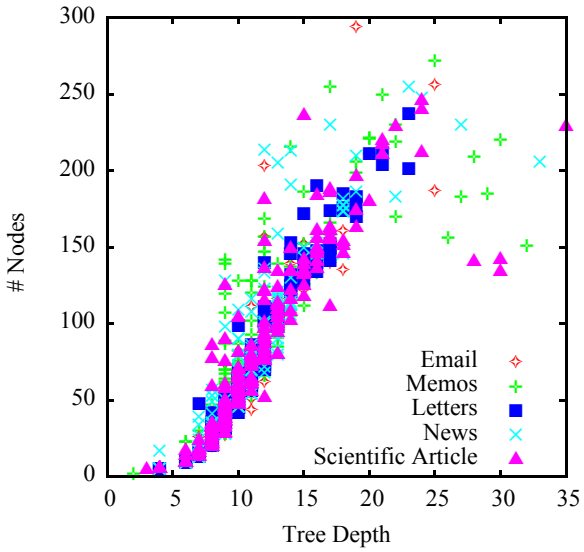

Figure 1: CDIP Parse Tree Complexity

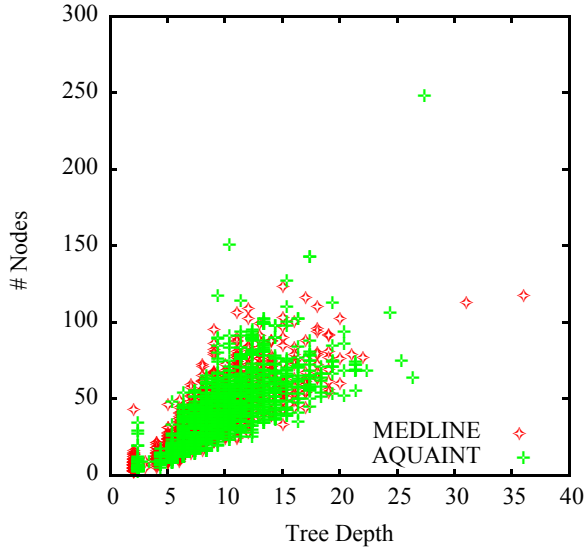

Figure 2: Clean Text Parse Tree Complexity

(roughly the number of words and containing phrases) in a sentence and the depth of the corresponding parse tree for a small sample from five document categories. Note that there

are detected sentences with extremely large numbers of nodes - roughly equivalent to the number of words on a full page of double spaced text. This is due to noise in the OCR that breaks words into fragments (thereby increasing the word count) and missed small punctuation (e.g., periods), fusing two or more sentences into one.

In comparison, Figure 2 shows two 'normal' sources - MEDLINE is a sample of 1000 sentences drawn from abstracts drawn from the NIH citation database and AQUAINT is a sample of 1000 sentences drawn from a collection of newswire stories used in various TREC evaluations. There are substantial differences for the OCR data - in parses with relatively high numbers of nodes, typical depth of trees (even when compared to science abstracts) and variability in tree depth at a given number of nodes. These differences are further accentuated when the outlier data point for AQUAINT at just less than 250 nodes is discounted - this is a pseudo-sentence generated by processing a table of sports scores. We are currently considering possible approaches to address OCR sentence parsing issues.

6 – Concept Recognition

This aspect of our work is intended to complement our named entity approach from the previous section. There are many situations where named entities (e.g., organizations and individuals) are not present or irrelevant to a particular information need. This can be interpreted as projecting from proper noun phrases to general noun phrases and their underlying semantics. Consider the term 'emphysema' as an example. In terms of the CDIP collection, academic characterizations of the concept occur not only in scientific journal articles and reports, but also in letters and memorandum. However, current information retrieval techniques are not typically capable of clustering sentences such as these two:

> Our fundamental approach to the understanding of pulmonary emphysema is to recognize, as Mitchell and others have emphasized clinically, that there are two main types, the panlobular type in which there is a primary destruction of the alveolar lobules and the absence of bronchitis. [docid=aee02a00,LETTER]

> The 1961 death certificates they found, show a frequent association of tuberculosis and emphysema, regardless of whether these diseases were primary or secondary causes of death. [dicid=ach1aa00, SCIENTIFIC ARTICLE]

together based on their language styles. Even when sharing the same format, some documents are more informative than the others. Considering the following sentences:

> Have two print ads-- Cowboys / emphysema and Mind if I smoke / care if I die? [docid=bwj36c00, EMAIL]

> What kind of company would continue to develop glitzy new marketing campaigns for cigarettes when "there is overwhelming medical and scientific consensus that cigarette smoking causes lung cancer, heart disease, emphysema and other serious diseases?"-- Is Philip Morris being equally 'candid' overseas? [docid=byc48c00, EMAIL]

They are both email messages from the collection and both mention the concept of interest, but the second sentence clearly conveys a more 'scientific tone' than the loosely structured first sentence. We could substantially facilitate information retrieval if we were able to distinguish an informative piece of text belonging to a desired document genre.

We look to concept recognition to identify a more abstract semantic meaning of a named entity. In the TREC genomics track, some groups have mapped specific terms such as "saggital section" or "transverse section" to a more abstract concept "sectional anatomy" [3]. Such a method adopts a bottom-up approach: mapping from leaves to root nodes in a knowledge taxonomy tree. In our current work, we intend to adopt a top-down approach to explore the concept. That is, given a specific concept such as "emphysema," we would like to learn how the concept is described. We also presume that a concept would be perceived differently according to its various target audience. Thus we would like to discover the useful context patterns supporting such a difference.

The following examples illustrate how "emphysema" is perceived differently even though they are expressing similar ideas:

> Emphysema develops when many of the small air sacs in the lungs are destroyed.

> Emphysema is a type of chronic obstructive pulmonary disease (COPD) involving damage to the alveoli.

Intuitively, from the wording and the style, we may predict that the first sentence is more likely from a news source and the second one from a scientific journal article, since we found more rarely used medical terms in the second one. Inversely, we may also assume that given two sentences from two different sources, the same referred concept would have different interpretations. Our goal here is to identify and model the context for each category, and be able to detect the shift of the context.

7 – Conclusions

As shown here, much of the information present in the CDIP collection is noisy and of varying reliability for high precision / high recall retrieval. We believe that contextualization will be beneficial in providing IR-based

constraints that map to E-discovery predicate clauses. Rather than treating each context feature individually, our future work will combine metadata related to named entities and build up a social network analysis to approach the dataset. Because the OCR performance over the collection is unreliable, it is difficult to model the collection by using only the OCR results. Our hypothesis in shifting into the social network domain is that similar, but OCR-problematic, documents will also demonstrate similar social network characteristics. Hence our goal is to be able to locate an informative document despite its poor OCR performance.

References

[1] Baron, J, Lewis, D, Oard, D. "TREC 2006 Legal Track Overview," Fifteenth Text Retrieval Conference Proceedings (TREC2006), Gaithersburg, MD.

[2] Bradshaw, S, Light, M. and D. Eichmann, "(Bee)Dancing on the Boundaries Between PIM and GIM," SIGIR Workshop on Personal Information Management, Seattle, WA August 10-11, 2006.

[3] Caporaso, J. G., W. A. Baumgartner, Jr., K. B. Cohen, H. L. Johnson, J. Paquette and L. Hunter, "Concept recognition and the TREC Genomics tasks," Fourteenth Text Retrieval Conference Proceedings (TREC2005), Gaithersburg, MD.

[4] Eichmann, D., "Extraction of Document Structure for Genomics Documents," Fifteenth Text Retrieval Conference Proceedings (TREC2006), Gaithersburg, MD.

[5] Lewis, D., Agam, G., Argamon, S., Frieder, O., Grossman, D. and Heard, J., "Building a Test Collection for Complex Document Information Processing," SIGIR 2006, p. 665-666.

[6] Lewis, D. "Description of IIT CDIP v.1 / TREC Legal 2006 metadata (6-Jun-06 version)," email message to [TREC-legal] mailing list, June 6, 2006.

[7] http://tobaccodocuments.org/profiles/people/