
Challenges at the Intersection of Semantic Computing with Law, Legal Reasoning, and Legal Practice

Kevin D. Ashley

Professor of Law and Intelligent Systems
Senior Scientist, Learning Research and Development Center
University of Pittsburgh
ashley@pitt.edu

Intersection of semantic computing with law, legal reasoning, and legal practice...

- addresses the derivation and matching of the semantics of computational content
- to that of *naturally expressed user intentions*
- relating to legal problem-solving or analysis
- in order to retrieve, manage, manipulate or create content
- based on its significance to the legal problem-solving or analysis,
- where "content" includes text, video, audio, services, networks, etc.

Challenges relating semantics of computational content to users' intentions re legal problem-solving in:

- evidentiary e-discovery:
 - process multitude of e-documents to meaning and significance to *litigators' intentions* re clients' legal claims and strategies;
- business compliance:
 - relate computational version of normative rules to *regulators' intentions* embodied in natural language legal codes, principles, and policies;
- legal information retrieval:
 - relate computer-processable legal texts, indices, ontologies to *legal researchers' intentions* in retrieving/ drawing inferences with legal rules and cases.
- e-commerce and semantic web-based legal services:
 - relate computationally-accessible resources to the *intentions of electronically contracting parties*.
- automated management of privacy and IP rights in data:
 - relate proposed data access to *intentions of data rights owners and users*.
- e-government and legal education:
 - relate computer-processable argument diagrams to the meanings and *intentions of legal arguers*.

e-Discovery

- *Pretrial discovery in lawsuits:*
 - Processing parties' requests for/access to materials in hands of opponents and others to reveal facts and develop evidence for trial.
- *e-discovery:*
 - Collecting, exchanging, and analyzing electronically stored information or ESI in pretrial discovery.
- Large law suits routinely involve millions of e-documents.
 - extremely *heterogeneous* documents produced in litigation;
 - not only corporate memoranda and agreements, but
 - full panoply of email and other internet-based communications.
- In e-discovery context, challenge is to represent litigators' intentions:
 - litigators' hypotheses (or theories) about relevance, and to
 - relate documents they review to those intentions.

Focus on litigators' hypotheses about relevance

- *Relevance hypothesis*:
 - more-or-less abstract description of subject matter that, if found in a document, would make that document relevant to the law suit. (Hogan et al., 2009)
- Part of senior litigators' "sensemaking":
 - "process of collecting, organizing and creating representations of complex information sets, all centered around some problem they need to understand."
(Bauer et al., 2008)
- Relevance hypotheses based on complaints, document requests (RFPs) (e.g., in the TREC Legal Track):
 - "There are documents showing that the Vice President of Marketing knew that cigarette advertisements were targeted to children by 1989," or
 - "There exist documents which refer to *'in-store,' 'on-counter,' 'point of sale,' or other retail marketing campaigns for cigarettes,*" or
 - "There are documents that are communications between Alice and her lawyer Bob between 1985 and 1989," or more generally,
 - "There are documents *of a particular kind, satisfying particular time constraints, satisfying particular social interaction constraints, that refer to particular concepts or phrases of interest.*"

User modeling/cognitive task analysis for relevance hypotheses

- Elicit litigators' relevance hypotheses in iterative user modeling procedure:
 1. User communicates info need
 2. Proxy formulates relevance hypothesis
 3. Assessor formulates search in data source
 4. Retrieves samples of results
 5. Assesses if responsive
 6. Proxy presents hypothesis + evidence to user
 7. User confirms / denies hypothesis
 8. Proxy reformulates hypothesis and iterates to 2. (Hogan, et al. 2009; 2010)

- Process yields:
 - a. **“use case”**:
 - user's objectives given RFP; don't produce too much or too little
 - b. **scope of conceptual boundaries of interest**:
 - legal and other concepts relevant to case
 - c. **nuance**:
 - level of specificity of relevant concepts of interest;
 - d. **linguistic variability**:
 - “variety of ways a concept can be expressed, whether lexically or syntactically.”
 - (Bauer, et al., 2008; Hogan, et al. 2009; 2010)

Example: iterative user modeling of relevance hypotheses

- RFP: *Plaintiffs request Defendants produce...*
 - *103. All documents which describe, refer to, ... or mention any “in-store,” “on-counter,” “point of sale,” or other retail marketing campaigns for cigarettes.*
- → Info need → Initial hypotheses:
 - *Are promotional offers relevant?*
 - *Is “buy one get one free” by itself sufficient for relevance?*
- → Documents:
 - *“Promotion Description: Utilize self-contained shipper. Shipper has 40 buy-2-get-1-free sleeves...”*
- → Query to/response from user:
 - *“buy some get some free” to be included in the theory of relevance:*
 - *Sales promotions: Special offers: Buy some get some free offers*
- → Queries:
 - *buy w/2 get w/2 free OR b ?[1-9] ?g ?[1-9] ?f*
 - *regular expression matches patterns of form “b1g1f” w/ w/o spaces and w/ digits (1-9) or letters often substituted for digits under OCR (1f).*

Emerging techniques to relate computational content semantics to users' intentions in e-discovery

1. Social network analysis to supplement / apply models of users' relevance theories in document analysis and retrieval.
 - ❑ Infer relevance of document content based on who has them, who sent them, and who likely read them.
 - ❑ Document senders, recipients, owners identify themselves via email.
 - ❑ Model network of knowledge and use resulting structure to draw inferences.
 - ❑ How each item moves through social organization helps map general flow of knowledge.
2. Machine learning to extend / apply users' theories of relevance
 - ❑ Examples of documents classified according to theory of relevance
 - ❑ used to train an automated classifier to identify and classify other documents.
3. Hypothesis ontology to generalize user modeling regarding relevance theories
 - ❑ vocabulary of objects (agents) and processes,
 - ❑ associated with recurrent areas of interest in e-Discovery such as knowledge transmission in corporate or commercial settings;
 - ❑ define[s] the relationships in which these entities can participate.

Enron email social network

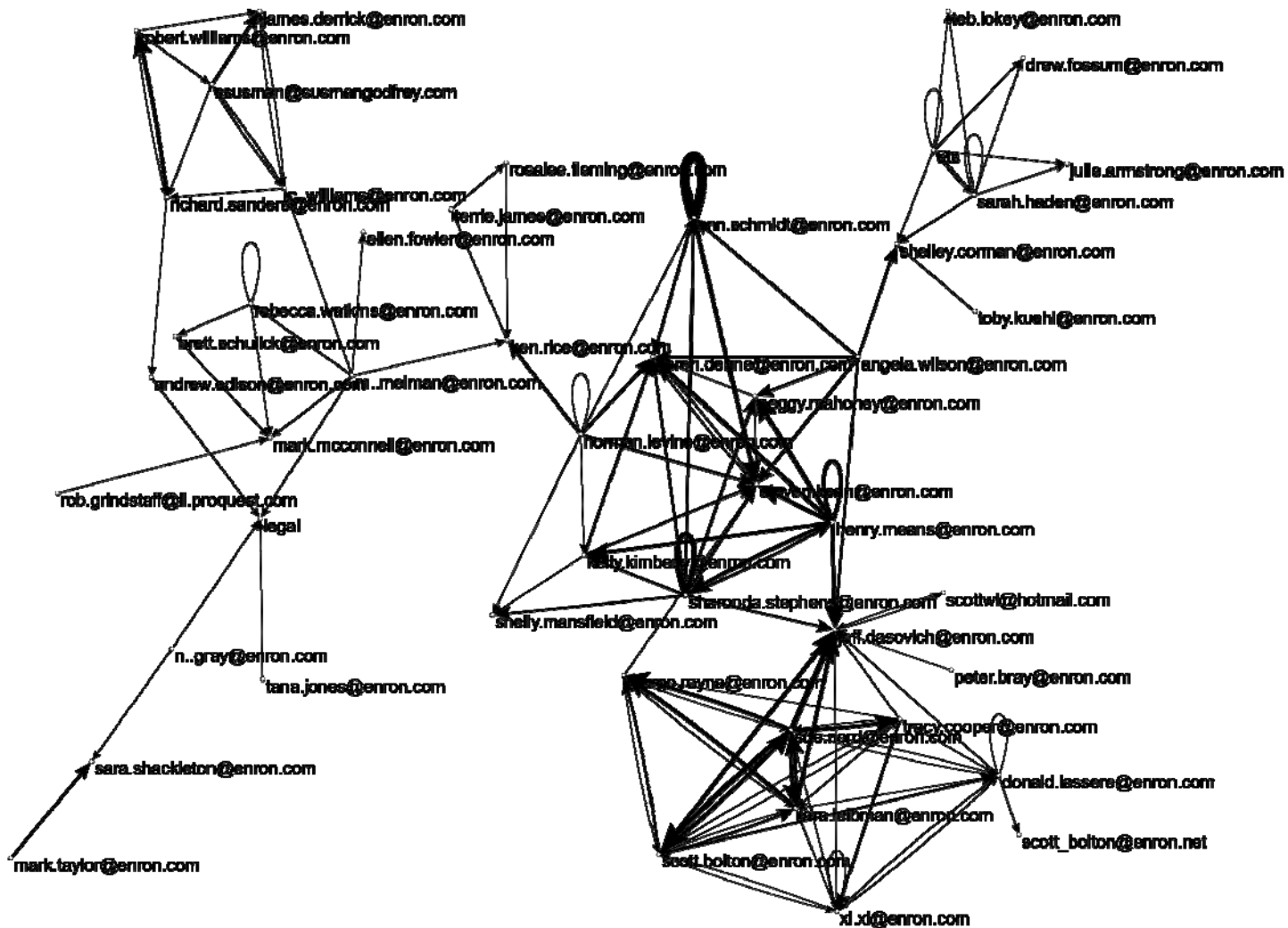


Fig. 7: Reduced network for the 'Blockbuster' query with width of lines indicating number of emails between nodes in the network.

Emerging techniques to relate computational content semantics to users' intentions in e-discovery

1. Social network analysis to supplement / apply models of users' relevance theories in document analysis and retrieval.
 - ❑ Document senders, recipients, owners identify themselves via email.
 - ❑ Model network of knowledge and use resulting structure to draw inferences.
 - ❑ How each item moves through social organization helps map general flow of knowledge.
 - ❑ Infer relevance of document content based on who has them, who sent them, and who likely read them.
2. Machine learning to extend / apply users' theories of relevance
 - ❑ Examples of documents classified according to theory of relevance
 - ❑ used to train an automated classifier to identify and classify other documents.
3. Hypothesis ontology to generalize user modeling regarding relevance theories
 - ❑ vocabulary of objects (agents) and processes,
 - ❑ associated with recurrent areas of interest in e-Discovery such as knowledge transmission in corporate or commercial settings;
 - ❑ define[s] the relationships in which these entities can participate.

Email categorization with machine learning

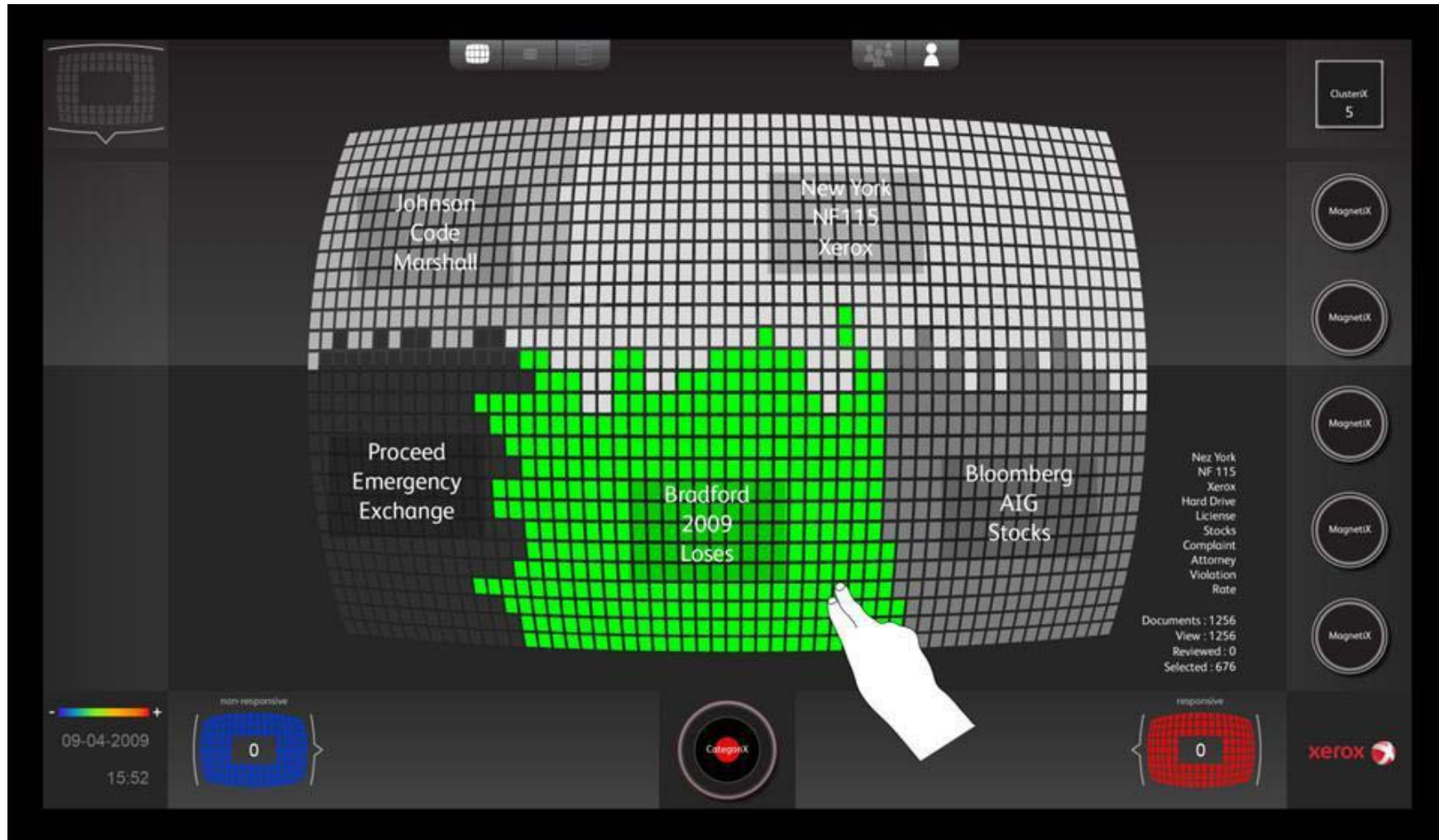


Figure 3: The wall view after ClusteriX applied

(Privault et al. 2010)

Email categorization with machine learning (cont.)

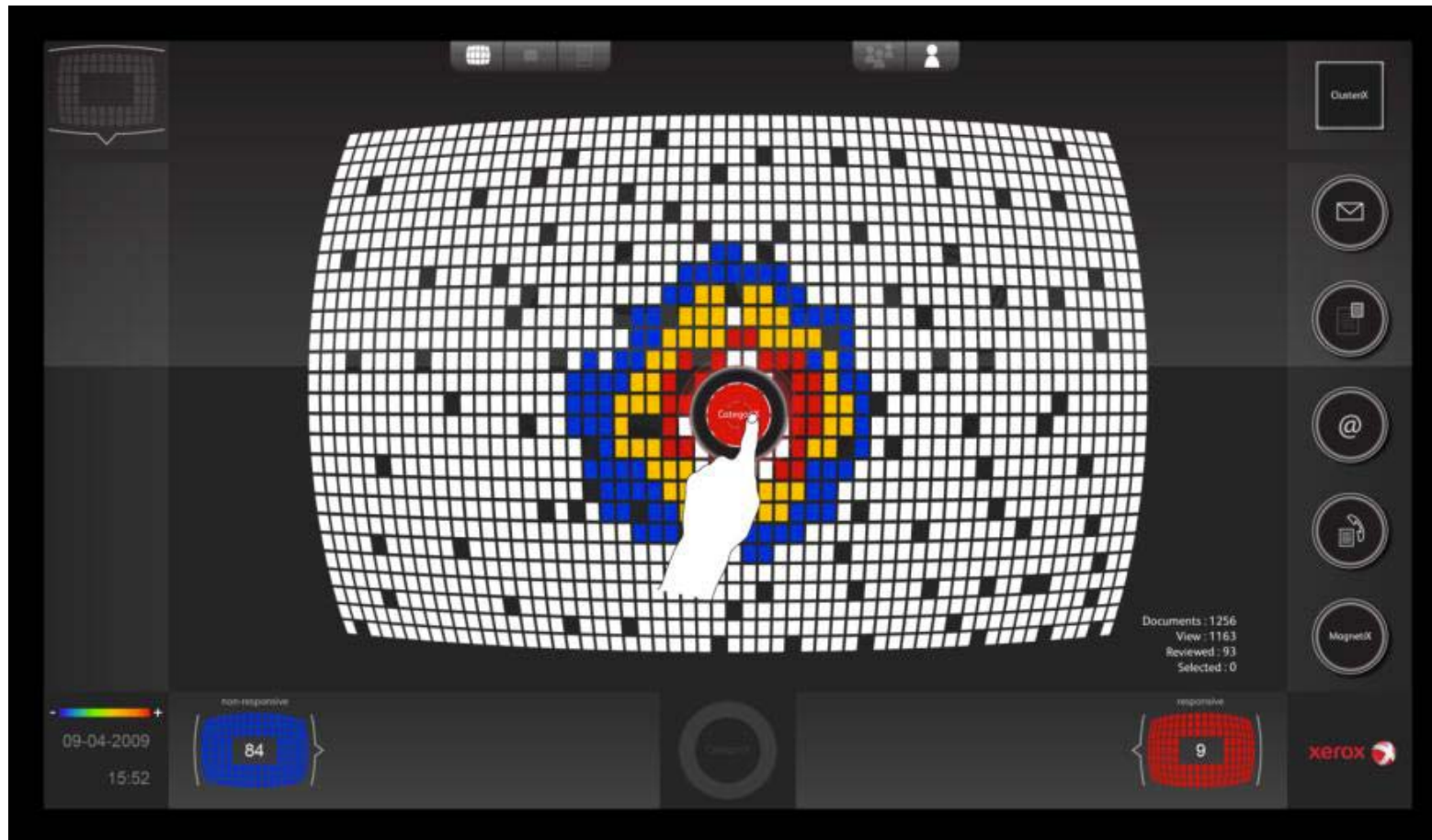


Figure 6: responsive documents grouped around the Categorix virtual magnet.

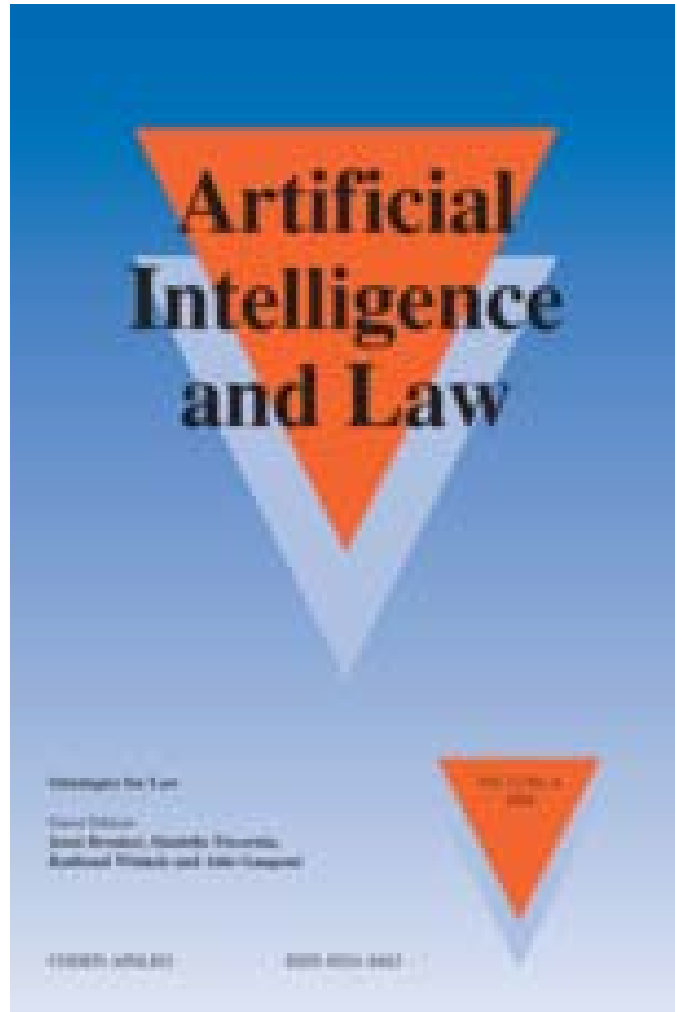
Emerging techniques to relate computational content semantics to users' intentions in e-discovery

1. Social network analysis to supplement / apply models of users' relevance theories in document analysis and retrieval.
 - ❑ Document senders, recipients, owners identify themselves via email.
 - ❑ Model network of knowledge and use resulting structure to draw inferences.
 - ❑ How each item moves through social organization helps map general flow of knowledge.
 - ❑ Infer relevance of document content based on who has them, who sent them, and who likely read them.
2. Machine learning to extend / apply users' theories of relevance
 - ❑ Examples of documents classified according to theory of relevance
 - ❑ used to train an automated classifier to identify and classify other documents.
3. Hypothesis ontology to generalize user modeling regarding relevance theories
 - ❑ vocabulary of objects (agents) and processes,
 - ❑ associated with recurrent areas of interest in e-Discovery such as knowledge transmission in corporate or commercial settings;
 - ❑ define[s] the relationships in which these entities can participate.

Research challenges re relevance hypotheses

1. How to use dynamics of social network over time and communications' time-stamps to refine inferences?
 - ❑ Static: if Alice and Bob correspond about nicotine addiction and similar communications between Charlie and Dana, if Bob and Dana correspond often they may have discussed addiction, as well.
 - ❑ Dynamic: if A, B, C, and D wrote about nicotine addiction within overlapping time frames, then documents between B and D deserve more attention than if A and B wrote to each other in 1981 and Bob and Dana corresponded in 1995.
2. How hypothesis-based retrieval system can explain why documents are relevant in terms of users' research hypotheses.
 - ❑ Complex justifications for document's inclusion or exclusion:
 - E.g., "The Vice President of Marketing communicated frequently with Sara between 1985 and 1989. Sara communicated with a third party, Tom, several times about cigarettes and children. This document between Sara and the Vice President of Marketing mentions children."
 - ❑ Explanation schemas need to be integrated with document clustering.

Upcoming resource on e-discovery



Special Issue on E-Discovery

Artificial Intelligence and Law

Editors: Kevin D. Ashley, Jason R. Baron, Jack G. Conrad

Issue No. 3 (Fall 2010)

1. Emerging AI & Law Approaches to Automating Analysis and Retrieval of Electronically Stored Information in Discovery Proceedings; Kevin Ashley & Will Bridewell
2. E-Discovery Revisited: The Need for Artificial Intelligence Beyond Information Retrieval; Jack G. Conrad
3. Evaluation of Information Retrieval for E-Discovery; Doug Oard, Jason Baron, David Lewis, Bruce Hedin & Stephen Tomlinson
4. Discovery-led refinement in e-discovery investigations: sensemaking, cognitive ergonomics and system design; Simon Attfield & Ann Blandford
5. Automation of Legal Sensemaking in e-Discovery; Chris Hogan, Bob Bauer & Dan Brassil
6. A new tangible user interface for machine learning document review; Caroline Privault, Jacki O'Neill, Jean-Michel Renders & Victor Ciriza
7. Network Based Filtering For Large E-Mail Collections in E-Discovery; Hans Henseler
8. Afterward; David D. Lewis

Conclusions

- Semantic computing and law, legal reasoning and legal practice do intersect!
 - deriving / matching semantics of computational content to naturally expressed user intentions in order to retrieve, manage, manipulate or create content based on its significance to the legal problem-solving or analysis.
- Challenges in:
 - *e-discovery*, business compliance, legal information retrieval, e-commerce and semantic web-based legal services, automated management of privacy and IP rights in data, e-government and legal education.
- Emerging AI & Law approach to e-discovery with semantic computing flavor:
 - Focuses on litigators' hypotheses (or theories) about relevance
 - Iterative user modeling / cognitive task analysis re relevance hypotheses
 - Emerging techniques to relate semantics of computational content to users' intentions re relevance using:
 - Social network analysis, Machine learning, Hypothesis ontology
 - Challenges:
 - Dynamic research challenges re relevance hypotheses
 - Explanation in hypothesis-based retrieval
- *Artificial Intelligence and Law* special issue (2010, no. 3) on e-Discovery

ICAIL-2011 in Pittsburgh

ICAIL 2011: The Thirteenth International Conference on Artificial Intelligence and Law | University of Pittsburgh School of Law

9/24/10 9:41 AM



ICAIL 2011: The Thirteenth International Conference on Artificial Intelligence and Law

Monday, June 6, 2011 - Friday, June 10, 2011

University of Pittsburgh School of Law
3900 Forbes Avenue
Pittsburgh, PA 15260



The 13th International Conference on AI and Law (ICAIL 2011) will be held in Pittsburgh, Pennsylvania, USA, June 6-10, 2011, at the University of Pittsburgh School of Law under the auspices of the International Association for Artificial Intelligence and Law (IAAIL), an organization devoted to promoting research and development in the field of AI and Law with members throughout the world.

The field of AI and Law is concerned with:

- the study of legal reasoning and argumentation, using computational methods
- the formal representation of norms, normative actions, normative systems and norm-governed societies and multi-agent systems
- the investigation of techniques from advanced information technology, using law as the illustrative domain
- applications of advanced information technology to support tasks in the legal domain.

ICAIL provides a forum for the presentation and discussion of the latest research results and practical applications and stimulates interdisciplinary and international collaboration. Previous ICAIL conferences have been held biennially since 1987, with proceedings published by ACM.

| More

ICAIL 2011 Conference Website:

<http://www.law.pitt.edu/icail2011>

June 6-10, 2011

University of Pittsburgh
School of Law

Submissions:

Workshop proposals: 12.6.2010

Papers: 1.10.2011

Other references

- Baron, J.R., & Thompson, P. (2007) The search problem posed by large heterogeneous data sets in litigation: Possible future approaches to research. Proceedings of the Eleventh International Conference on Artificial Intelligence and Law pp. 141–147. Stanford, CA: ACM Press.
- Bauer, R., Jade, T., Hedin, B., and Hogan, C. (2008) Automated Legal Sensemaking: The Centrality of Relevance and Intentionality. In Proceedings 2nd Intl. Workshop on Supporting Search & Sensemaking for Electronically Stored Information in Discovery, London. June. <http://eprints.ucl.ac.uk/9131/1/9131.pdf>
- Hogan, C., Bauer, R., and Brassil, D. (2009) Human-Aided Computer Cognition for e-Discovery. in Proceedings of 12th International Conference on Artificial Intelligence and Law (ICAIL-09). pp. 194-201. Barcelona. June.
-