PEOPLE + DATA + COMPUTATION

Working Productively with Data

Jeffrey Heer
Joe Hellerstein
The last few decades have seen the rise of formal theories of statistics, "legitimizing" variation by confining it by assumption to random sampling, often assumed to involve tightly specified distributions, and restoring the appearance of security by emphasizing narrowly optimized techniques and claiming to make statements with "known" probabilities of error.
While some of the influences of statistical theory on data analysis have been helpful, others have not.
Exposure, the effective laying open of the data to display the unanticipated, is to us a major portion of data analysis. Formal statistics has given almost no guidance to exposure; indeed, it is not clear how the informality and flexibility appropriate to the exploratory character of exposure can be fitted into any of the structures of formal statistics so far proposed.
It is too much to ask for close and effective guidance for data analysis from any highly formalized structure, either now or in the near future.

Data analysis can gain much from formal statistics, but only if the connection is kept adequately loose.
Visualization
Acquisition
Cleaning
Integration
Visualization
Modeling
Presentation
Dissemination
Acquisition
↓
Cleaning
↓
Integration
↓
Visualization
↓
Modeling
↓
Presentation
↓
Dissemination
Effective statistical models for syntactic and semantic disambiguation

Student: Kristina Nikolova Toutanova
Advisor: Christopher D. Manning


Keywords: Syntactic, Semantic, Tree kernels, Parsing

Abstract:

This thesis focuses on building effective statistical models for disambiguation of sophisticated syntactic and semantic natural language (NL) structures. We advance the state of the art in several domains by (i) choosing representations that encode domain knowledge more effectively and (ii) developing machine learning algorithms that deal with the specific properties of NL disambiguation tasks—sparsity of training data and large, structured spaces of hidden labels. For the task of syntactic disambiguation, we propose a novel representation of parse trees that connects the words of the sentence with the hidden syntactic structure in a direct way. Experimental evaluation on parse selection for a Head Driven Phrase Structure Grammar shows the new representation achieves superior performance compared to previous models. For the task of disambiguating the semantic role structure of verbs, we build a more accurate model, which captures the knowledge that the semantic frame of a verb is a joint structure with strong dependencies between arguments. We achieve this using a Conditional Random Field without Markov independence assumptions on the sequence of semantic role labels. To address the sparsity problem in machine learning for NL, we develop a method for incorporating many additional sources of information, using Markov chains in the space of words. The Markov chain framework makes it possible to combine multiple knowledge sources, to learn how much to trust each of them, and to chain inferences together. It achieves large gains in the task of disambiguating prepositional phrase attachments.

Interactive Topic Model Assessment

with Jason Chuang, Dan Ramage, Sonal Gupta & Chris Manning
Oh, the humanities!
Effective statistical models for syntactic disambiguation

Student: Kristina Nikolova Toutanova
Advisor: Christopher D. Manning

Computer Science (2005)

Keywords: Syntactic, Semantic, Tree-Tagged

Abstract:

This thesis focuses on building effective statistical models for syntactic disambiguation of sophisticated natural language (NL) structures. We advance the state-of-the-art by (i) choosing representations that encode meaning effectively and (ii) developing machine learning models that leverage the specific properties of NL disambiguation. We use large, structured datasets and develop methods to learn from distributional properties of syntactic disambiguation, we propose a novel model that connects the words of the sentence into a tree structure in a direct way. Experiments show that our Head Driven Phrase Structure Grammar achieves superior performance compared to existing methods. The task of disambiguating the semantic role of a verb is a joint structure problem. We achieve this using a novel tree-based model that captures Markov independence assumptions on semantic role labels. To address the sparsity problem, we develop a method for incorporating additional information, using Markov chains in a dynamic chain framework makes it possible to disambiguate prepositional phrases together. It achieves large gains over previous disambiguation models.
“Word Borrowing” via Labeled LDA
“Word Borrowing” via Labeled LDA
“Word Borrowing” via Labeled LDA
“Word Borrowing” via Labeled LDA
“Word Borrowing” via Labeled LDA
Real-World Topical Analysis
Real-World Topical Analysis

Figure 7: Applications over time

Applications

Figure 8: Six applied topics over time

looked at trends over time for the following applications: Machine Translation, Spelling Correction, Dialogue Systems, Information Retrieval, Call Routing, Speech Recognition, and Biomedical applications.

Figure 7 shows a clear trend toward an increase in applications over time. The figure also shows an interesting bump near 1990. Why was there such a sharp temporary increase in applications at that time? Figure 8 shows details for each application, making it clear that the bump is caused by a temporary spike in the Speech Recognition topic.

In order to understand why we see this temporary spike, Figure 9 shows the unsmoothed values of the Speech Recognition topic prominence over time. Figure 9 clearly shows a huge spike for the years 1989–1994. These years correspond exactly to the DARPA Speech and Natural Language Workshop, held at different locations from 1989–1994. That workshop contained a significant amount of speech until its last year (1994), and then it was revived in 2001 as the Human Language Technology workshop with a much smaller emphasis on speech processing. It is clear from Figure 9 that there is still some speech research appearing in the Anthology after 1995, certainly more than the period before 1989, but it's equally clear that speech recognition is not an application that the ACL community has been successful at attracting.

Differences and Similarities Among COLING, ACL, and EMNLP

The computational linguistics community has two distinct conferences, COLING and ACL, with different histories, organizing bodies, and philosophies. Traditionally, COLING was larger, with parallel sessions and presumably a wide variety of topics, while ACL had single sessions and a more narrow scope. In recent years, however, ACL has moved to parallel sessions, and the conferences are of similar size. Has the distinction in breadth of topics also been blurred? What are the differences and similarities in topics and trends between these two conferences?

More recently, the EMNLP conference grew out of the Workshop on Very Large Corpora, sponsored by the Special Interest Group on Linguistic Data and corpus-based approaches to NLP (SIGDAT).
Real-World Topical Analysis

Figure 7: Applications over time

Applications
- Machine Translation
- Spelling Correction
- Dialogue Systems
- Information Retrieval
- Call Routing
- Speech Recognition
- Biomedical

Figure 8: Six applied topics over time

- Figure 7 shows a clear trend toward an increase in applications over time. The figure also shows an interesting bump near 1990. Why was there such a sharp temporary increase in applications at that time? Figure 8 shows details for each application, making it clear that the bump is caused by a temporary spike in the Speech Recognition topic.

In order to understand why we see this temporary spike, Figure 9 shows the unsmoothed values of the Speech Recognition topic prominence over time.

Figure 9 clearly shows a huge spike for the years 1989–1994. These years correspond exactly to the DARPA Speech and Natural Language Workshop, held at different locations from 1989–1994. That workshop contained a significant amount of speech until its last year (1994), and then it was revived in 2001 as the Human Language Technology workshop with a much smaller emphasis on speech processing. It is clear from Figure 9 that there is still some speech research appearing in the Anthology after 1995, certainly more than the period before 1989, but it's equally clear that speech recognition is not an application that the ACL community has been successful at attracting.

Differences and Similarities Among COLING, ACL, and EMNLP

The computational linguistics community has two distinct conferences, COLING and ACL, with different histories, organizing bodies, and philosophies. Traditionally, COLING was larger, with parallel sessions and presumably a wide variety of topics, while ACL had single sessions and a more narrow scope. In recent years, however, ACL has moved to parallel sessions, and the conferences are of similar size. Has the distinction in breadth of topics also been blurred? What are the differences and similarities in topics and trends between these two conferences?

More recently, the EMNLP conference grew out of the Workshop on Very Large Corpora, sponsored by the Special Interest Group on Linguistic Data and corpus-based approaches to NLP (SIGDAT) [Talley et al. 2011].

Tags, plus ~215 categorical designations

NIH Sleep Research

<table>
<thead>
<tr>
<th>Topic</th>
<th>Keywords</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Circadian Rhythms</td>
<td>circadian, clock, rhythms, suprachiasmatic clock, sleep/awake cycle</td>
</tr>
<tr>
<td>2 Sleep Disorders</td>
<td>sleep, fatigue, insomnia, older, depression, sleep apnea, obstructive, respiratory</td>
</tr>
<tr>
<td>3 Neurobiology Sleep/Arousal</td>
<td>sleep, hypocretin, orexin, sleep paralysis, REM sleep, narcolepsy</td>
</tr>
<tr>
<td>4 Sleep Disordered Breathing</td>
<td>sleep apnea, obstructive, respiratory</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Tag</th>
<th>FIC</th>
<th>NCCAM</th>
<th>NCI</th>
<th>NCMH</th>
<th>NCRR</th>
<th>NEI</th>
<th>NHLBI</th>
<th>NIGMS</th>
<th>NID</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>31-32 51-58 28-32</td>
<td>13-18 13-15</td>
<td>4-6</td>
<td>50-52</td>
<td>9-1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>72-73 17 27</td>
<td>36-37 27</td>
<td>9-10</td>
<td>2</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>17-31 50-55</td>
<td>35-39 21-25</td>
<td>7-10</td>
<td>3</td>
<td>6</td>
<td>1-5</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>82-89 6-7</td>
<td>1-2 3-7</td>
<td>12-14 21-29</td>
<td>0-5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Total: 192 128 101 73 68 60 56 24
Current Practices
Current Practices

[Topic Words: vegf angiogenesis vascular_endothelial_growth_factor angiogenic end antiangiogenic anti_angiogenic vegf_a tumor_angiogenesis vegfr2 growth signaling bind

Title Words: angiogenesis, vegf, vascular_endothelial_growth_factor, angiogenic, tumor, neovascularization, angiopoietin, signaling, vegfr, vascular, human

Phrases: vascular_endothelial_growth_factor vegf, vegf angiogenesis, vegf receptor,]
## Current Practices

<table>
<thead>
<tr>
<th>Topic</th>
<th>Keywords</th>
</tr>
</thead>
<tbody>
<tr>
<td>Anaphora Resolution</td>
<td>resolution, anaphora, pronoun, discourse, antecedent, pronouns, coreference, reference, definite, algorithm</td>
</tr>
<tr>
<td>Automata</td>
<td>string, state, set, finite, context, rule, algorithm, strings, language, symbol</td>
</tr>
<tr>
<td>Biomedical</td>
<td>medical, protein, gene, biomedical, wkh, abstracts, medline, patient, clinical, biological</td>
</tr>
<tr>
<td>Call Routing</td>
<td>call, caller, routing, calls, destination, vietnamese, routed, router, destinations, gorin</td>
</tr>
<tr>
<td>Categorial Grammar</td>
<td>proof, formula, graph, logic, calculus, axioms, axiom, theorem, proofs</td>
</tr>
<tr>
<td>Centering*</td>
<td>centering, cb, discourse, cf, utterance, center, utterances, theory, coherence, entities, local</td>
</tr>
<tr>
<td>Classical MT</td>
<td>japanese, method, case, sentence, analysis, english, dictionary, figure</td>
</tr>
<tr>
<td>Classification/Tagging</td>
<td>features, data, corpus, set, feature, table, word, tag, al, test</td>
</tr>
<tr>
<td>Comp. Phonology</td>
<td>vowel, phonological, syllable, phoneme, stress, phonetic, phonology, pronunciation, vowels, phonemes</td>
</tr>
<tr>
<td>Comp. Semantics*</td>
<td>semantic, logical, semantics, john, sentence, interpretation, scope, logical</td>
</tr>
</tbody>
</table>

[Hall et al. 2008]
Current Practices

Topic Words: vegf angiogenesis vascular_endothelial_growth_factor angiogenic endotherm
antiangiogenic anti_angiogenic vegf a tumor_angiogenesis vegfr2 growth signaling isoforms biology.

Title Words: neovascularization angiogenesis neovascularization angiogenic

Anaphora Resolution resolution anaphora pronoun discourse antecedent pronouns coreference reference definite algorithm

Automata string state set finite context rule algorithm strings language symbol.

Biomedical medical protein gene biomedical wkh abstracts medline patient clinical biological

Call Routing call caller routing calls destination vietnamese routed router destinations gorin

Categorial Grammar proof formula graph logic calculus axioms axiom theorem proofs lambek

Centering centering cb discourse cf utterance center utterances theory coherence entities local

Classical MT japanese method case sentence analysis english dictionary figure japan word

Classification1Tagging features data corpus set feature table word tag al test

Compfi Phonology vowel phonological syllable phoneme stress phonetic phonology pronunciation vowels phonemes

Compfi Semanticsw semantic logical semantics john sentence interpretation scope logic form set

Dialogue Systems user dialogue system speech information task spoken human utterance language

Discourse Relations discourse text structure relations rhetorical relation units coherence texts rst

Discourse Segmentfi segment segmentation segments chain chains boundaries boundary seg cohesion lexical

Events1Temporal event temporal time events tense state aspect reference relations relation

French Function de le des les en une est du par pour

Generation generation text system language information knowledge natural figure domain input

Genre Detection genre stylistic style genres fiction humor register biber authorship registers

Infofi Extraction system text information muc extraction template names patterns pattern domain

Information Retrieval document documents query retrieval question information answer term text web

Lexical Semantics semantic relations domain noun corpus relation nouns lexical ontology patterns

MUC Terrorism slot incident tgt target id hum phys type fills perp

Metaphor metaphor literal metonymy metaphors metaphorical essay metonymic essays qualia analogy

Morphology word morphological lexicon form dictionary analysis morphology lexical stem arabic

Named Entitiesw entity named entities ne names ner recognition ace nes mentions mention

Paraphrase1RTE paraphrases paraphrase entailment paraphrasing textual para rte pascal entailed dagan

Parsing parsing grammar parser parse rule sentence input left grammars np

PlanzBased Dialogue plan discourse speaker action model goal act utterance user information

Probabilistic Models model word probability set data number algorithm language corpus method

Prosody prosodic speech pitch boundary prosody phrase boundaries accent repairs intonation

Semantic Rolesw semantic verb frame argument verbs role roles predicate arguments

Yale School Semantics knowledge system semantic language concept representation information network concepts base

Sentiment subjective opinion sentiment negative polarity positive wiebe reviews sentence opinions

Speech Recognition speech recognition word system language data speaker error test spoken

Spell Correction errors error correction spelling ocr correct corrections checker basque corrected detection

Statistical MT english word alignment language source target sentence machine bilingual mt

Statistical Parsing dependency parsing treebank parser tree parse head model al np

Summarization sentence text evaluation document topic summary summarization human summaries score

Syntactic Structure verb noun syntactic sentence phrase np subject structure case clause

TAG Grammarsw tree node trees nodes derivation tag root figure adjoining grammar

Unification feature structure grammar lexical constraints unification constraint type structures rule

WSDw word senses wordnet disambiguation lexical semantic context similarity dictionary

Word Segmentation chinese word character segmentation corpus dictionary korean language table system

WordNetw synset wordnet synsets hypernym ili wordnets hypernyms eurowordnet hyponym ewn wn

[Hall et al. 2008]
How might we better support human-in-the-loop verification of topic models?
Termite | Topic Model Visualization

Representative Documents

A Comparison of the Readability of Graphs Using Node-Link and Matrix Visualization
Mohammad Ghoniem Jean-Daniel Fekete Philippe Castagliola

Using Multilevel Call Matrices in Large Software Projects
Frank van Ham

Improving the Readability of Clustered Social Networks using NodeTrix
Nathalie Henry Anastasia Bezerianos Jean-Daniel Fekete

MatrixExplorer: a Dual-Representation System to Explore Social Networks
Nathalie Henry Jean-Daniel Fekete

NodeTrix: a Hybrid Visualization of Social Networks
Nathalie Henry Jean-Daniel Fekete Michael J. McGuffin

The need to visualize large social networks is growing as hardware and many new data sets become available. Unfortunately, the visualizations resolve the basic dilemma of being readable both for the global structure of local communities. To address this problem, we present NodeTrix, which combines the advantages of two traditional representations: node-link of a network, while arbitrary portions of the network can be shown. NodeTrix visualization by dragging selections to and from node-link to create a NodeTrix representation to explore the dataset and create meaning. Finally, we present a case study applying NodeTrix to the analysis to illustrate the capabilities of NodeTrix as both an exploration tool and a tool for researchers.

Visualizing Causal Semantics using Animations
Nivedita R. Kadaba Parrang P. Trani Jason Leboe

Balancing Systematic and Flexible Exploration of Social Networks
Adam Perer Ben Shneiderman

Social network analysis (SNA) has emerged as a powerful method to explore networks. However, interactive exploration of networks is currently patterns and comprehend the structure of networks with many nodes and overwhelming visual output which leads to structural and visual patterns that are difficult to interpret. NodeTrix allows for flexible exploration through visualizations of measures to gain an overview of networks using link structure, find cohesive subgroups, and focus on networks by viewing different link types separately, or find patterns in an overview. For each operation, a stable node layout is maintained to perform comparisons. SocialAction offers analysts a strategy beyond opportunities for exploring social networks.
1. Graph Visualization
   Graph, network, node-link diagram, layout, adjacency matrix, reordering
   Asymmetric Relations in Longitudinal Social Networks
   Multi-Level Graph Layout on the GPU
   Balancing Systematic and Flexible Exploration of Social Networks
   Parallel Edge Splatting for Scalable Dynamic Graph Visualization

2. Text Visualization
   Text, topics, sentiment analysis
   The Shape of Shakespeare: Visualizing Text Using Implicit Surf
   ThemeRiver: visualizing theme changes over time
   From Metaphor to Method: Cartographic Perspectives on Inform
   Participatory Visualization with Words
   FacetAtlas: Multifaceted Visualization for Rich Text Corpora
   Mapping Text with Phrase Nets

3. Multidimensional Visualization
   Parallel coordinates, small multiples, splom, scatterplot matrix, multidimensional projections, embeddings, MDS, PCA
   Rolling the Dice: Multidimensional Visual Exploration using So
   Scattering Points in Parallel Coordinates
   Multidimensional Detective
   Improved Similarity Trees and their Application to Visual Display
   Steerable, Progressive Multidimensional Scaling

4. Tree Visualization
   Treemap, node-link diagram, hierarchies
   SpaceTree: supporting exploration in large node-link trees, design
   Browsing Zoomable Treemaps: Structure-Aware Multi-Scale Na
   Interactive Visualization of Genealogical Graphs

5. Software Visualization
   Algorithm animation, traces, logs
   code_swarm: A Design Study in Organic Software Visualization
   The Visual Code Navigator: An Interactive Toolset for Source Code
   Using MultiLevel Call Matrices in Large Software Projects

6. Topic: Significant and coherent area of research
   Exemplary terms: techniques, methods, systems, people...
   Separate the terms by commas or semicolons
   Exemplary documents: 3 or more papers
   Drag and drop from InfoVis proceedings

2011 InfoVis Conference
Providence, Rhode Island

Theory and Foundations
- Quality Metrics in High-Dimensional Data Visualization: An Overview and Systematization
  Enrico BERTINI, Andraad TUATU, Daniel KEIM
- Benefiting InfoVis with Visual Difficulties
  Jessica HULLMAN, Eyat ALDER, Priti SHAH
- Product Plots
  Hadley WICKHAM, Heike HOFMANN
- Visualization Rhetoric: Framing Effects in Narrative Visualization
  Jesse HULLMAN, Nick DIKAPPOULOS
- Adaptive Privacy-Preserving Visualization Using Parallel Coordinates
  Aris DASGUPTA, Robert KOSARA

Techniques
- Context-Preserving Visual Links
  Markus STEINBERGER, Manuela WALDNER
  Marc STREIT, Alexander LEX
  Dieter SCHMALSTIEG
- Design Study of LineSets, a Novel Set Visualization Technique
  Basak ALPER, Nathalie RICHE, Gonzalo RAMOS
  Mary CZERWINSKI
- Developing and Evaluating Quilts for the Depiction of Large Layered Graphs
  Juhee BAE, Benjamin WATSON
- Arc Length-based Aspect Ratio Selection
  Justin TALBOT, John CERTH, Pat HANRAHAN
- Asymmetric Relations in Longitudinal Social Networks
  Ulrike BRANDIS, Bobo NICK

Systems and Frameworks
- VisBricks: Multiform Visualization of Large, Inhomogeneous Data
  Alexander LEX, Hans-Joerg SCHULZ
  Marc STREIT, Christian PARIL
  Dieter SCHMALSTIEG
- VisBricks: A Bridge to Frameworks
  Alexander LEX, Hans-Joerg SCHULZ
  Marc STREIT, Christian PARIL
  Dieter SCHMALSTIEG
Prominent issue is the lack of a recognizable correspondence between the highest quality LDA topic model and the set of expert-generated InfoVis topics. Another interesting phenomenon is the absence of a number of topics, which exhibit coherent textual descriptions in our survey but are missing from the LDA model. Another interesting observation is that the majority of LDA topics align well with the exerts’ clustering engines. In general, topics on topological and hierarchical visualization frameworks, such as trees and tree maps, are represented clearly in LDA. However, topics related to multi-dimensional visualization, such as Parallel Coordinates, are less clearly identifiable. Additionally, LDA generated multiple redundant topics (e.g., four latent topics corresponding to experts’ concepts of animation).
Vis are the exerts'...
LDA generated multiple redundant topics (e.g., four latent topics corresponding to experts' concepts of InfoVis research areas). We also found that 22 of the generated topics are “junk” topics that do not usefully help organize InfoVis research areas.
Missing concepts
A prominent issue is the lack of a recognizable correspondence between the highest quality LDA topic model and the experts' concepts of InfoVis research areas. We found that 22 of the generated topics are "junk" topics that do not usefully help organize InfoVis research areas.


Psychology 4, 1983.

Feltovich, P.J., J. P. M. J., and Swanson, D. Designing model-driven visualizations for text analysis. In NIPS Workshop on Challenges of Data Visualization.


The prominent issue is the lack of a recognizable visual representation. Figure 5 shows a correspondence between the highest quality LDA topic model and InfoVis research areas. We found that 22 of the generated topics are "junk" topics that do not help organize the research areas. The clustering engines perform well in comparing LSA, PMI, and GLSA similarity measures on common corpora. In a recent study, experts and novices were compared using a symbolic skills representation of physics problems. A model-driven visualization for text analysis was presented at the VAST challenge. The role and predictability of medical knowledge in diagnostic expertise were also explored. Perception and information scent were modeled. A survey of web development techniques for assessing textual topic models was conducted. A visualization tool, LSA View, was developed. The predictability and information in semantic processing were analyzed. A spreading-activation theory was applied to a spreading activation model. A visualization tool for UNIX command line was also presented.
The prominent issue is the lack of a recognizable correspondence between the highest quality LDA topic model and the set of expert-generated InfoVis topics. Two notable omissions are the expert-generated topics with lower quality, which exhibit coherent textual descriptions in our survey but are missing from the LDA model. Another key issue is the number of redundant topics, such as four latent topics corresponding to experts' concepts of *trees* and *treemaps*.

<table>
<thead>
<tr>
<th>Expert Topics</th>
<th>Latent Topics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Graphs</td>
<td>23, 25, 26</td>
</tr>
<tr>
<td>Networks</td>
<td>2, 3, 4, 6</td>
</tr>
<tr>
<td>Trees</td>
<td></td>
</tr>
<tr>
<td>Treemaps</td>
<td></td>
</tr>
<tr>
<td>Multi Dimensional</td>
<td></td>
</tr>
<tr>
<td>Parallel Coords.</td>
<td></td>
</tr>
<tr>
<td>Text</td>
<td></td>
</tr>
<tr>
<td>GeoVis</td>
<td></td>
</tr>
<tr>
<td>BioVis</td>
<td></td>
</tr>
<tr>
<td>Time Series</td>
<td></td>
</tr>
<tr>
<td>Uncertainty</td>
<td></td>
</tr>
<tr>
<td>Narrative</td>
<td></td>
</tr>
<tr>
<td>Software</td>
<td></td>
</tr>
<tr>
<td>Devices</td>
<td></td>
</tr>
<tr>
<td>Evaluation</td>
<td></td>
</tr>
<tr>
<td>Perception</td>
<td></td>
</tr>
<tr>
<td>Cognition</td>
<td></td>
</tr>
<tr>
<td>Theory</td>
<td></td>
</tr>
<tr>
<td>Collaboration</td>
<td></td>
</tr>
<tr>
<td>Social</td>
<td></td>
</tr>
<tr>
<td>For the Masses</td>
<td></td>
</tr>
<tr>
<td>Toolkits</td>
<td></td>
</tr>
<tr>
<td>Systems</td>
<td></td>
</tr>
<tr>
<td>Interact. Theories</td>
<td></td>
</tr>
<tr>
<td>Interact. Tech.</td>
<td></td>
</tr>
<tr>
<td>Animation</td>
<td></td>
</tr>
<tr>
<td>Overview and Detail</td>
<td></td>
</tr>
<tr>
<td>Multiple Views</td>
<td></td>
</tr>
</tbody>
</table>
A prominent issue is the lack of a recognizable view of the experts' views. LDA generated multiple redundant topics (e.g., four latent topics corresponding to experts' concepts of GeoVis, Time Series, Text, and Parallel Coords.). We also found that 22 of the generated topics are "junk" topics that do not usefully help organize InfoVis research areas.

Repeated concepts
The prominent issue is the lack of a recognizable correspondence between the highest quality LDA topic model and the expert-generated InfoVis topics. We found that 22 of the generated topics are "junk" topics that do not usefully help organize InfoVis research areas. Two notable omissions are...
Resolved/Fused Concepts vs. Number of Latent Topics

(alpha=5/N, beta=0.25)
Resolved/Fused Concepts vs. Number of Latent Topics

(alpha=5/N, beta=0.25)

% of Concepts

Resolved
Repeated
Fused & Repeated

Number of Latent Topics
Expert-based evaluation of over 10,000 models, across inference algorithms and parameter settings.

Chuang et al., ICML 2013
STEPPING BACK
STEPPING BACK: A POST-GRADUATE EXERCISE

Your chosen field’s biggest blind spot?
STEPPING BACK: A POST-GRADUATE EXERCISE

Your chosen field’s biggest blind spot?
STEPPING BACK: 
A POST-GRADUATE EXERCISE

Your chosen field’s biggest blind spot?

“A little disdain is not amiss; 
a little scorn is alluring.”
— William Congreave, 1670–1729
BATCH PROCESSING AND UI
BATCH PROCESSING AND UI

This query needs to use large tables in the database. It may run slowly. Consider submitting this query to QueryBack in order to avoid waiting.

Use QueryBack...  Run Anyway  Cancel

http://control.cs.berkeley.edu

CONTROL

• On-Line processing of large datasets
  – constant, useful feedback for long-running (data-intensive) operations
  – progressive refinement of answers
  – online user control

• A blend of
  – data processing, statistics, UI

http://control.cs.berkeley.edu

CONTROL

• On-Line processing of large datasets
  – constant, useful feedback for long-running (data-intensive) operations
  – progressive refinement of answers
  – online user control

• A blend of
  – data processing, statistics, UI
GOALS FOR ONLINE PROCESSING

- Maximize 1st derivative of the “mirth index”
- Mirth subject to dynamic redefinition
- Need FEEDBACK and CONTROL
## ONLINE AGGREGATION

### [SIGMOD ’97, ‘99]

#### Table: GPA per College

<table>
<thead>
<tr>
<th>College</th>
<th>Count</th>
<th>GPA</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>190</td>
<td>2.9353933</td>
</tr>
<tr>
<td>B</td>
<td>208</td>
<td>3.0174129</td>
</tr>
<tr>
<td>D</td>
<td>186</td>
<td>3.5308642</td>
</tr>
<tr>
<td>E</td>
<td>167</td>
<td>3.0384614</td>
</tr>
<tr>
<td>H</td>
<td>177</td>
<td>3.1497006</td>
</tr>
<tr>
<td>K</td>
<td>191</td>
<td>3.0500000</td>
</tr>
<tr>
<td>L</td>
<td>185</td>
<td>3.0730338</td>
</tr>
<tr>
<td>M</td>
<td>209</td>
<td>3.1683416</td>
</tr>
<tr>
<td>N</td>
<td>175</td>
<td>3.2116663</td>
</tr>
<tr>
<td>P</td>
<td>177</td>
<td>3.1225805</td>
</tr>
<tr>
<td>Q</td>
<td>169</td>
<td>3.3609271</td>
</tr>
<tr>
<td>R</td>
<td>184</td>
<td>3.5167599</td>
</tr>
<tr>
<td>V</td>
<td>183</td>
<td>3.2289157</td>
</tr>
<tr>
<td>Z</td>
<td>22</td>
<td>3.0227273</td>
</tr>
</tbody>
</table>

#### GPA Distribution Graph

- **GPA per College**
- **Confidence**: 95%
CLOUDS [IEEE COMPUTER 8/99]
POTTER’S WHEEL [VLDB 01]
HADOOP WINTER
HADOOP WINTER AND A NEW SPRING
Scalable Approximate Query Processing With The DBO Engine

Christopher Jermaine, Subramanian Arumugam, Abhijit Pol, Alin Dobra
CISE Department, University of Florida
Gainesville, FL, USA
{cjermain, sa2, apol, adobra}@cise.ufl.edu

MapReduce Online

Tyson Condie, Neil Conway, Peter Alvaro, Joseph M. Hellerstein
UC Berkeley
Khaled Elmeleegy, Russell Sears
Yahoo! Research

Incremental, Approximate Database Queries and Uncertainty for Exploratory Visualization

Danyel Fisher
Microsoft Research

Online Aggregation for Large MapReduce Jobs

Niketan Pansare¹, Vinayak Borkar², Chris Jermaine¹, Tyson Condie³
¹Rice University, ²UC Irvine, ³Yahoo! Research
np6@rice.edu, vborkar@ics.uci.edu, cmj4@rice.edu, tcondie@yahoo-inc.com

Stat! – An Interactive Analytics Environment for Big Data
Mike Barnett¹, Badrish Chandramouli¹, Robert DeLine¹, Steven Drucker¹, Danyel Fisher¹, Jonathan Goldstein¹, Patrick Morrison², John Platt¹
¹Microsoft Research
Redmond, Washington, USA
{mbarnett, badris, rdeline, sdrucker, danyelf, jongold, jplatt}@microsoft.com
²North Carolina State University
Raleigh, North Carolina, USA
pjmorris@ncsu.edu
ENGAGING PRACTITIONERS

To gain insight, we interviewed 35 analysts:
ENGAGING PRACTITIONERS

To gain insight, we interviewed 35 analysts:

25 Companies
Healthcare
Retail, Marketing
Social networking
Media
Finance, Insurance

Various titles
Data analyst
Data scientist
Software engineer
Consultant
Chief technical officer
ENGAGING PRACTITIONERS

To gain insight, we interviewed 35 analysts:

25 Companies
Healthcare
Retail, Marketing
Social networking
Media
Finance, Insurance

Various titles
Data analyst
Data scientist
Software engineer
Consultant
Chief technical officer

Enterprise Data Analysis and Visualization: An Interview Study
Sean Kandel, Andreas Paepcke, Joseph Hellerstein, Jeffrey Heer
IEEE Visual Analytics Science & Technology (VAST), 2012
The last set of analysts performed almost all operations in a spread.

... and then produced visualizations using the statistical package during ex-

... resident on one machine (as opposed to distributed). Scripters

... and application of algorithms was more easily done when dealing with

... observed. Advanced modeling was potentially enabled by the

... not know how to create scripts that run at scale.

... comfortable writing scripts in a scripting language, but typically do

... data warehouse by IT staff and stored in an expected format. Some of

... such as filtering and aggregating data, but typically could not perform

... such as R or Matlab. They were able to perform simple manipulations

4.1.2 Scripters

... data.

... perform flexible data manipulation within visualization tools they only

... of the database and into an Excel format that I can start

... All data is in a relational database. When I get it, it's out

... challenges and tools. The matrix displays interviewees (grouped by archetype and sector) and their corresponding chal-

... Hackers faced the most diverse set of challenges, corresponding to the diversity of their workflows and toolset. Application users

... scripters typically relied on the IT team to perform certain tasks and therefore did not perceive them as challenges.
“I spend more than half of my time integrating, cleansing and transforming data without doing any actual analysis. Most of the time I’m lucky if I get to do any ‘analysis’ at all.”
“I spend more than half of my time integrating, cleansing and transforming data without doing any actual analysis. Most of the time I’m lucky if I get to do any ‘analysis’ at all.”

Lost productivity
“I spend more than half of my time integrating, cleansing and transforming data without doing any actual analysis. Most of the time I’m lucky if I get to do any ‘analysis’ at all.”

Lost productivity
“I spend more than half of my time integrating, cleansing and transforming data without doing any actual analysis. Most of the time I’m lucky if I get to do any ‘analysis’ at all.”

Lost productivity

“Most of the time once you transform the data ... the insights can be scarily obvious.”
“I spend more than half of my time integrating, cleansing and transforming data without doing any actual analysis. Most of the time I’m lucky if I get to do any ‘analysis’ at all.”

Lost productivity

“Most of the time once you transform the data ... the insights can be scarily obvious.”

Lost accessibility
“I spend more than half of my time integrating, cleansing and transforming data without doing any actual analysis. Most of the time I’m lucky if I get to do any ‘analysis’ at all.”

Lost productivity

“Most of the time once you transform the data ... the insights can be scarily obvious.”

Lost accessibility
“I spend more than half of my time integrating, cleansing and transforming data without doing any actual analysis. Most of the time I’m lucky if I get to do any ‘analysis’ at all.”

Lost productivity

“Most of the time once you transform the data ... the insights can be scarily obvious.”

Lost accessibility
“I spend more than half of my time integrating, cleansing and transforming data without doing any actual analysis. Most of the time I’m lucky if I get to do any ‘analysis’ at all.”

Lost productivity

“Most of the time once you transform the data … the insights can be scarily obvious.”

Lost accessibility
“It’s easy to just think you know what you are doing and not look at data at every intermediary step. An analysis has 30 different steps. It’s tempting to just do this then that and then this. You have no idea in which ways you are wrong and what data is wrong.”

Interactivity and Visualization
Analytic Productivity
Remove drudgery, restore time
Analytic Productivity
  Remove drudgery, restore time

Data Accessibility
  Enable self-service data manipulation
CONTROL

- On-Line processing of large datasets
  - constant, useful feedback for long-running (data-intensive) operations
  - progressive refinement of answers
  - online user control

- A blend of
  - data processing, statistics, UI
CONTROL

- On-Line processing of large datasets
  - constant, useful feedback for long-running (data-intensive) operations
  - progressive refinement of answers
  - online user control
- A blend of
  - data processing, statistics, UI
TRADITIONAL LANDSCAPE

- Decision-Makers
- Quants
- IT Staff

Raw Data | Actionable Data | Decisions
NEW LANDSCAPE

LOB Data
Partner Data
Data Vendors
Public Data

BI Tools
Analytic DBMS, MapReduce, Hive, Pig, etc.

Enterprise
Data

IT Staff

Quants

Decision-Makers

Raw Data
Actionable Data
Decisions

Spreadsheets
BI Tools
Stat Packages
HDFS
Analytic DBMS, MapReduce, Hive, Pig, etc.
NEW LANDSCAPE

- Raw Data
  - LOB Data
  - Partner Data
  - Data Vendors
  - Public Data

- IT Staff
  - Enterprise Data

- Quants
  - HDFS
  - Analytic DBMS, MapReduce, Hive, Pig, etc.

- Decision-Makers
  - BI Tools
  - Stat Packages
  - Spreadsheets
PEOPLE + DATA + COMPUTATION