accenture

Technology Labs

Research Directions in Enterprise Knowledge Management

Rayid Ghani

Enterprise Search Today

accenture

accenture technology labs | Search | Advanced Search

Accenture Technology Labs
SIT - **Technology** > **Accenture Technology Labs Accenture Technology Labs** Asset Monitoring & Optimization Practice **Accenture Technology Labs** Asset Monitoring & Optimization Practice
https://kx.accenture.com/Organizations/Pages/AccentureTechnologyLabs.aspx
Content Current Date - 6/14/2005
Last Index Date - 2/13/2007
[View duplicates]

Accenture Technology Labs Patent Profiles
Accenture Technology Labs Patent Profiles **Accenture Technology Labs** Patent Profiles **Accenture Technology Labs**
https://kx.accenture.com/TechMeth/Pages/LabsPatent-AllProfiles.aspx
Content Current Date - 4/7/2005
Last Index Date - 12/29/2006
[View duplicates]

Accenture Technology Labs: Contacts - Meet the **Labs**
Accenture Technology Labs: **Accenture Technology Labs** falls under Systems Integration & **Technology** under the direction of the director of the **Accenture Technology Labs**, responsible for setting strategic direction and managing
https://kx.accenture.com/Organizations/Pages/AccentureTechnologyLabs-Contacts.aspx
Content Current Date - 5/24/2005
Last Index Date - 7/6/2007

Accenture Technology Labs - Sell & Deliver
paper describing **Accenture Technology Labs**' "Command and Control (C2) Interactive Wall" and how it Short Video Clips of **Technology** Prototypes and Solutions from **Accenture Technology Labs** showcasing the **Accenture Technology Labs** latest prototypes and solutions.
https://kx.accenture.com/Organizations/Pages/AccentureTechnologyLabs-SellDeliver.aspx
Content Current Date - 6/13/2005
Last Index Date - 12/29/2006

Accenture Technology Labs India
SIT - **Technology** > **Accenture Technology Labs Accenture Technology Labs** India in Collaboration - **Accenture Technology Labs** creates a Virtual Corridor between **Accenture** offices i
https://kx.accenture.com/Organizations/Pages/AccentureTechnologyLabs-Bangalore.aspx
Content Current Date - 3/20/2006
Last Index Date - 6/28/2007

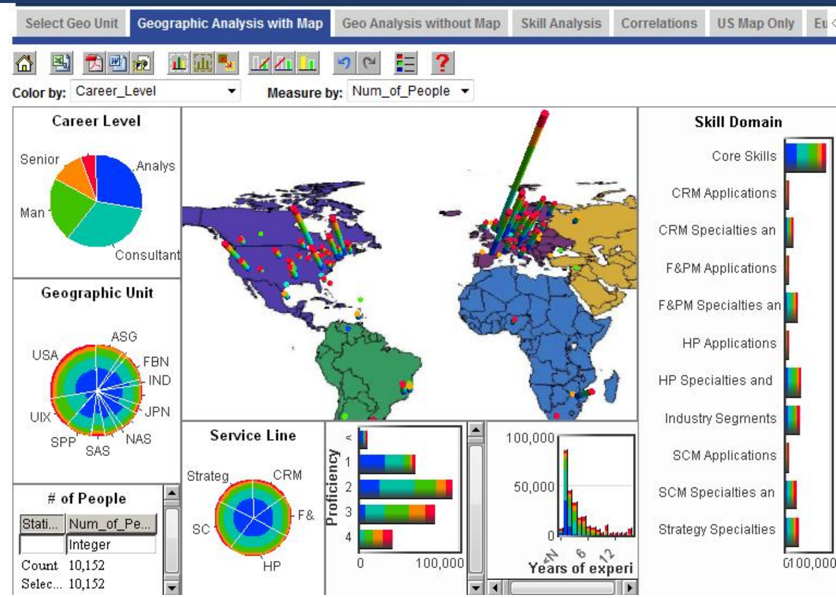
Accenture Technology Labs - About the Organization
SIT - **Technology** > **Accenture Technology Labs Accenture Technology Labs**, the **technology** research and development (R&D) organization within Acce **Accenture Technology Labs** falls under Systems Integration & **Technology** under the direction of the
https://kx.accenture.com/Organizations/Pages/AccentureTechnologyLabs-About.aspx

Enterprise IR is different from Web IR

- Scale
- Hyperlinks
- Static ranking
- Structured data
- Version Control
- Expectations (THE document vs a document)
- Redundancy
- Security/access control
- Users (roles/needs/context)
- Tasks (Business processes)

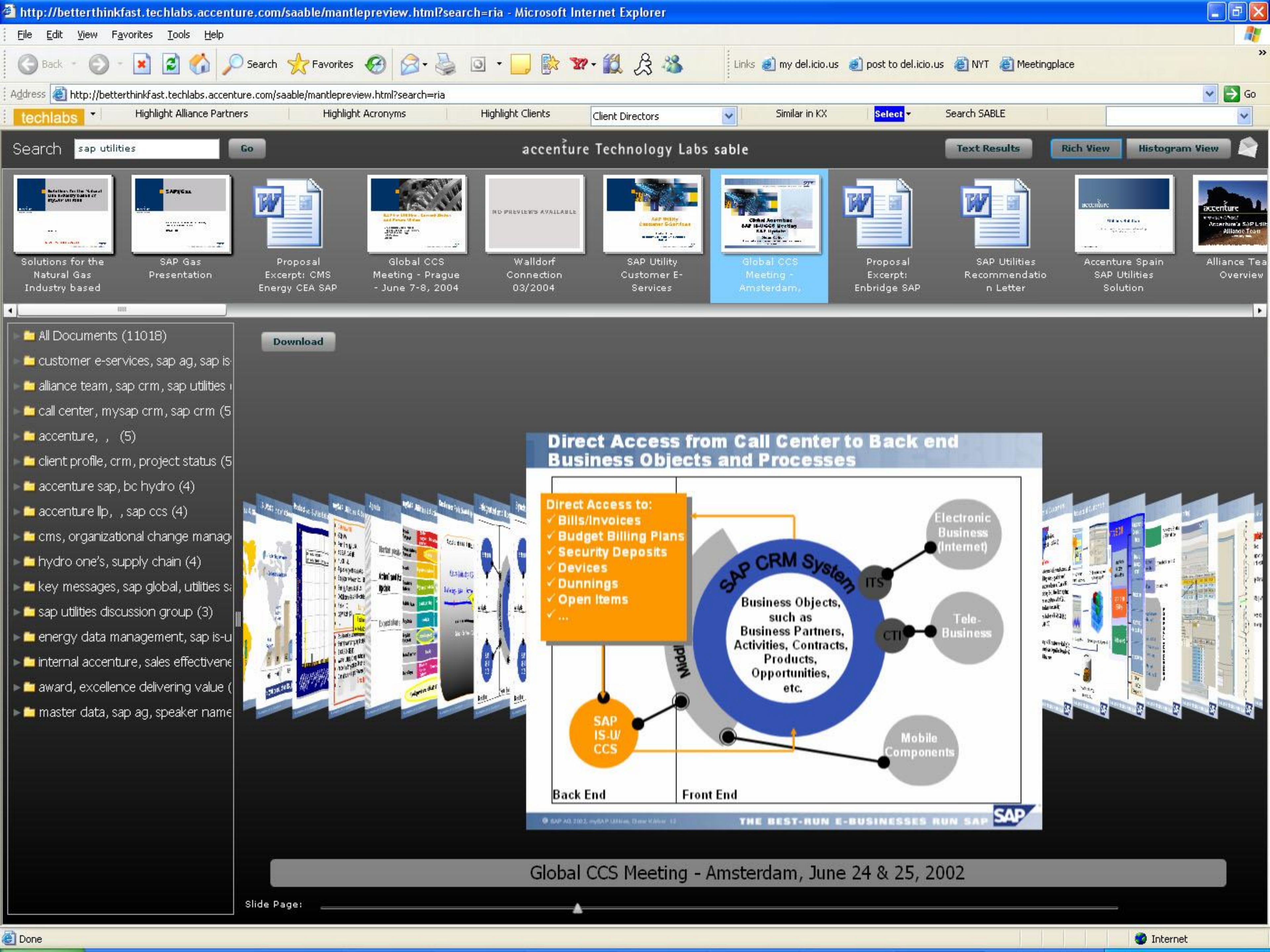
Enterprise Tasks involve specialized business processes and roles

- Proposal Writing
- Marketing
- Selling/Estimation
- Risk management
- Requirements Analysis
- Software Testing
- Training/Learning
- Outsourcing
- Vendor evaluation
- Procurement
- Business Intelligence
- Project Staffing
- Recruiting



Collaborative Document Development Process

- Identify requirements
- Identify collaborators/team
- Develop high level themes
- Create outline with sections
- Assign sections to individuals
- Support individuals finding content to complete sections
- Support checkpoints and alerts
- Consolidate drafts of sections and provide support to ensure consistency, compliance, theme integration, tracking changes
- Consolidate to produce final document
- Review for consistency, compliance, integration
- finalize



http://betterthinkfast.techlabs.accenture.com/saable/mantlepreview.html?search=ria - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites
Links my del.icio.us post to del.icio.us NYT Meetingplace
Address http://betterthinkfast.techlabs.accenture.com/saable/mantlepreview.html?search=ria
Go
techlabs
Highlight Alliance Partners
Highlight Acronyms
Highlight Clients
Client Directors
Similar in KX
Select
Search SABLE
Search
sap utilities
Go
accenture Technology Labs sable
Text Results
Rich View
Histogram View
Solutions for the Natural Gas Industry based
SAP Gas Presentation
Proposal Excerpt: CMS Energy CEA SAP
Global CCS Meeting - Prague - June 7-8, 2004
Walldorf Connection 03/2004
SAP Utility Customer E-Services
Global CCS Meeting - Amsterdam,
Proposal Excerpt: Enbridge SAP
SAP Utilities Recommendation Letter
Accenture Spain SAP Utilities Solution
Alliance Team Overview
All Documents (11018)
customer e-services, sap ag, sap is
alliance team, sap crm, sap utilities
call center, mysap crm, sap crm (5
accenture, , (5)
client profile, crm, project status (5
accenture sap, bc hydro (4)
accenture llp, , sap ccs (4)
cms, organizational change manag
hydro one's, supply chain (4)
key messages, sap global, utilities s
sap utilities discussion group (3)
energy data management, sap is-u
internal accenture, sales effectivene
award, excellence delivering value (
master data, sap ag, speaker name
Download
Direct Access from Call Center to Back end Business Objects and Processes
Direct Access to:
✓ Bills/Invoices
✓ Budget Billing Plans
✓ Security Deposits
✓ Devices
✓ Dunnings
✓ Open Items
✓ ...
SAP CRM System
Business Objects, such as Business Partners, Activities, Contracts, Products, Opportunities, etc.
Middl
ITS
CTI
Electronic Business (Internet)
Tele-Business
Mobile Components
SAP IS-U/ CCS
Back End
Front End
THE BEST-RUN E-BUSINESSES RUN SAP
SAP
Global CCS Meeting - Amsterdam, June 24 & 25, 2002
Slide Page:
Done
Internet

Areas of Improvement

Access to specific content instead of large monolithic documents

Task and Context-sensitive access to knowledge

Enterprise KM

Document Analysis and Review Tools

Process-based Support

Areas of Improvement

Access to specific content instead of large monolithic documents

Task and Context-sensitive access to knowledge

Enterprise KM

Document Analysis and Review Tools

Process-based Support

From monolithic documents to reusable pieces of information

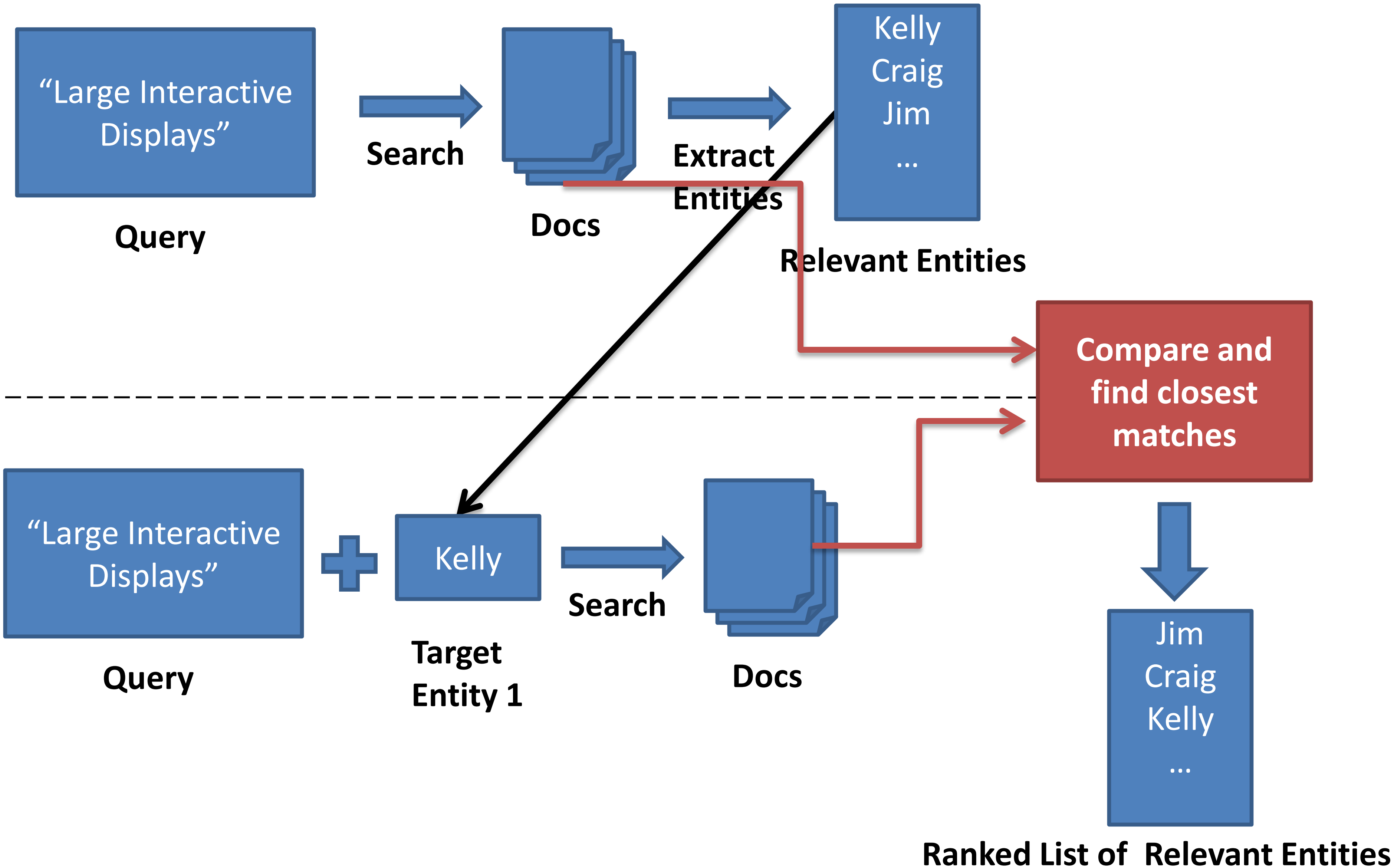
- Entities (experts for example - demo)
- Document Chunking (demo)
- Images (demo)

Motivating business task

- Task: Developing a consulting project proposal

- Information Needs
  - *People* that have worked on similar proposals
  - *Clients* for which Accenture has done similar work
  - *Vendors* that Accenture has used for similar work
  - *Alliances* Accenture has with companies that we can partner with
  - *People* available to work on the project
  - …

Entity Retrieval & Ranking



Online experiments

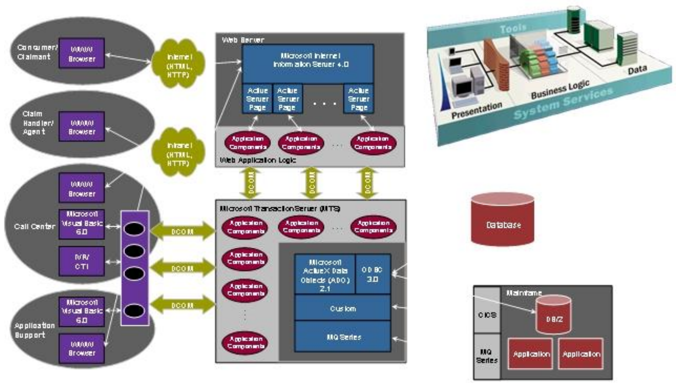
- Expert Search as an extension to enterprise search

- IM bot to find and interact with experts

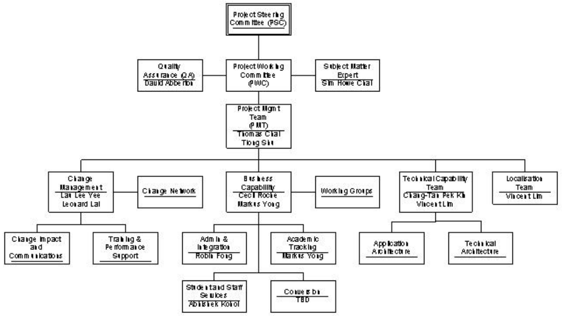
Document Chunking

- Approach:
  - Determine boundary of sections by segmenting documents
  - Use clustering to find similar sections
  - Classify sections into predetermined section classes
- Output:
  - List of labeled individual sections from the document

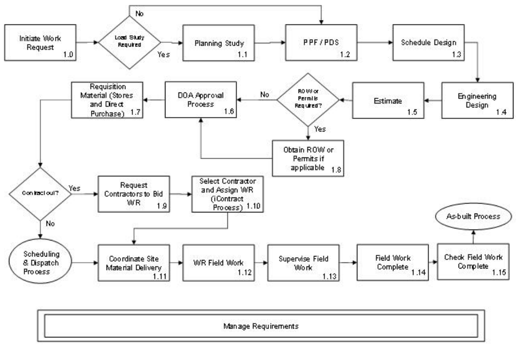
Enterprise Image Classification and Retrieval



Architecture Diagram



Org Chart

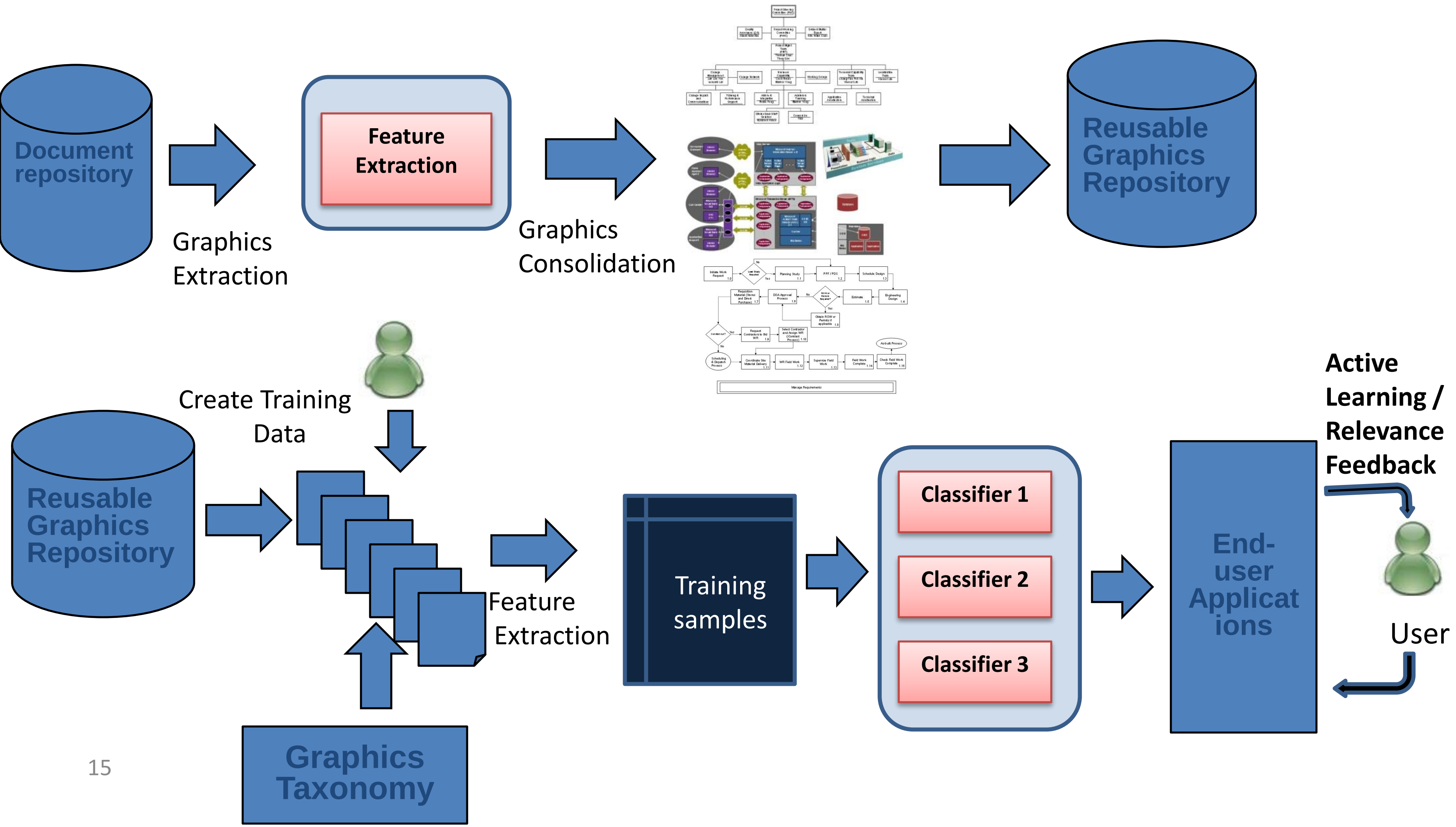


Process Flow

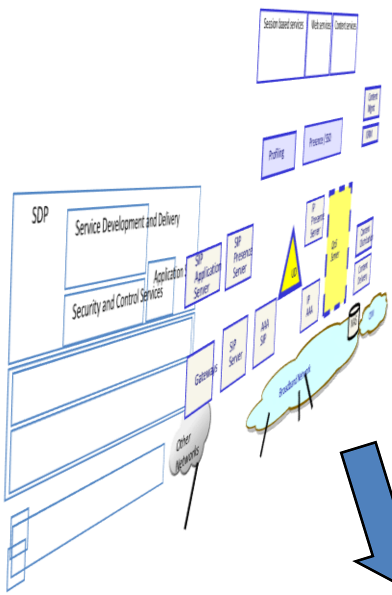
[Demo](#)

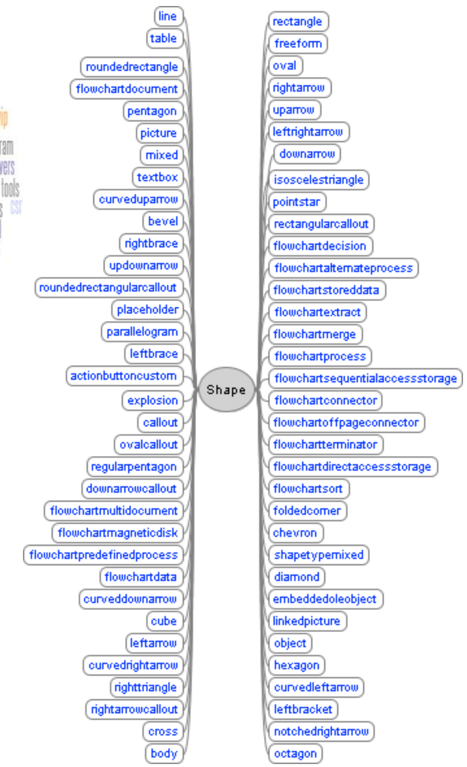
- Approach
  - Image extraction from office Documents
  - Supervised learning (6-10 categories) with structural, visual, and text features of the image
  - Index image using words and image category

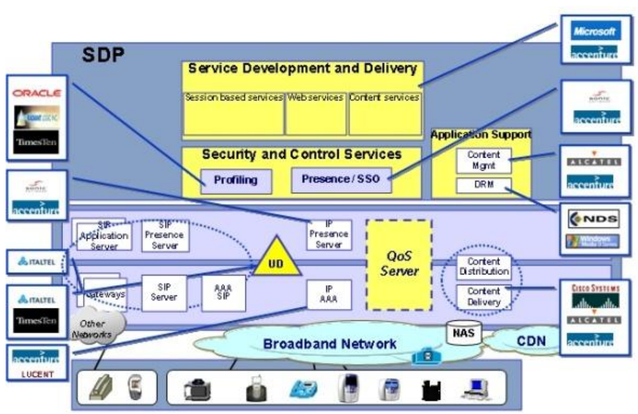
Overview of Approach



Feature Extraction







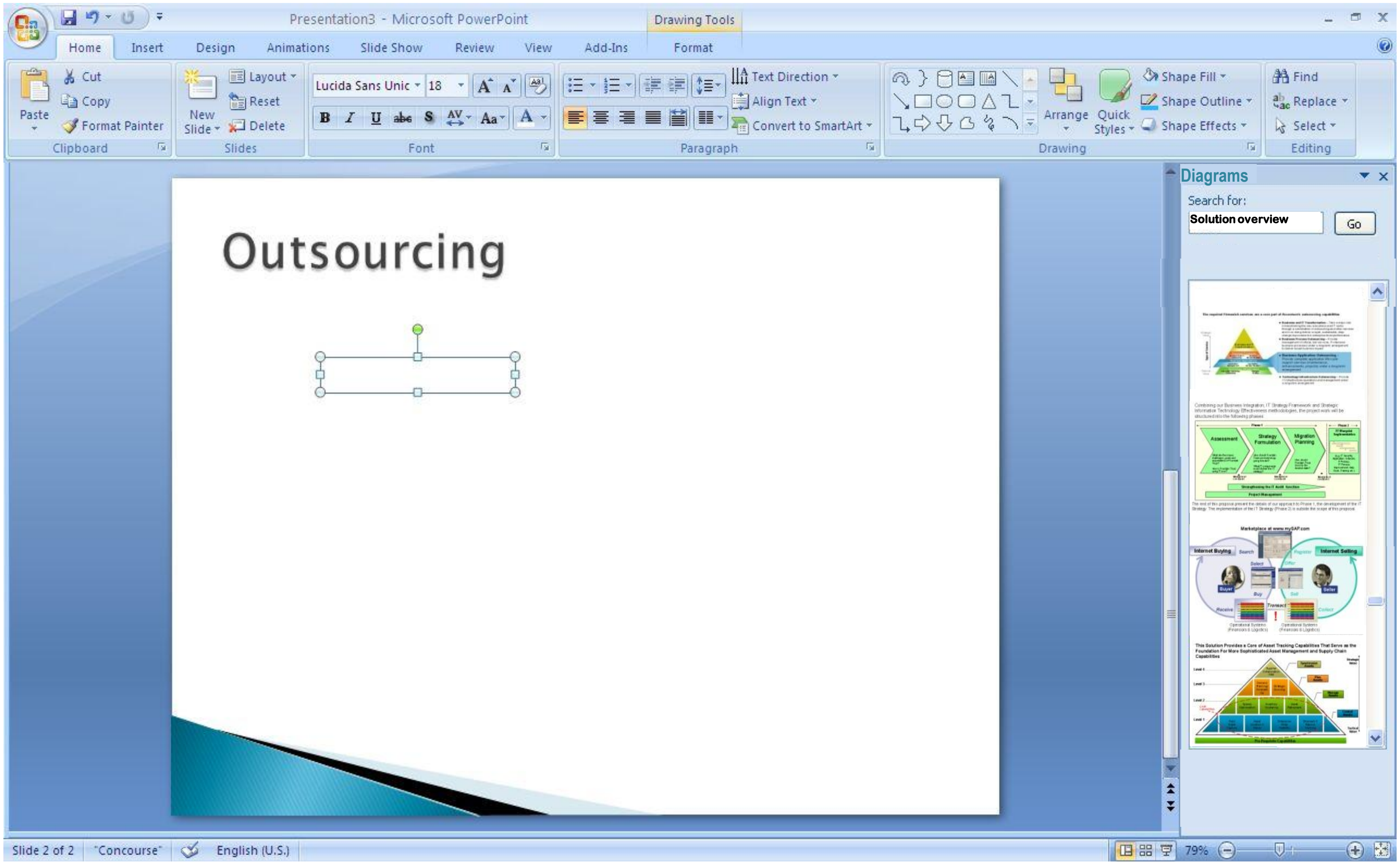
Local Color Structure

Main Layout of Colors

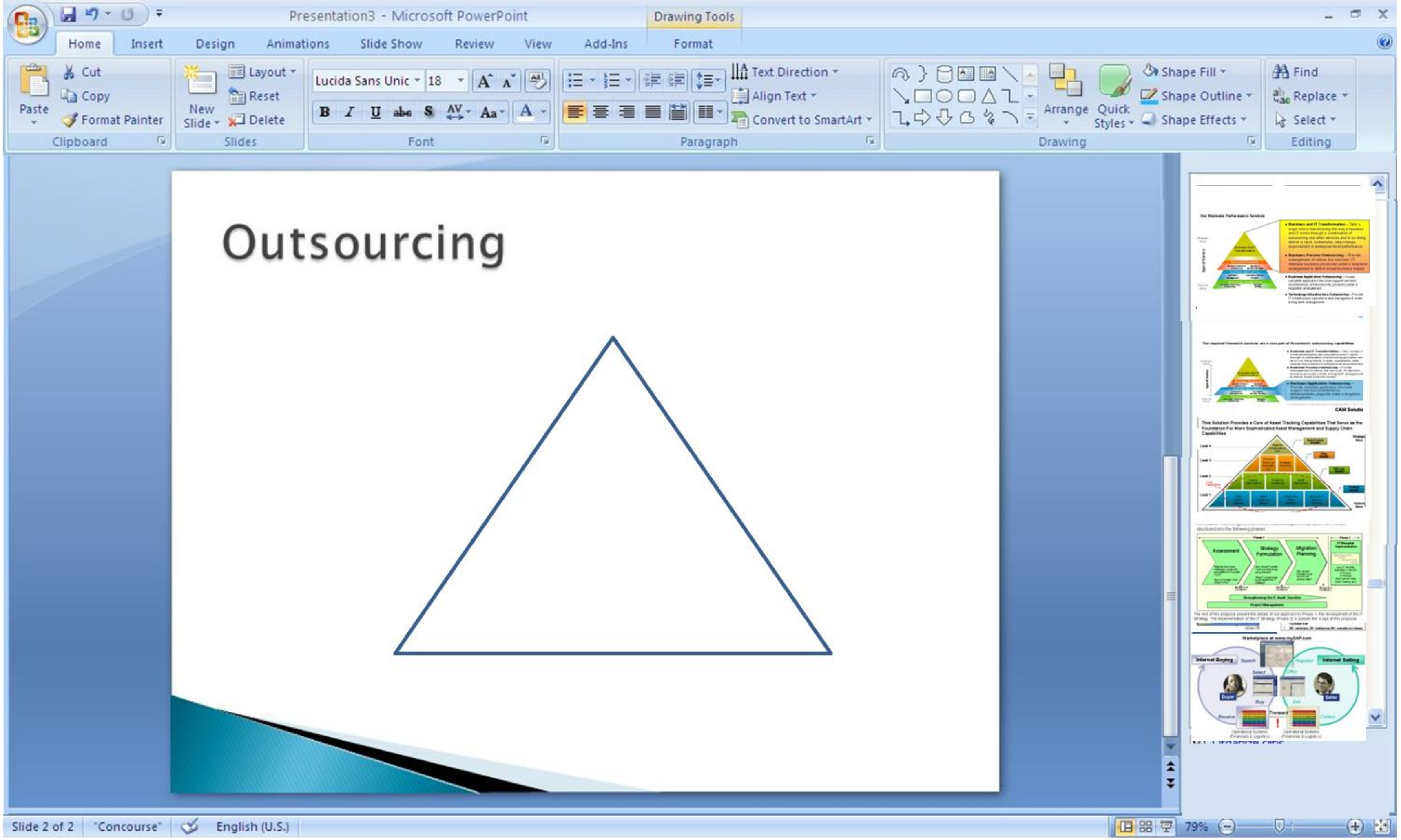
Histogram of Edges

Texture -Gabor Filters

Task specific application



3. Task specific application



Areas of Improvement

Access to specific content instead of large monolithic documents

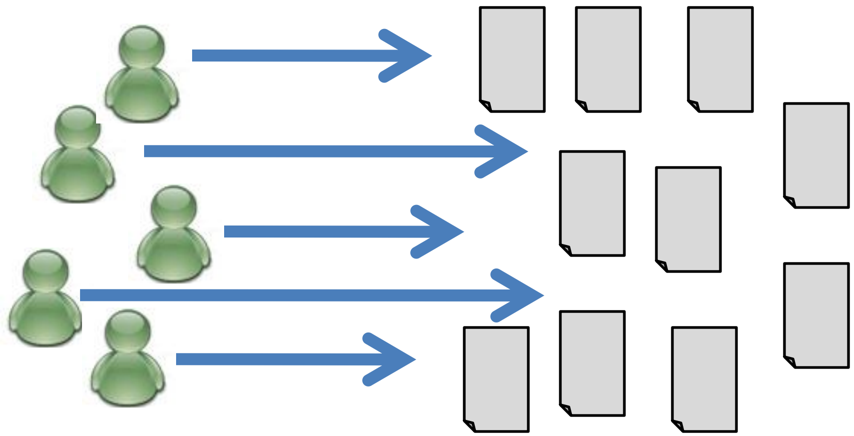
Task and Context-sensitive access to knowledge

Enterprise KM

Document Analysis and Review Tools

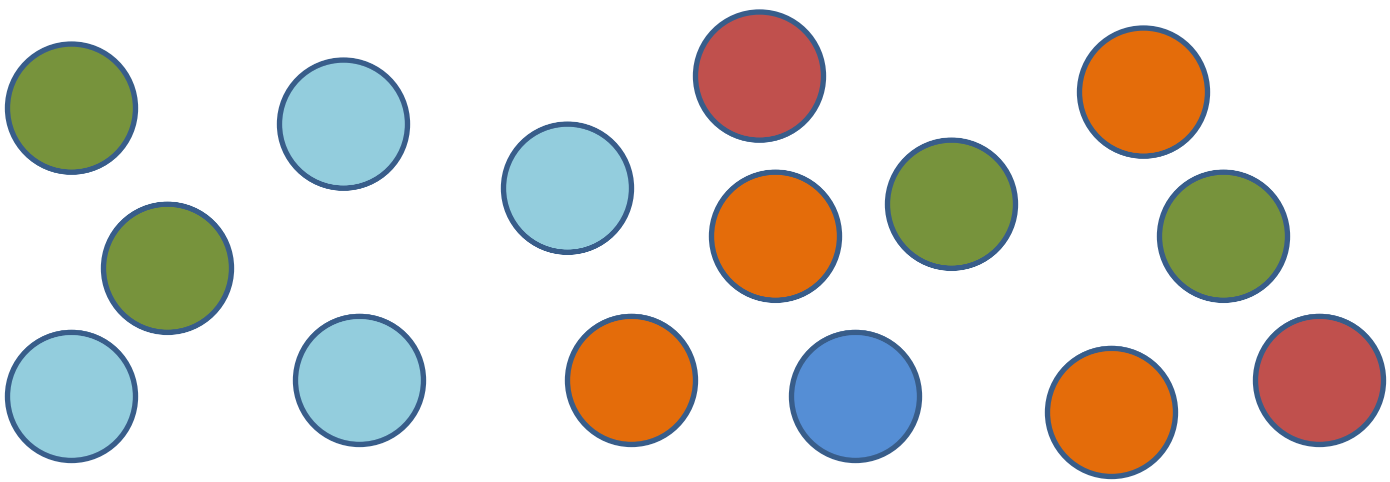
Process-based Support

Context

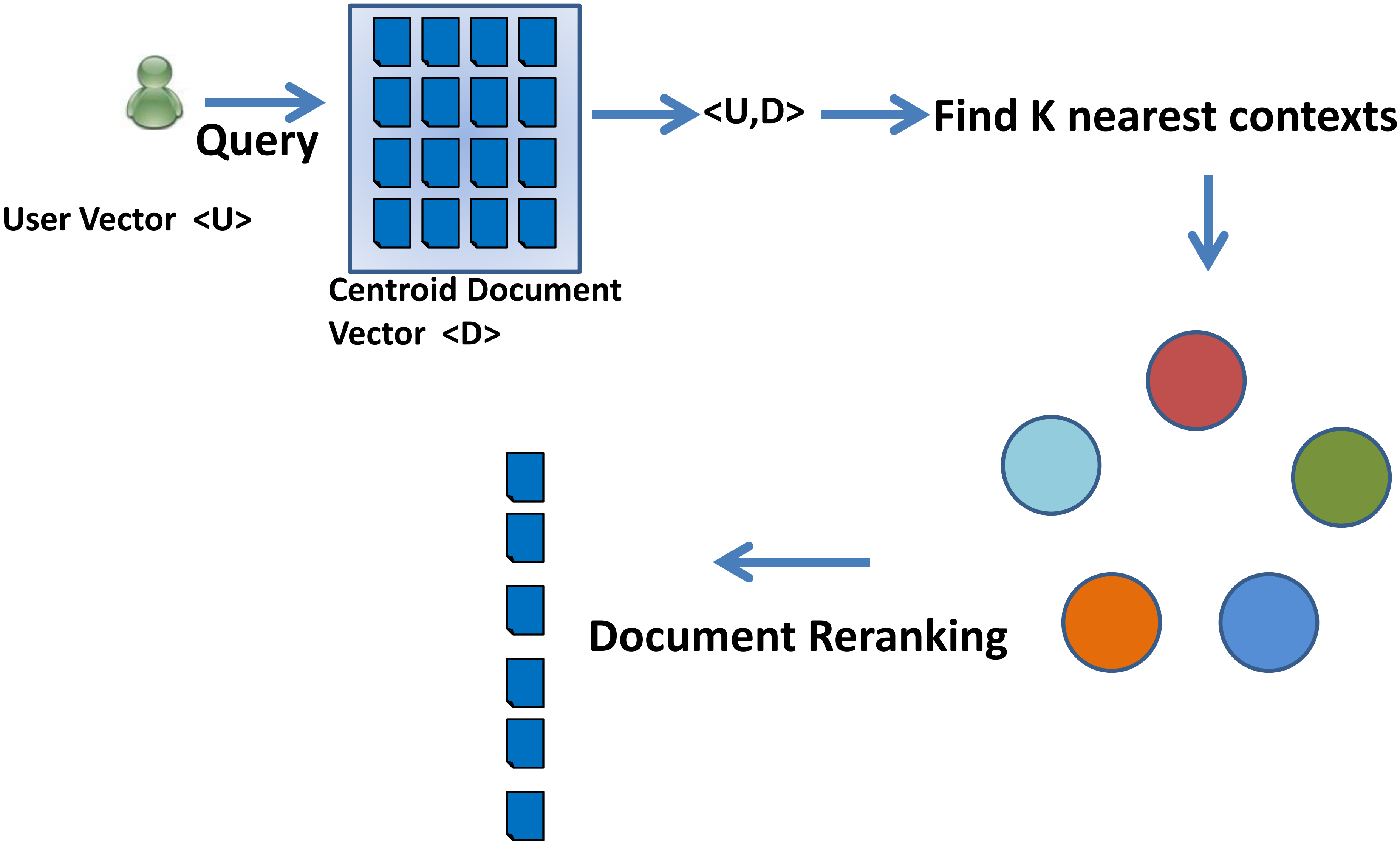


Context = **User Role** + **Document Attributes**

Super Contexts = Cluster All contexts over time



Context-sensitive ranking

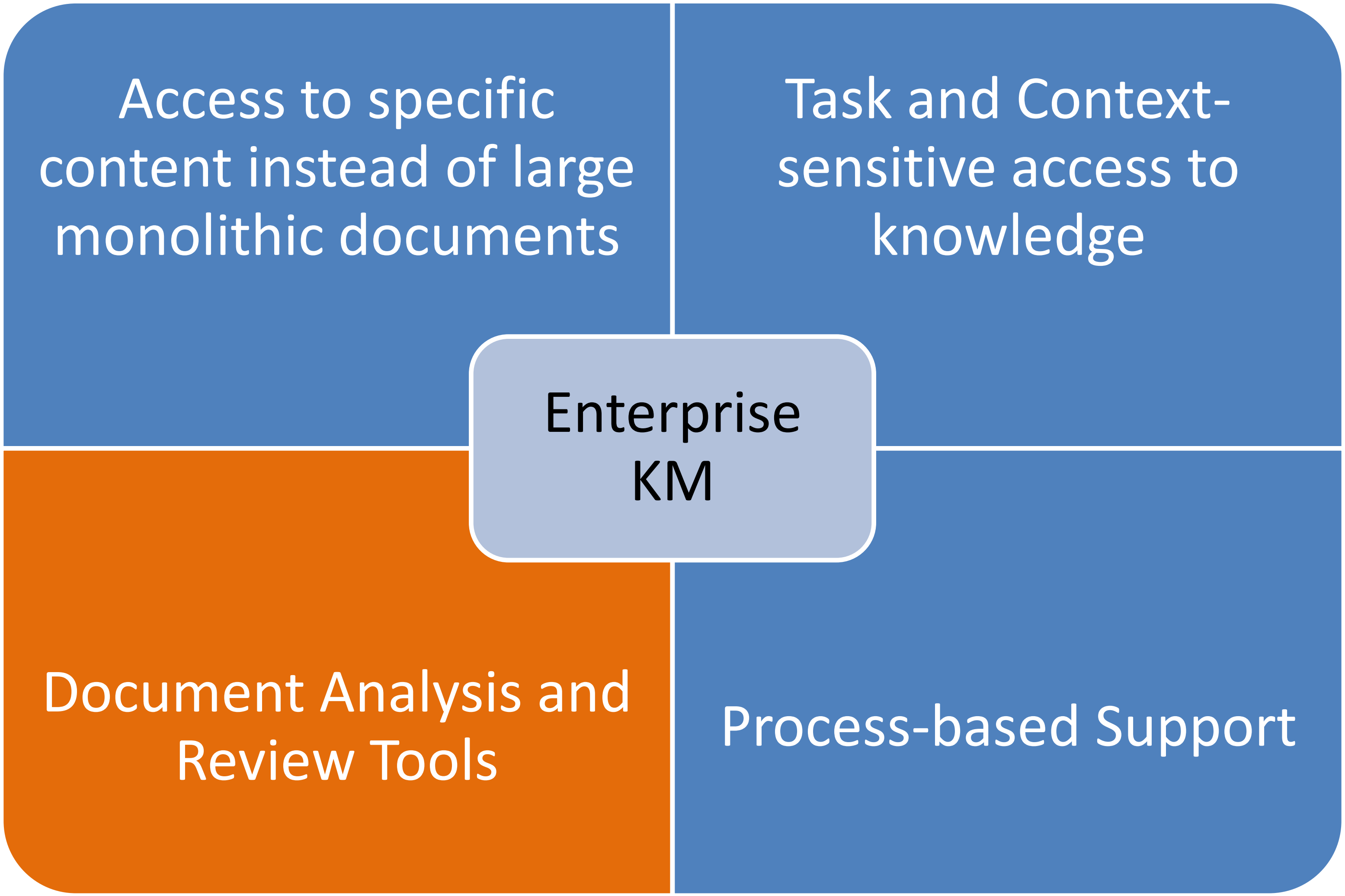


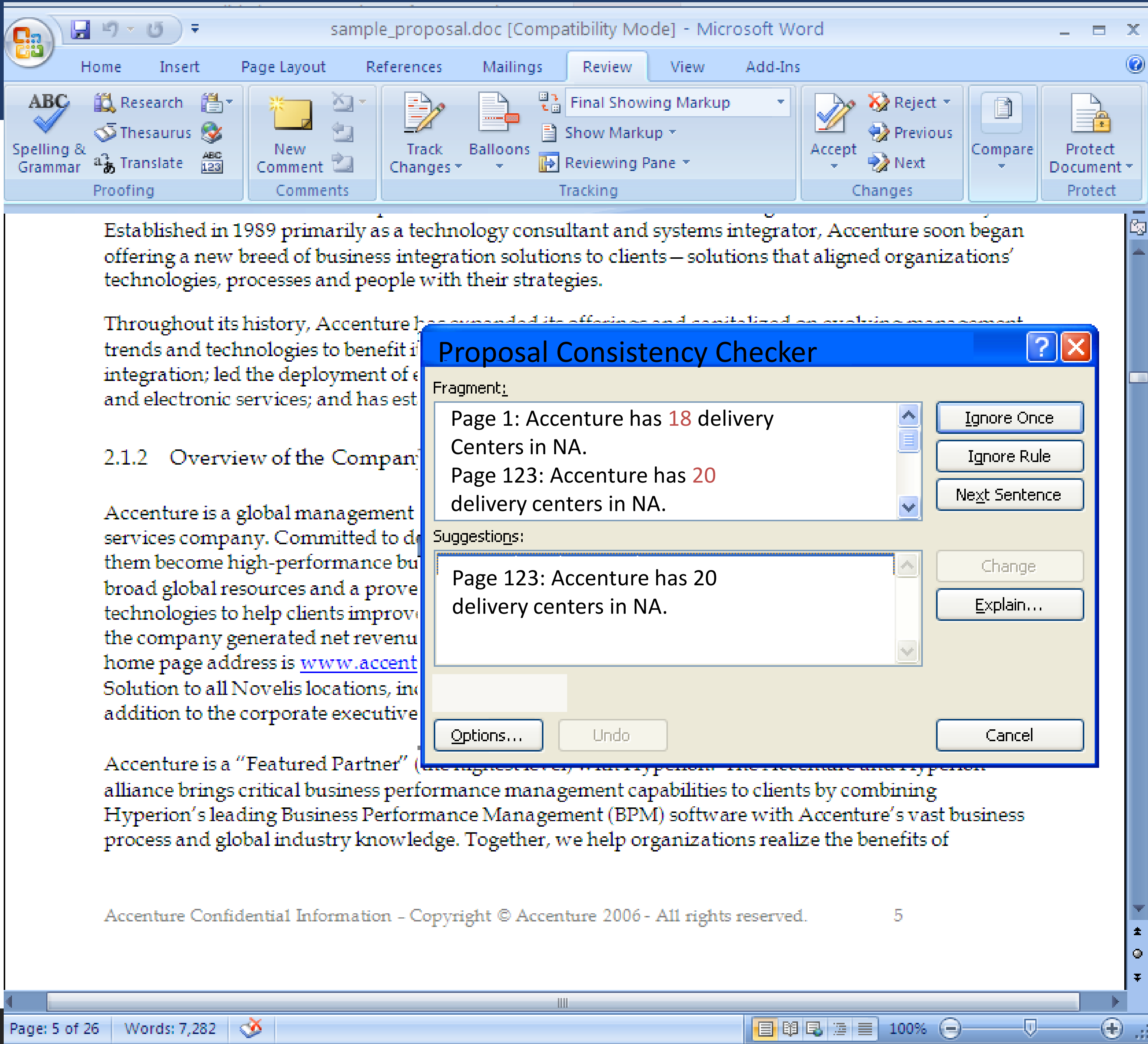
Contextual Access to Information

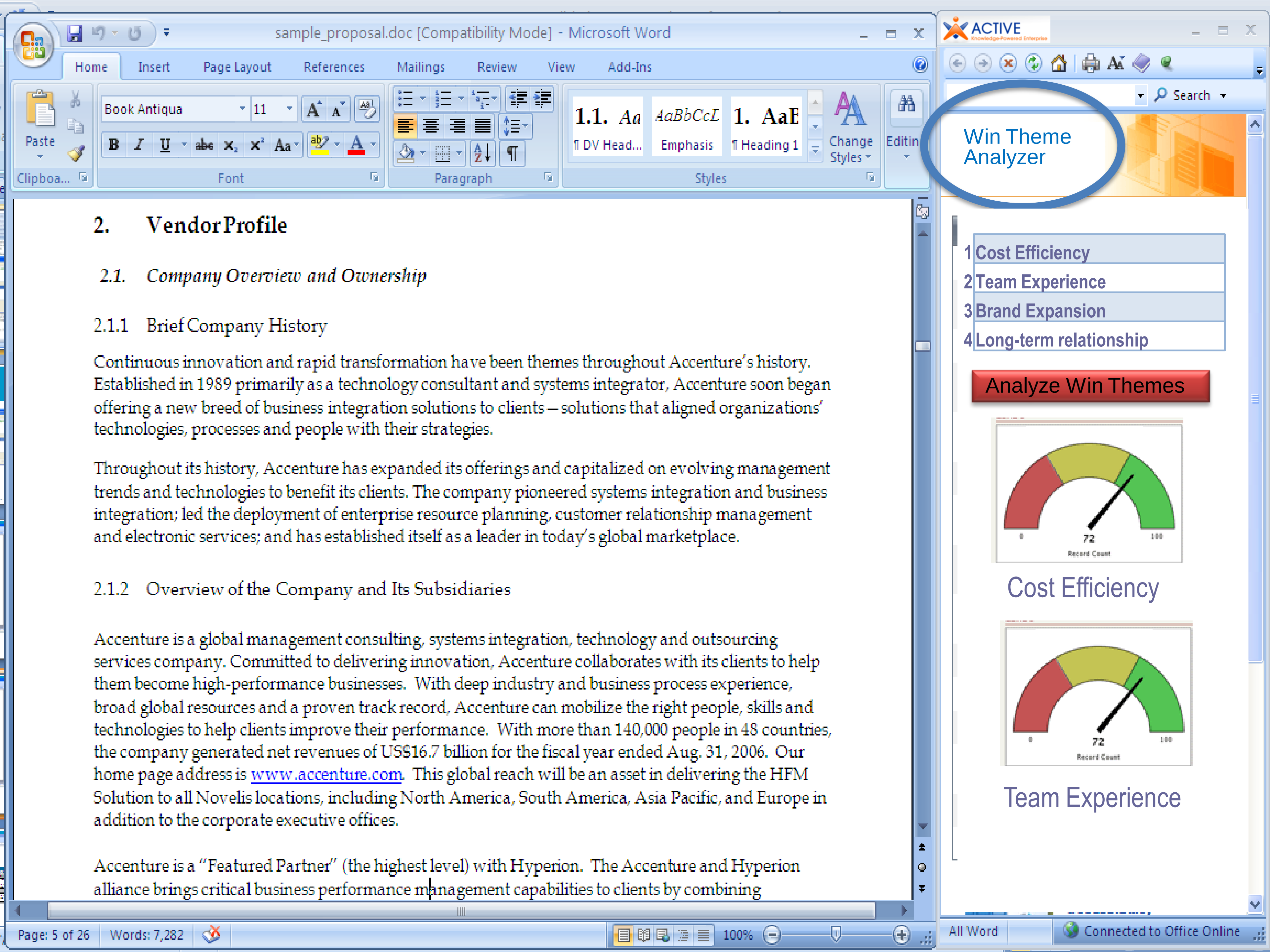
- Current search engines are based on keywords and do not work well with long (sentence-like) querie
- Need t ... s to be use ... s differe ... examp ... rs)
- Long Query (prototype demo)

We are writing a response to a proposal for an online retailer that is exploring new marketing capabilities to improve their customer and marketing results through direct mail, email, web and outbound telemarketing programs. The focus is on flexible delivery of an end-to-end CRM solution, as well as alternate delivery methods that could appeal to the client. The proposed end-to-end option will leverage our alliances with Acxiom, Unica and Teradata

Areas of Improvement




sample_proposal.doc [Compatibility Mode] - Microsoft Word
Home
Insert
Page Layout
References
Mailings
Review
View
Add-Ins
Spelling & Grammar
Research
Thesaurus
Translate
Proofing
New Comment
Comments
Track Changes
Balloons
Final Showing Markup
Show Markup
Reviewing Pane
Tracking
Accept
Reject
Previous
Next
Changes
Compare
Protect Document
Protect
Established in 1989 primarily as a technology consultant and systems integrator, Accenture soon began offering a new breed of business integration solutions to clients – solutions that aligned organizations' technologies, processes and people with their strategies.
Throughout its history, Accenture h
trends and technologies to benefit i
integration; led the deployment of e
and electronic services; and has est
2.1.2 Overview of the Compan
Accenture is a global management
services company. Committed to d
them become high-performance bu
broad global resources and a prove
technologies to help clients improv
the company generated net revenu
home page address is www.accent
Solution to all Novelis locations, in
addition to the corporate executive
Accenture is a "Featured Partner" (
alliance brings critical business performance management capabilities to clients by combining Hyperion's leading Business Performance Management (BPM) software with Accenture's vast business process and global industry knowledge. Together, we help organizations realize the benefits of
Accenture Confidential Information - Copyright © Accenture 2006 - All rights reserved.
5
Proposal Consistency Checker
Fragment:
Page 1: Accenture has 18 delivery Centers in NA.
Page 123: Accenture has 20 delivery centers in NA.
Ignore Once
Ignore Rule
Next Sentence
Suggestions:
Page 123: Accenture has 20 delivery centers in NA.
Change
Explain...
Options...
Undo
Cancel
Page: 5 of 26
Words: 7,282
100%



2. Vendor Profile

2.1. *Company Overview and Ownership*

2.1.1 Brief Company History

Continuous innovation and rapid transformation have been themes throughout Accenture's history. Established in 1989 primarily as a technology consultant and systems integrator, Accenture soon began offering a new breed of business integration solutions to clients – solutions that aligned organizations' technologies, processes and people with their strategies.

Throughout its history, Accenture has expanded its offerings and capitalized on evolving management trends and technologies to benefit its clients. The company pioneered systems integration and business integration; led the deployment of enterprise resource planning, customer relationship management and electronic services; and has established itself as a leader in today's global marketplace.

2.1.2 Overview of the Company and Its Subsidiaries

Accenture is a global management consulting, systems integration, technology and outsourcing services company. Committed to delivering innovation, Accenture collaborates with its clients to help them become high-performance businesses. With deep industry and business process experience, broad global resources and a proven track record, Accenture can mobilize the right people, skills and technologies to help clients improve their performance. With more than 140,000 people in 48 countries, the company generated net revenues of US$16.7 billion for the fiscal year ended Aug. 31, 2006. Our home page address is www.accenture.com. This global reach will be an asset in delivering the HFM Solution to all Novelis locations, including North America, South America, Asia Pacific, and Europe in addition to the corporate executive offices.

Accenture is a "Featured Partner" (the highest level) with Hyperion. The Accenture and Hyperion alliance brings critical business performance management capabilities to clients by combining

Areas of Improvement

Enterprise KM

Access to specific content instead of large monolithic documents

Task and Context-sensitive access to knowledge

Document Analysis and Review Tools

Process-based Support

Publishing Enterprise Content Today

Gather
- KM team gathers project materials
- Filters out “irrelevant” stuff

Upload
- They bundle up docs & tag them
- Bundles are uploaded

Tag
- Create indices and tag with enterprise taxonomies to organize bundles

Copyright © 2007 Accenture All Rights Reserved.

Encouraging Content Submission

- Constantly monitor laptop hard drives and index new documents
- Classification system to tag new documents

Some Metadata fields

Industry	• Automotive, Financial, Electronics, Government, Retail, Forestry, Products,…
Offerings (OGs)	• Products, Strategy, CRM, Supply Chain, Resources,…
Doc type	• Offering, Proposal, Credential,…
Business Function	• Accounting, Learning, HR, Outsourcing, Program Management,…

Level 2 subcategories branch out x10

Ordering Sequential Prediction Tasks

- Determine order of tasks to maximize overall performance
- Intractable
- Approximation (Lad et al. SDM 2009):
  - Find Pairwise Preferences
  - Combine to form an optimal ordering

Distributed Active Learning

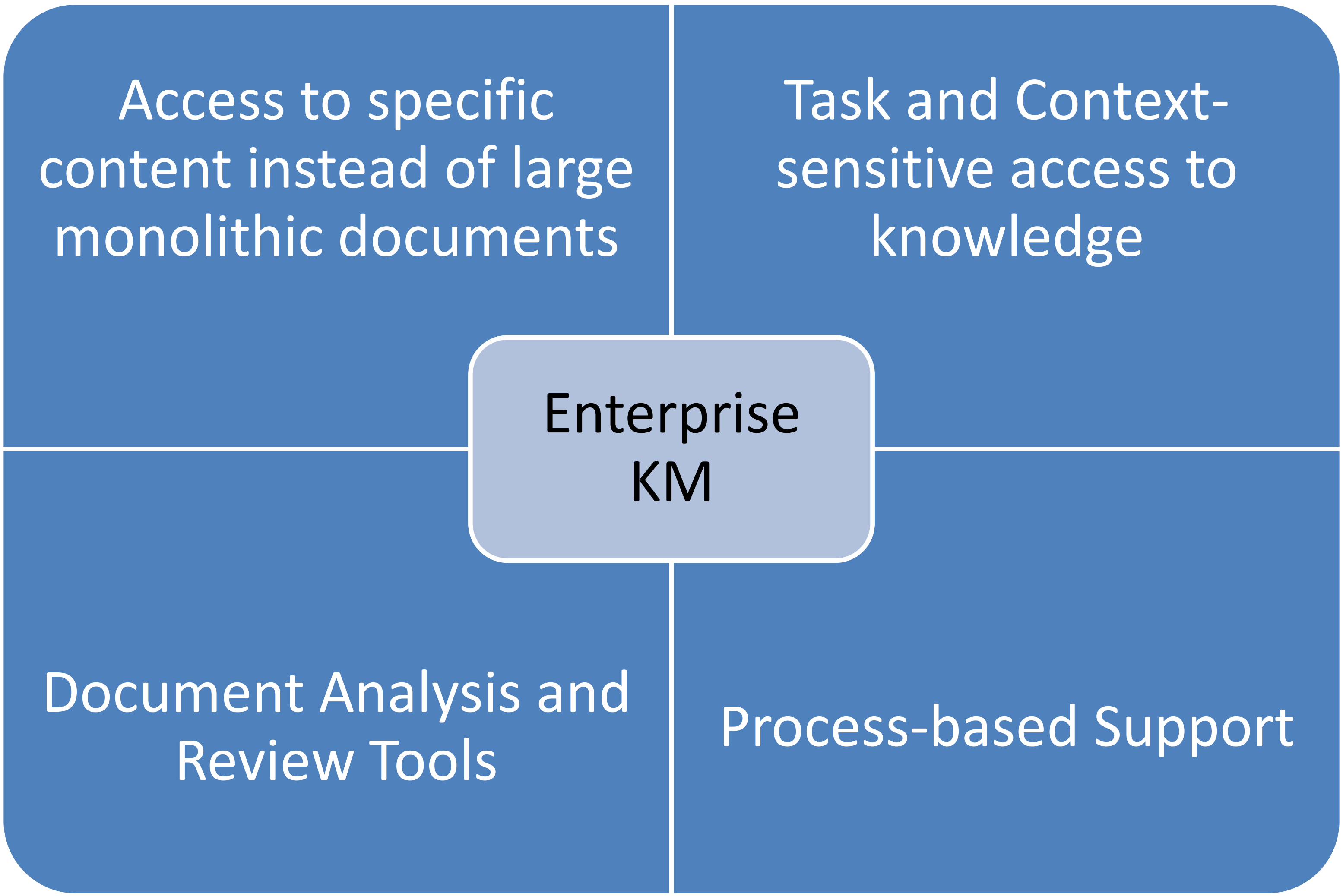
- Querying the user offers many interesting problems for active learning
  - Online only? No pool of instances available.
  - Query for all contributions by a user – Structured example?
    - Constraints/Dependencies among labels for contributions

- Querying KM team members might even have structure

Confidentiality Issues

- Structured Data – Guessing Anonymity (Rachlin et al.PinKDD 2008)
- Challenge: Unlike structured data, text does not contain identified sensitive attributes
- Tension: Redact documents while still preserving some aspect of the utility (Cumby, ECML 2009 Demo)
- K-confusability

Summary


Access to specific content instead of large monolithic documents
Task and Context-sensitive access to knowledge
Enterprise KM
Document Analysis and Review Tools
Process-based Support

Enterprise KM: A different perspective

The Consumer’s Life

The Consumer's Life

Crowd:

Loose Organization
Scalable Efforts
Transparency

Te
Na
Bo
Co

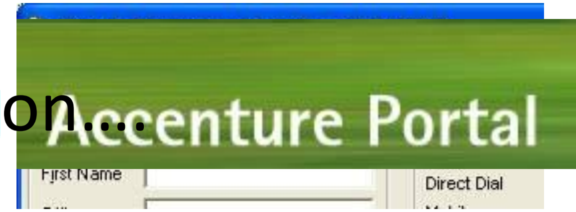
Knowledge:

Pervasive
Dense
Rich
Accessible

Work Life

Work Life

Staffing…Communities…



accenture

Advanced Search

Knowledge:

- Sparse
- Siloed
- Stale

Te

A

R

Internal

Crowd:

A burden, rather than an opportunity.

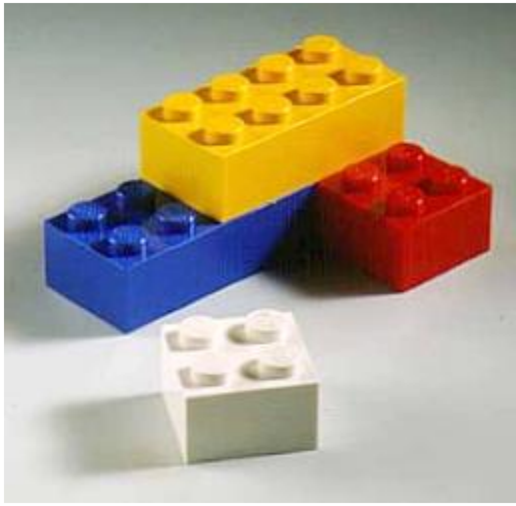
Personalized Knowledge Models

Vs.

MyKM

In the enterprise: Top Down:
"This is what you will use"

In the outside world: Bottom Up:
"I choose or make what I'll use"



"Lego blocks" that will let our community make the tools they can use to provide and consume enterprise information.

Clothing Retailer

Product Database : Table

	SKU	UPC	Item	Quantity	Price
	005654	1212234455	DKNYJTee	1	$49.00

Lauren by Ralph Lauren Catlin Twill Pant
$89.50

A sophisticated pant in substantial cotton twill is contoured for a sleek silhouette.

details | **shipping & returns**

- Cotton
- Machine washable
- Imported
- Web ID: 273846

Overall Rating ★★★★½ **4.5** out of 5

6 of **6** (100%) customers would recommend this product to a friend.

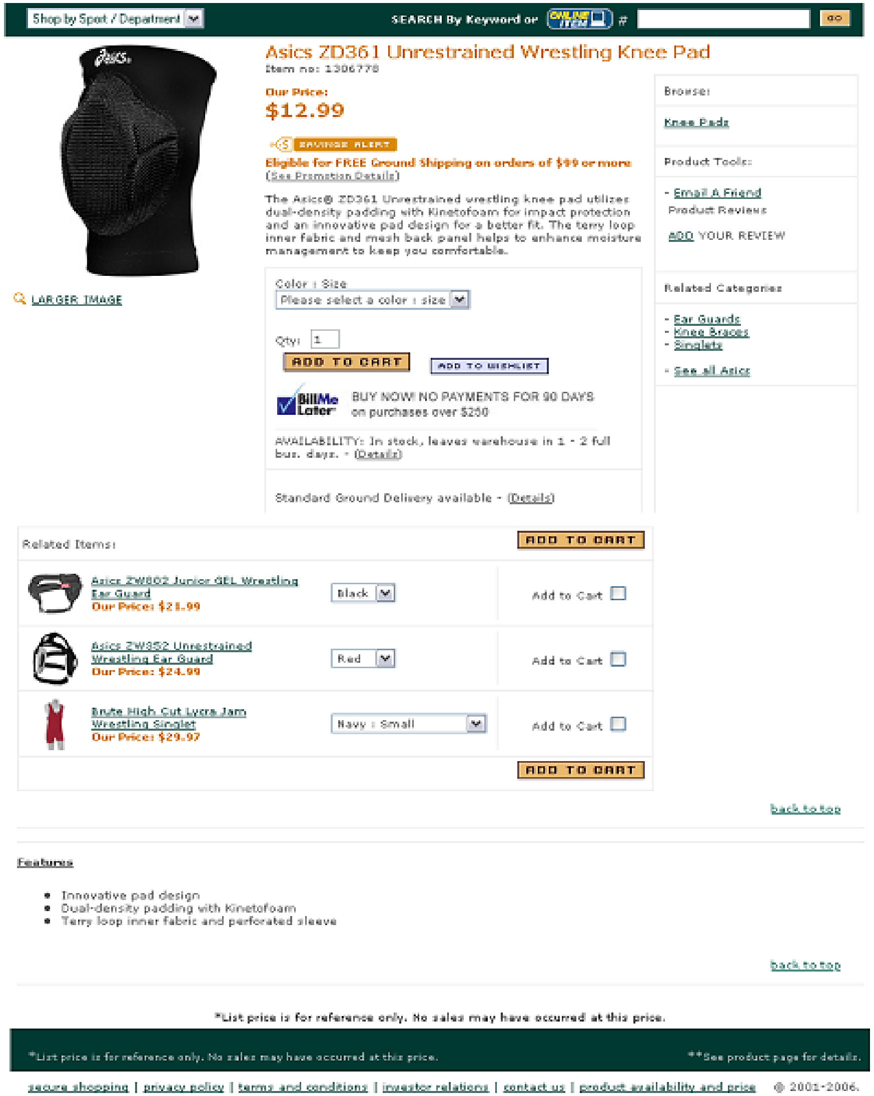
Read all reviews   Write a review

Share this Product:

size	color	qty
Select Size	Select Color	1

size chart   reset selection

Sporting Goods



Sporting Goods

The Asics® ZD361 Unrestrained wrestling knee pad utilizes dual-density padding with Kinetofoam for impact protection and an innovative pad design for a better fit. The terry loop inner fabric and mesh back panel helps to enhance moisture management to keep you comfortable.

Attribute	***Value***
padding	**dual-density**
#material# inner fabric	**terry loop**
sleeve	**perforated**
pad design	**innovative**

- Dual-density padding with Kinetofoam
- Terry loop inner fabric and perforated sleeve
- Innovative pad design

Challenge: How do you describe these products?

Extreme Case: eBay

All Categories

Search

Search titles & descriptions

Browse Categories

Category: All Categories | Format: All Items | Listings: All Active | Location: Available on: eBay.com | Show

Show number of items in category | Show category numbers

Antiques (213451)
- Antiquities (5061)
- Architectural & Garden (13710)
- Asian Antiques (25012)
- Books & Manuscripts (6562)
- Decorative Arts (33460)
- Ethnographic (4220)
- Furniture (16049)
- Home & Hearth (913)
- Linens & Textiles (Pre-1930) (9232)
- Maps, Atlases & Globes (9777)
- Maritime (2634)
- Mercantile, Trades & Factories (744)
- Musical Instruments (Pre-1930) (341)
- Periods & Styles (5977)
- Primitives (13542)
- Restoration & Care (46)
- Rugs & Carpets (19060)
- Science & Medicine (Pre-1930) (1805)
- Sewing (Pre-1930) (2260)
- Silver (36947)
- Reproduction Antiques (1369)
- Other (4730)

See all Antiques categories...

Collectibles (1974383)
- Advertising (132014)
- Animals (93120)
- Animation Art & Characters (94804)
- Arcade, Jukeboxes & Pinball (12238)
- Autographs (8956)
- Banks, Registers & Vending (5340)
- Barware (10342)
- Beads (1514)
- Bottles & Insulators (13618)
- Breweriana, Beer (31822)
- Casino (15407)
- Clocks (11449)
- Comics (147817)
- Cultures & Ethnicities (59523)
- Decorative Collectibles (205113)
- Disneyana (67015)
- Fantasy, Mythical & Magic (22536)
- Historical Memorabilia (92234)
- Holiday & Seasonal (35719)
- Kitchen & Home (57285)
- Knives, Swords & Blades (65112)
- Lamps, Lighting (17714)
- Linens, Fabric & Textiles (22115)
- Metalware (9781)

Jewelry & Watches (2105770)
- Children's Jewelry (13765)
- Designer Brands (38434)
- Engagement & Wedding (105485)
- Ethnic, Regional & Tribal (49543)
- Fashion Jewelry (831412)
- Fine Jewelry (121625)
- Handcrafted, Artisan Jewelry (67967)
- Jewelry Boxes, Cases & Display (25253)
- Jewelry Design & Repair (57839)
- Loose Beads (211374)
- Loose Diamonds & Gemstones (100458)
- Men's Jewelry (70898)
- Vintage & Antique Jewelry (112976)
- Watches (255139)
- Other (3821)
- Wholesale Lots (39781)

See all Jewelry & Watches categories...

Music (1852118)
- Accessories (3390)
- Cassettes (17288)
- CDs (1534322)
- DVD Audio (1403)
- Records (256823)

Having a product attribute database allows….

- Transfer learning from campaigns to related products and categories and **scale** to a large number of products efficiently

- Dealing with the Cold-start problem especially for fast-changing categories

- Ad networks and aggregators to learn across client

ebay®
The World's Online Marketplace®

YAHOO!® SHOPPING AUCTIONS

craigslist

amazon.com.

onSale®
World's Premier Marketplace™

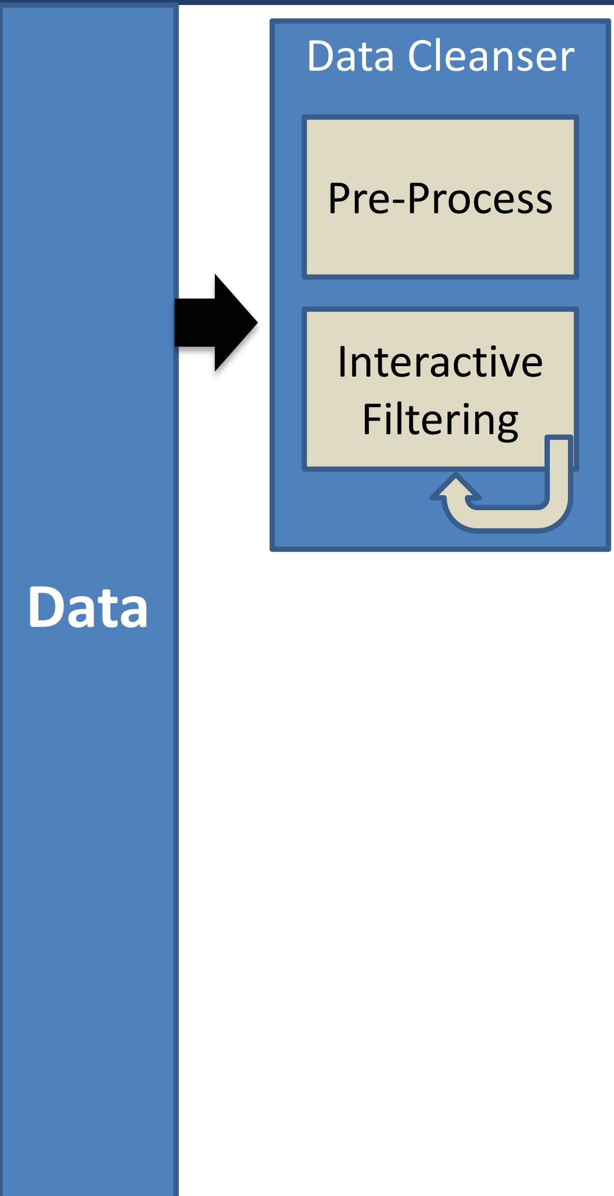
uBid.
The Brand Name Marketplace

What can businesses do with such enriched databases?

- **Assortment optimization**
  - **Sample Questions: Do people buy Bic Gel Grip Pen because of the brand being Bic or the Gel grip?**
- **Supply chain management**
  - **Sample Question: Are these two products comparable?**
- **Procurement**
  - **Sample Question: What should I buy and from whom?**
- **Marketing**
  - **Sample Question: How should I talk about my products?**
- **Competitive Intelligence:**
  - **Sample Question: How many high-end TVs does my competitor sell as compared to me?**
- **Product lifecycle planning**
  - **Sample Question: How do products change over time?**

Challenges

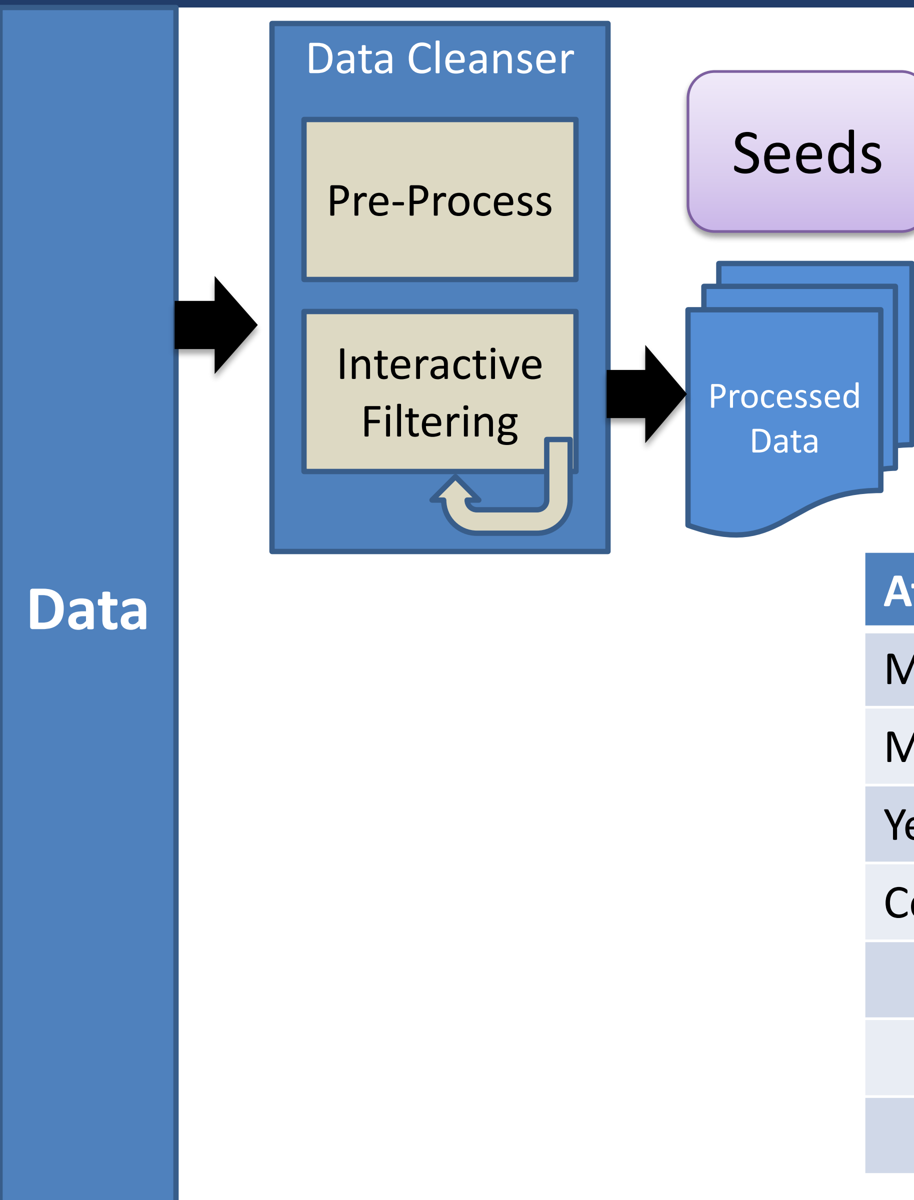
- Signal/noise ratio
  - Extra information
- Labeled data is noisy/misleading
- Heterogeneous sources (title, desc, tables)
- Adhoc sellers
- Typos
- Loooooong sentences


Data
Data Cleanser
Pre-Process
Interactive Filtering

Pre-Processing

Extract text content from Source data while preserving formatting and layout → Remove Extraneous Sections → Remove Extraneous words (and text following them) → Remove Low-Frequency Words (2<n<10) → Interactive Description Processing using unsupervised clustering (10<k<50)

Keep	Cluster ID	Descriptive Words	Size
☐	1	guarantee, free, worldwide, ship, anywhere	35432
X	2	Chevy, grand, front, wheel, silver	21211
X	3	Tire, rubber, tread, michelin, goodyear	5000
X	4	#number#, model, make, year, all	4323
☐	5	Return, carefully, package, accepted	3244



Attributes	Values	Neither
Make	2000	And
Model	Fiat	Is
Year	Bmw	Over
Condition	Goodyear	…
	…	…
	…	…
	…	…

Unsupervised Seed Augmentation

- Goal: High Precision and Low Recall

- Intuition: Attributes and Values often occur as pairs of consecutive words where the 1st word is the value and 2nd is the attribute
  - Back pocket, side pocket, front pocket

- 2^{nd} word shouldn't occur with the same word every time (phrase) and also shouldn't occur with "too many" words
  - Pittsburgh Steelers, Fifth Avenue, running across

Unsupervised Seed Generation

- All bigrams $w_j w$ are considered candidates for attribute-value pairs:
  - w_j is a potential value and w is a potential attribute
- Let w_j (with $0<j<k$) be the set of unique words that occur just before w
- Sort all w_j from highest to lowest $P(w_j|w)$
- Retain all w_j such that the sum of the highest-ranking $p(w_j/w) = z$ (we use $z=0.5$)

- Compute cumulative mutual information:
  - Let $p(w, w_{1...k}) = \sum_{j=1}^{k} p(w, w_j)$. Then,

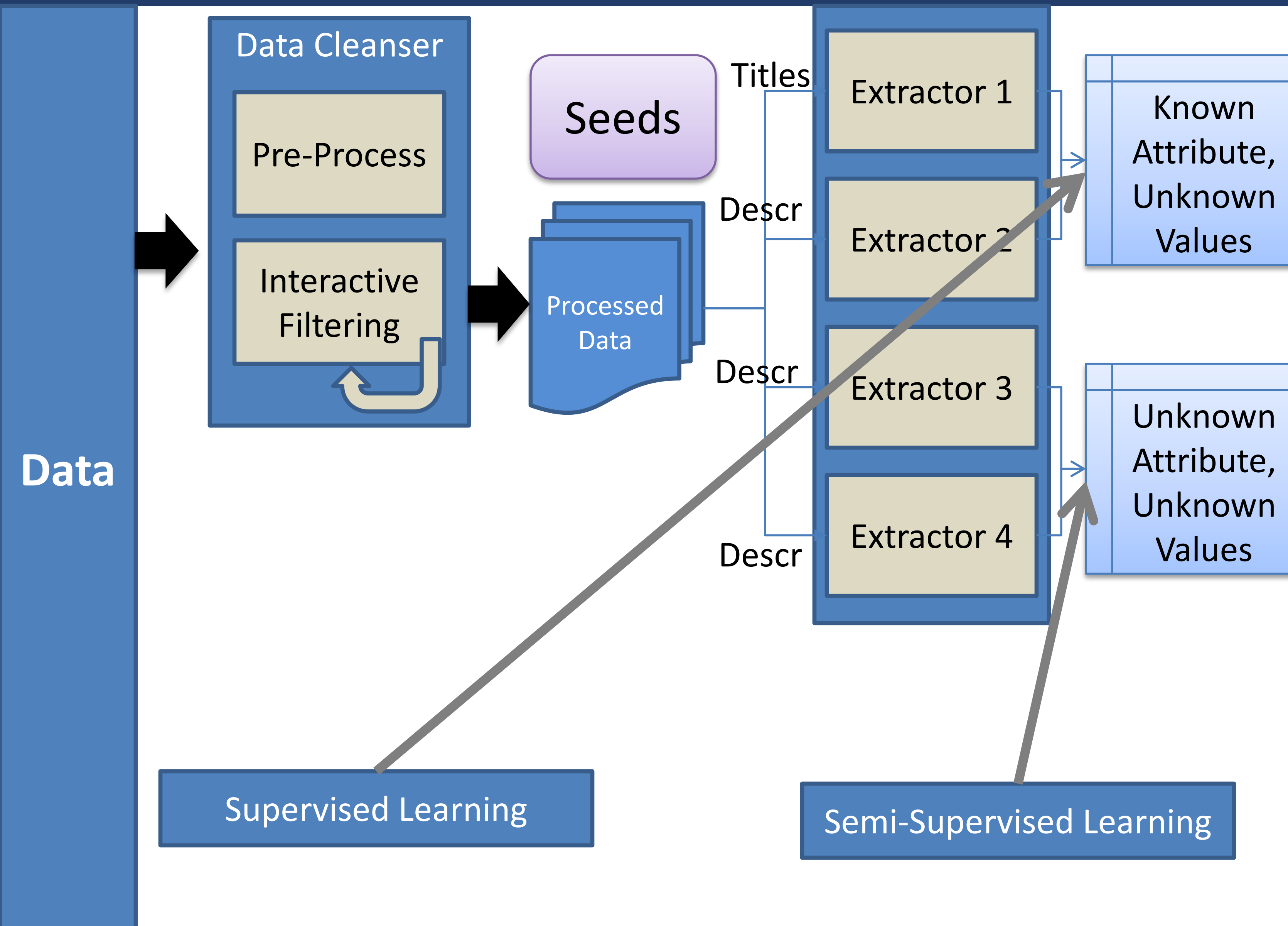
$$cmi(w_{1...k}; w) = \log \frac{p(w, w_{1...k})}{(\lambda * \sum_{j=1}^{k} p(w_j)) * ((\lambda - 1) * p(w))}$$

where $0 < \lambda < 1$.

- Retain as pairs those whose c.m.i. exceeds a threshold.

Examples of extracted attribute-value pairs

Attribute	**Values**
case	carrying, storage
compartment	main, racquet
pocket	ball, welt, side-seam, key


Data
Data Cleanser
Pre-Process
Interactive Filtering
Seeds
Processed Data
Titles
Descr
Descr
Descr
Extractor 1
Extractor 2
Extractor 3
Extractor 4
Known Attribute, Unknown Values
Unknown Attribute, Unknown Values
Supervised Learning
Semi-Supervised Learning

Titles

- Supervised Learning
- Sequential Classification

Descriptions

- Supervised Learning from Title: Apply to Description
- Supervised Learning from Title: Apply to Description
- Semi-Supervised Learning from Title: Apply to Description
- Learning Patterns from Tables

Classification: Co-EM with Naïve

- Split data into two views:
  - View1: stemmed word itself and POS tag
  - View2: 8-word context and their POS tags

- Training data for first classification iteration:
  - Seeds extracted in previous step
  - Lists of generic attributes: countries, colors, materials

Classification: Co-EM with Naïve Bayes

- General Co-EM procedure:
  1. Initialize based on labeled data
  2. Train *view1* classifier, label *view1* of unlabeled data
  3. Use *view1* labels to train *view2* classifier
  4. Label *view2* of unlabeled data using view2 classifier
  5. Repeat 2, 3, and 4
  6. Use both classifiers to get final probabilities on unlabeled data

Classification: Co-EM with Naïve Bayes

- When labeling *view2, e*stimate word and class probabilities using:
  - Current assignments of classes to *view1* words (use probability distribution over all classes)
  - Co-occurrence counts of *view1* words with each *view2* word (how often did a *view1* word appear in *view1*'s context)
- Reverse procedure when labeling *view1*

Training View2 from View1

- Estimate new *view2* class prior probabilities:

$$P(c_k) = \frac{1 + \sum_i^{n1} cnt(view1_i) * P(c_k|view1_i)}{numclasses + \sum_i^{n1} cnt(view1_i)}$$

- Estimate new *view2* word probabilities:

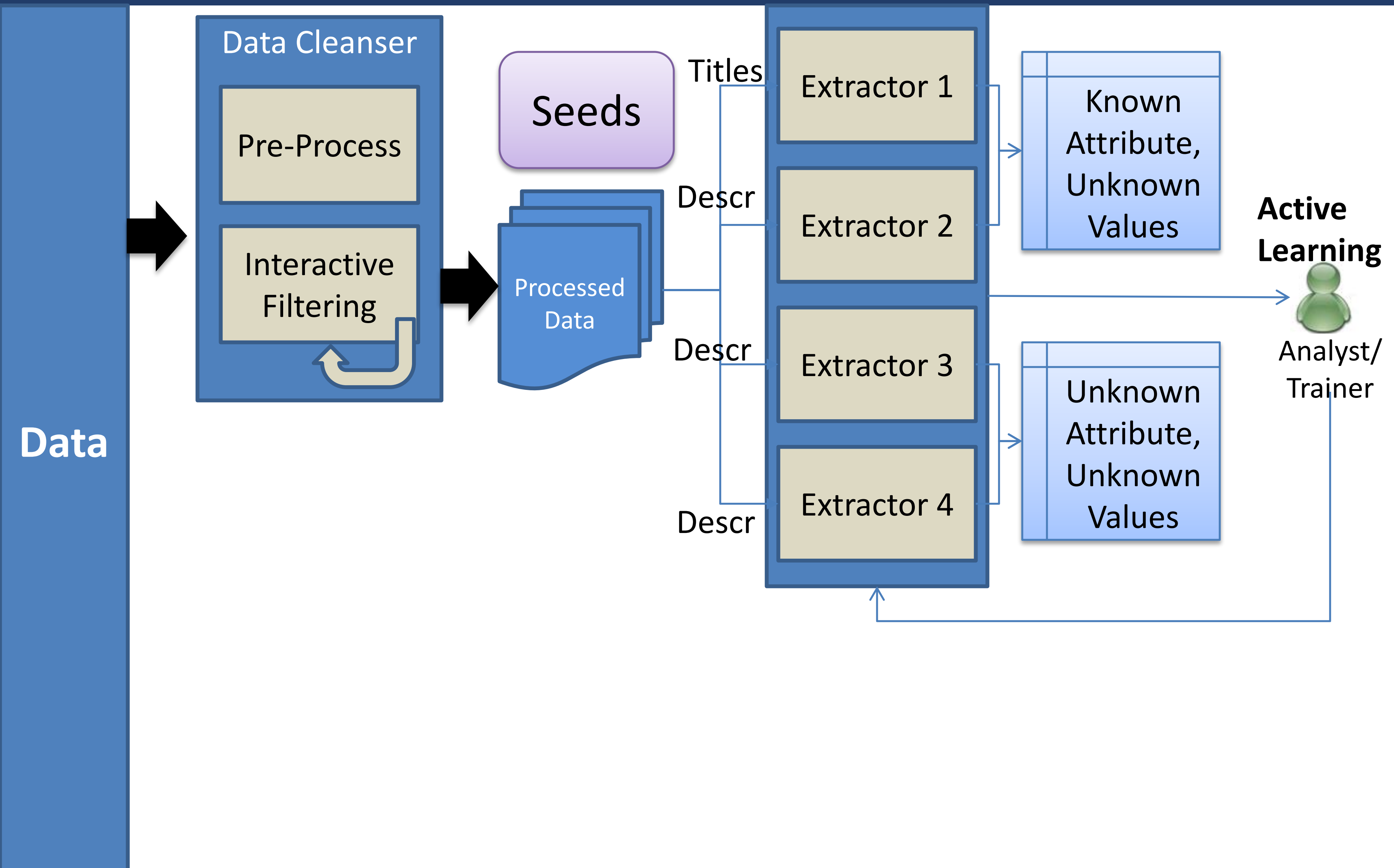
$$P(view2_j|c_k) = \frac{1 + \sum_{i=1}^{n1} cooc(view1_i, view2_j) * P(c_k|view1_i)}{n_2 + \sum_{i=1}^{n1} cooc(view1_i, view2_j)}$$

- Label *view2* words:

$$P(c_k|view2_i) \propto P(c_k) * P(view2_i|c_k)$$

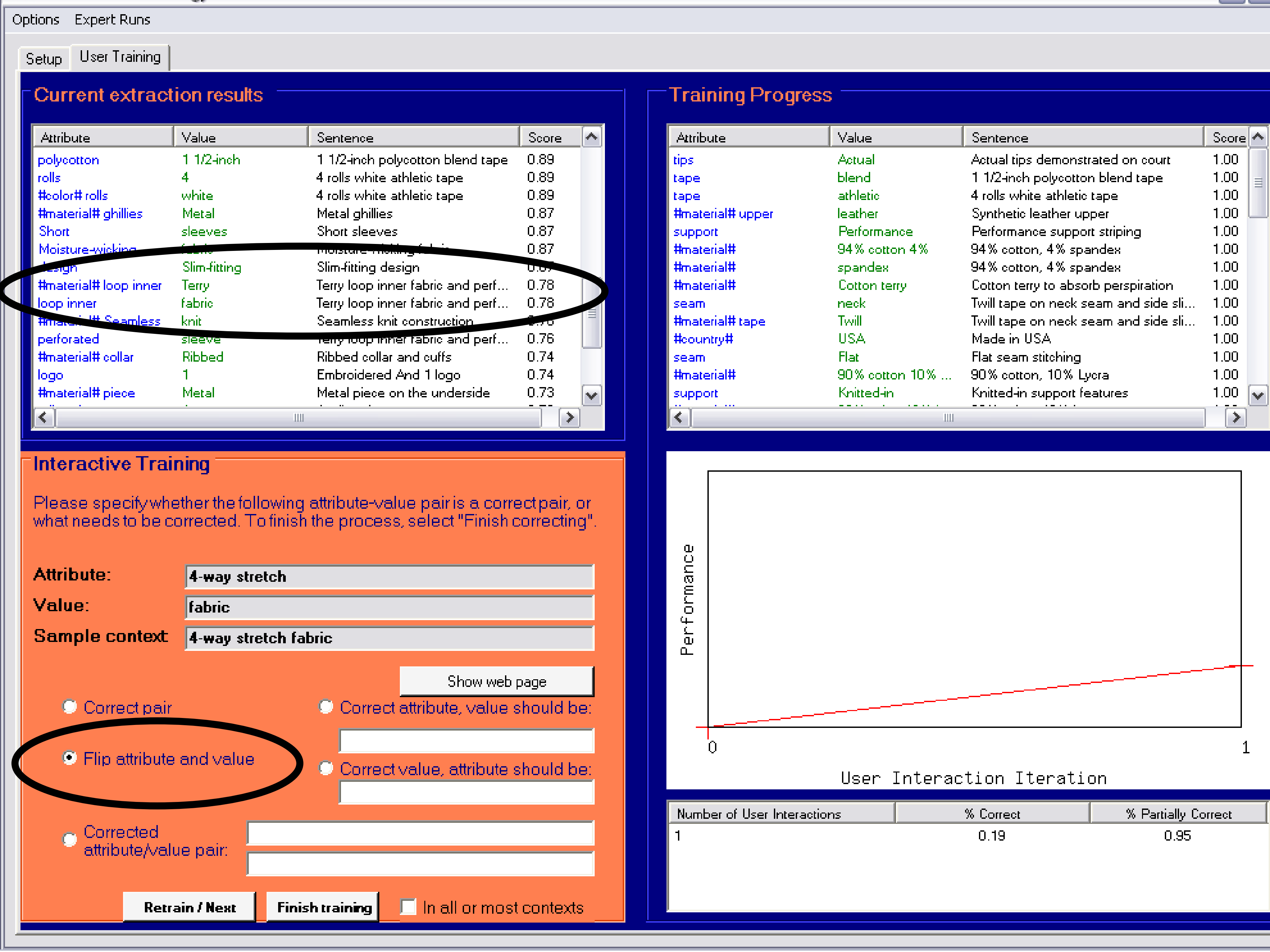
- Use both view1 and view2 classifiers to label unlabeled data

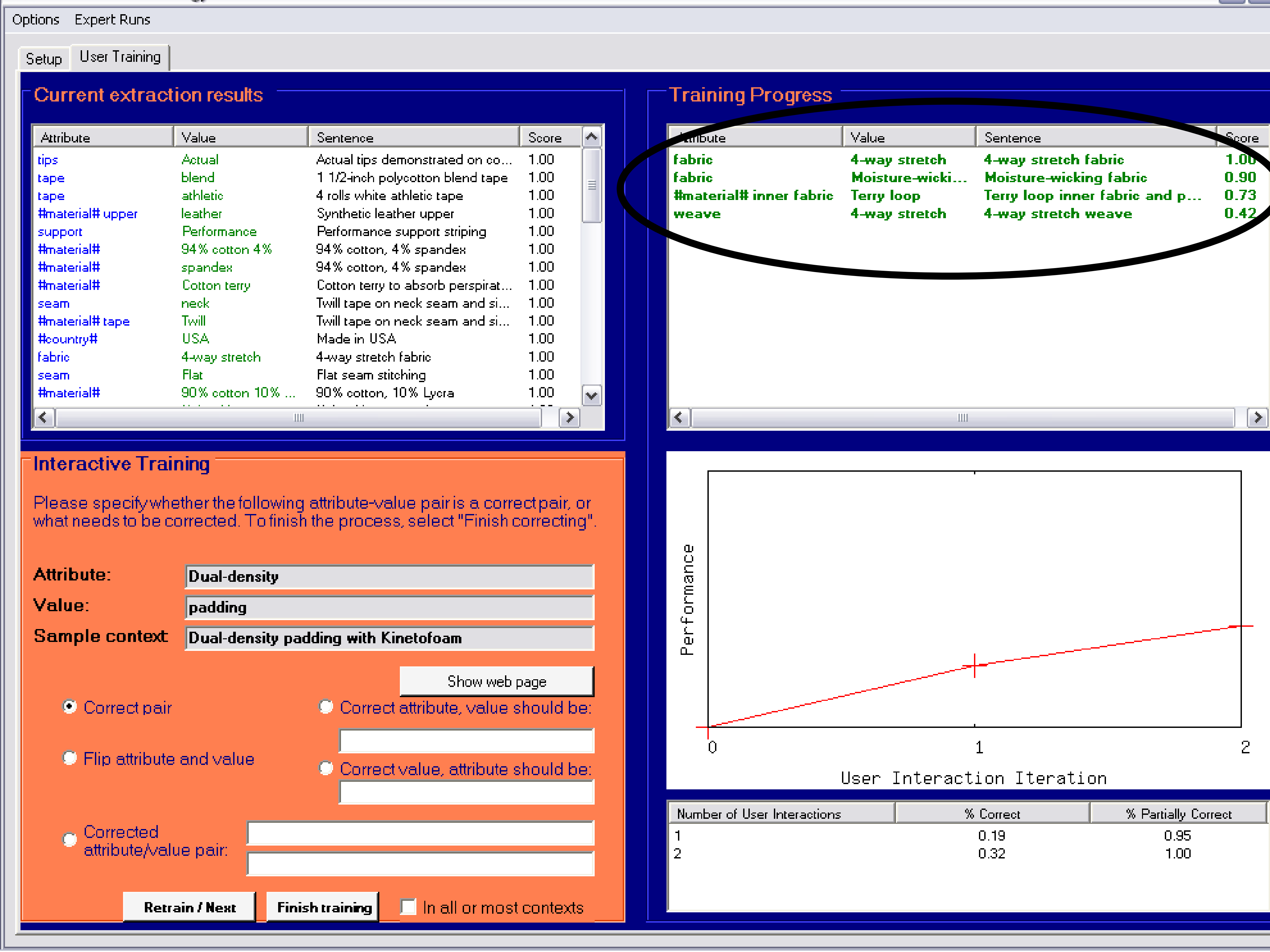
$$P(c_k \mid < view1_i, view2_j >) = \frac{P(c_k \mid view1_i) + P(c_k \mid view2_j)}{2}$$


Data
Data Cleanser
Pre-Process
Interactive Filtering
Seeds
Processed Data
Titles
Descr
Descr
Descr
Extractor 1
Extractor 2
Extractor 3
Extractor 4
Known Attribute, Unknown Values
Unknown Attribute, Unknown Values
Active Learning
Analyst/ Trainer

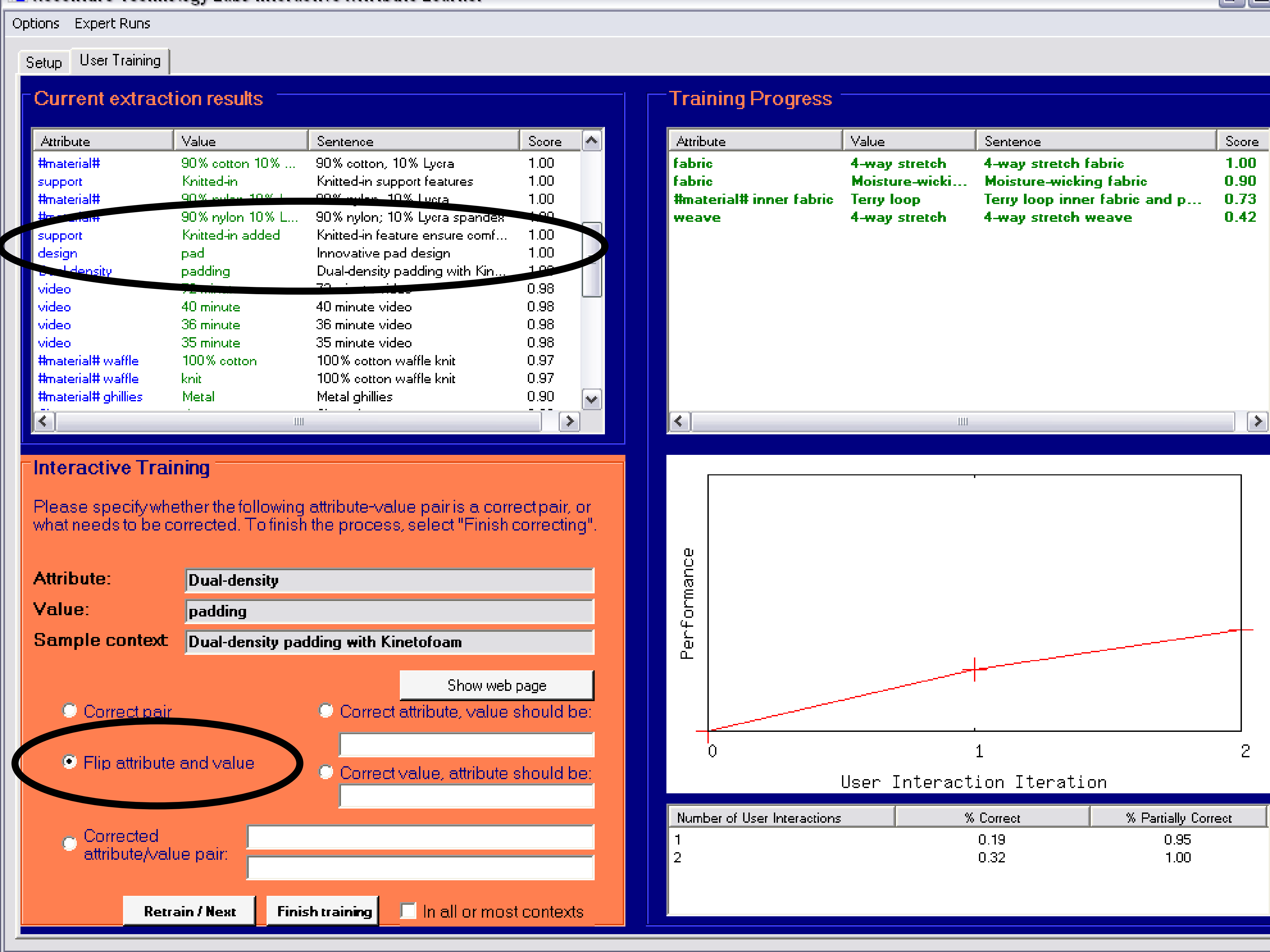
Active Learning

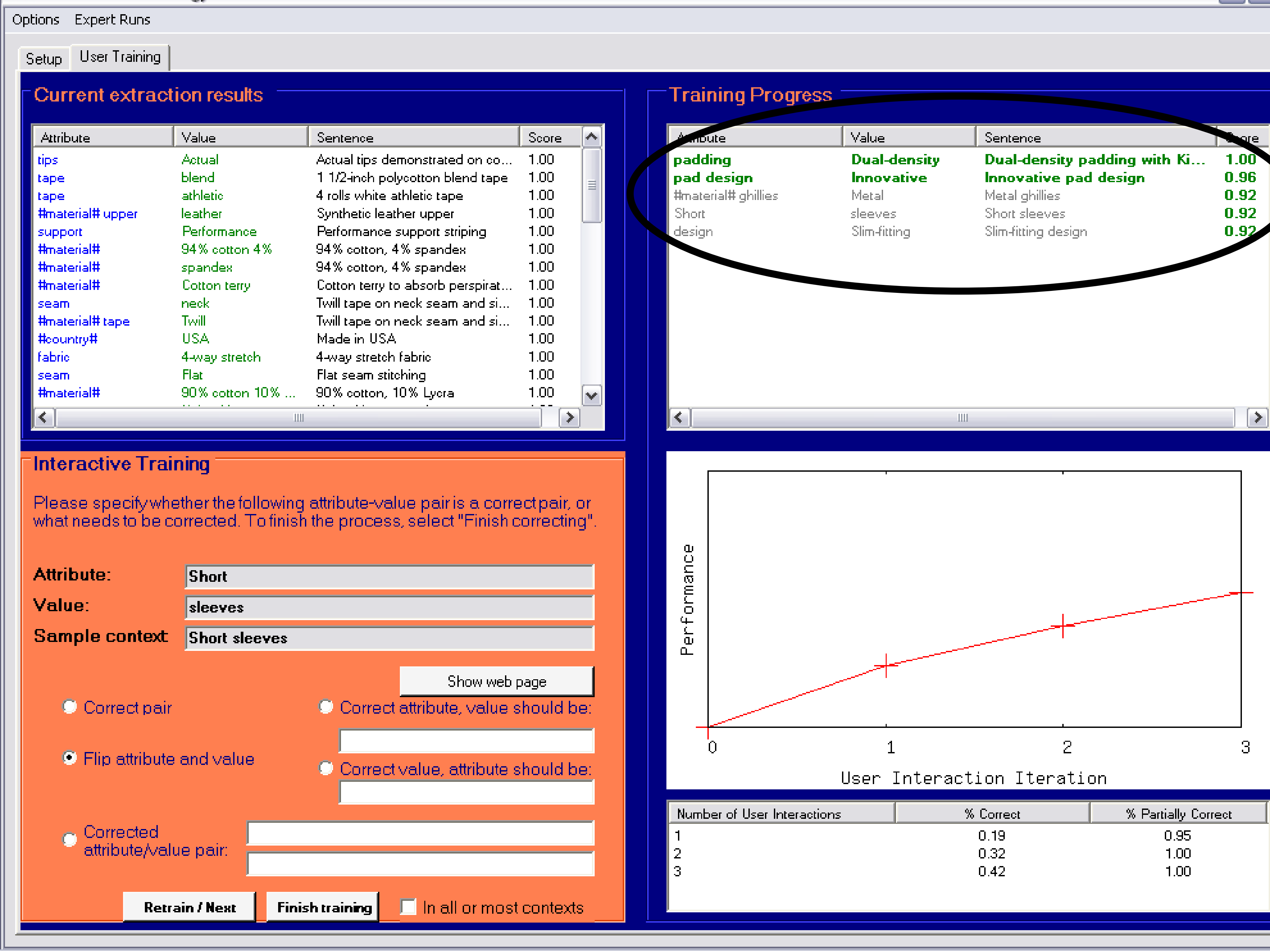
- Muli-view Semi-supervised + Active Learning
- Variety of metrics
  - Frequency
  - KL
  - Frequency weighted KL
- Interactive version of co-EM + NB
  - Orders of magnitude faster (from 15 minutes to 10 seconds)


Options Expert Runs
Setup
User Training
Current extraction results
Attribute | Value | Sentence | Score
polycotton | 1 1/2-inch | 1 1/2-inch polycotton blend tape | 0.89
rolls | 4 | 4 rolls white athletic tape | 0.89
#color# rolls | white | 4 rolls white athletic tape | 0.89
#material# ghillies | Metal | Metal ghillies | 0.87
Short | sleeves | Short sleeves | 0.87
Moisture-wicking | fabric | Moisture-wicking fabric | 0.87
design | Slim-fitting | Slim-fitting design
#material# loop inner | Terry | Terry loop inner fabric and perf... | 0.78
loop inner | fabric | Terry loop inner fabric and perf... | 0.78
#material# Seamless | knit | Seamless knit construction
perforated | sleeve | Terry loop inner fabric and perf... | 0.76
#material# collar | Ribbed | Ribbed collar and cuffs | 0.74
logo | 1 | Embroidered And 1 logo | 0.74
#material# piece | Metal | Metal piece on the underside | 0.73
Training Progress
Attribute | Value | Sentence | Score
tips | Actual | Actual tips demonstrated on court | 1.00
tape | blend | 1 1/2-inch polycotton blend tape | 1.00
tape | athletic | 4 rolls white athletic tape | 1.00
#material# upper | leather | Synthetic leather upper | 1.00
support | Performance | Performance support striping | 1.00
#material# | 94% cotton 4% | 94% cotton, 4% spandex | 1.00
#material# | spandex | 94% cotton, 4% spandex | 1.00
#material# | Cotton terry | Cotton terry to absorb perspiration | 1.00
seam | neck | Twill tape on neck seam and side sli... | 1.00
#material# tape | Twill | Twill tape on neck seam and side sli... | 1.00
#country# | USA | Made in USA | 1.00
seam | Flat | Flat seam stitching | 1.00
#material# | 90% cotton 10% ... | 90% cotton, 10% Lycra | 1.00
support | Knitted-in | Knitted-in support features | 1.00
Interactive Training
Please specify whether the following attribute-value pair is a correct pair, or what needs to be corrected. To finish the process, select "Finish correcting".
Attribute: 4-way stretch
Value: fabric
Sample context 4-way stretch fabric
Show web page
Correct pair
Correct attribute, value should be:
Flip attribute and value
Correct value, attribute should be:
Corrected attribute/value pair:
Retrain / Next
Finish training
In all or most contexts
Performance
0
1
User Interaction Iteration
Number of User Interactions | % Correct | % Partially Correct
1 | 0.19 | 0.95



Options Expert Runs
Setup
User Training
Current extraction results
Attribute | Value | Sentence | Score
tips | Actual | Actual tips demonstrated on co... | 1.00
tape | blend | 1 1/2-inch polycotton blend tape | 1.00
tape | athletic | 4 rolls white athletic tape | 1.00
#material# upper | leather | Synthetic leather upper | 1.00
support | Performance | Performance support striping | 1.00
#material# | 94% cotton 4% | 94% cotton, 4% spandex | 1.00
#material# | spandex | 94% cotton, 4% spandex | 1.00
#material# | Cotton terry | Cotton terry to absorb perspirat... | 1.00
seam | neck | Twill tape on neck seam and si... | 1.00
#material# tape | Twill | Twill tape on neck seam and si... | 1.00
#country# | USA | Made in USA | 1.00
fabric | 4-way stretch | 4-way stretch fabric | 1.00
seam | Flat | Flat seam stitching | 1.00
#material# | 90% cotton 10% ... | 90% cotton, 10% Lycra | 1.00
Training Progress
Value | Sentence | Score
fabric | 4-way stretch | 4-way stretch fabric | 1.00
fabric | Moisture-wicki... | Moisture-wicking fabric | 0.90
#material# inner fabric | Terry loop | Terry loop inner fabric and p... | 0.73
weave | 4-way stretch | 4-way stretch weave | 0.42
Interactive Training
Please specify whether the following attribute-value pair is a correct pair, or what needs to be corrected. To finish the process, select "Finish correcting".
Attribute: Dual-density
Value: padding
Sample context Dual-density padding with Kinetofoam
Show web page
Correct pair
Correct attribute, value should be:
Flip attribute and value
Correct value, attribute should be:
Corrected attribute/value pair:
Retrain / Next
Finish training
In all or most contexts
Performance
0
1
2
User Interaction Iteration
Number of User Interactions | % Correct | % Partially Correct
1 | 0.19 | 0.95
2 | 0.32 | 1.00


Options
Expert Runs
Setup
User Training
Current extraction results
Attribute
Value
Sentence
Score
#material#
90% cotton 10% ...
90% cotton, 10% Lycra
1.00
support
Knitted-in
Knitted-in support features
1.00
90% nylon 10% L...
90% nylon; 10% Lycra spandex
support
Knitted-in added
Knitted-in feature ensure comf...
1.00
design
pad
Innovative pad design
1.00
padding
Dual-density padding with Kin...
video
0.98
video
40 minute
40 minute video
0.98
video
36 minute
36 minute video
0.98
video
35 minute
35 minute video
0.98
#material# waffle
100% cotton
100% cotton waffle knit
0.97
#material# waffle
knit
100% cotton waffle knit
0.97
#material# ghillies
Metal
Metal ghillies
0.90
Training Progress
Attribute
Value
Sentence
Score
fabric
4-way stretch
4-way stretch fabric
1.00
fabric
Moisture-wicki...
Moisture-wicking fabric
0.90
#material# inner fabric
Terry loop
Terry loop inner fabric and p...
0.73
weave
4-way stretch
4-way stretch weave
0.42
Interactive Training
Please specify whether the following attribute-value pair is a correct pair, or what needs to be corrected. To finish the process, select "Finish correcting".
Attribute:
Dual-density
Value:
padding
Sample context
Dual-density padding with Kinetofoam
Show web page
Correct pair
Correct attribute, value should be:
Flip attribute and value
Correct value, attribute should be:
Corrected attribute/value pair:
Retrain / Next
Finish training
In all or most contexts
Performance
0
1
2
User Interaction Iteration
Number of User Interactions
% Correct
% Partially Correct
1
0.19
0.95
2
0.32
1.00


Options
Expert Runs
Setup
User Training
Current extraction results
Attribute
Value
Sentence
Score
tips
Actual
Actual tips demonstrated on co...
1.00
tape
blend
1 1/2-inch polycotton blend tape
1.00
tape
athletic
4 rolls white athletic tape
1.00
#material# upper
leather
Synthetic leather upper
1.00
support
Performance
Performance support striping
1.00
#material#
94% cotton 4%
94% cotton, 4% spandex
1.00
#material#
spandex
94% cotton, 4% spandex
1.00
#material#
Cotton terry
Cotton terry to absorb perspirat...
1.00
seam
neck
Twill tape on neck seam and si...
1.00
#material# tape
Twill
Twill tape on neck seam and si...
1.00
#country#
USA
Made in USA
1.00
fabric
4-way stretch
4-way stretch fabric
1.00
seam
Flat
Flat seam stitching
1.00
#material#
90% cotton 10% ...
90% cotton, 10% Lycra
1.00
Training Progress
Value
Sentence
padding
Dual-density
Dual-density padding with Ki...
1.00
pad design
Innovative
Innovative pad design
0.96
#material# ghillies
Metal
Metal ghillies
0.92
Short
sleeves
Short sleeves
0.92
design
Slim-fitting
Slim-fitting design
0.92
Interactive Training
Please specify whether the following attribute-value pair is a correct pair, or what needs to be corrected. To finish the process, select "Finish correcting".
Attribute:
Short
Value:
sleeves
Sample context
Short sleeves
Show web page
Correct pair
Correct attribute, value should be:
Flip attribute and value
Correct value, attribute should be:
Corrected attribute/value pair:
Retrain / Next
Finish training
In all or most contexts
Performance
0
1
2
3
User Interaction Iteration
Number of User Interactions
% Correct
% Partially Correct
1
0.19
0.95
2
0.32
1.00
3
0.42
1.00

Experimental Results

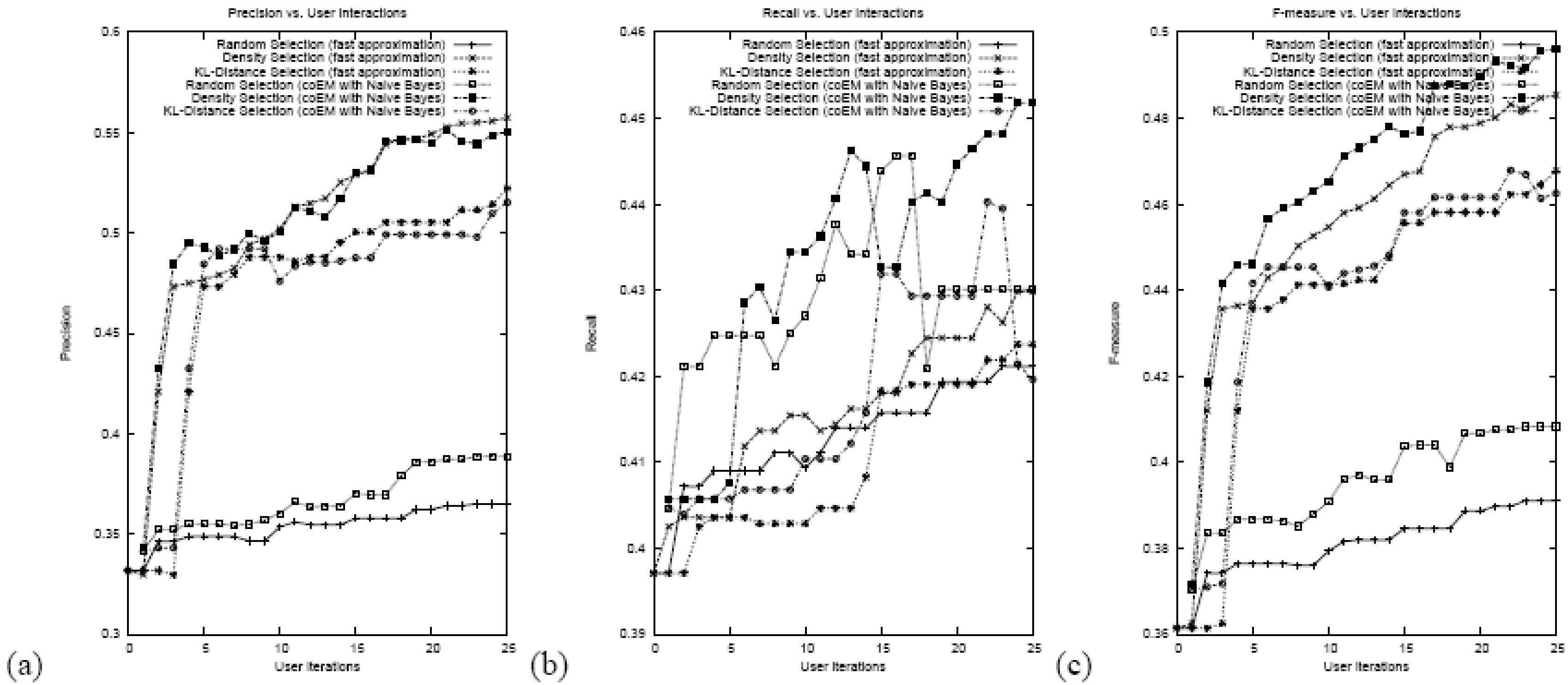


Fig. 1. Precision, Recall, and F-measure for fast algorithm compared to coEM with naive Bayes. The y-value for k indicates the recall, precision, and F-measure after the k^{th} user interaction.

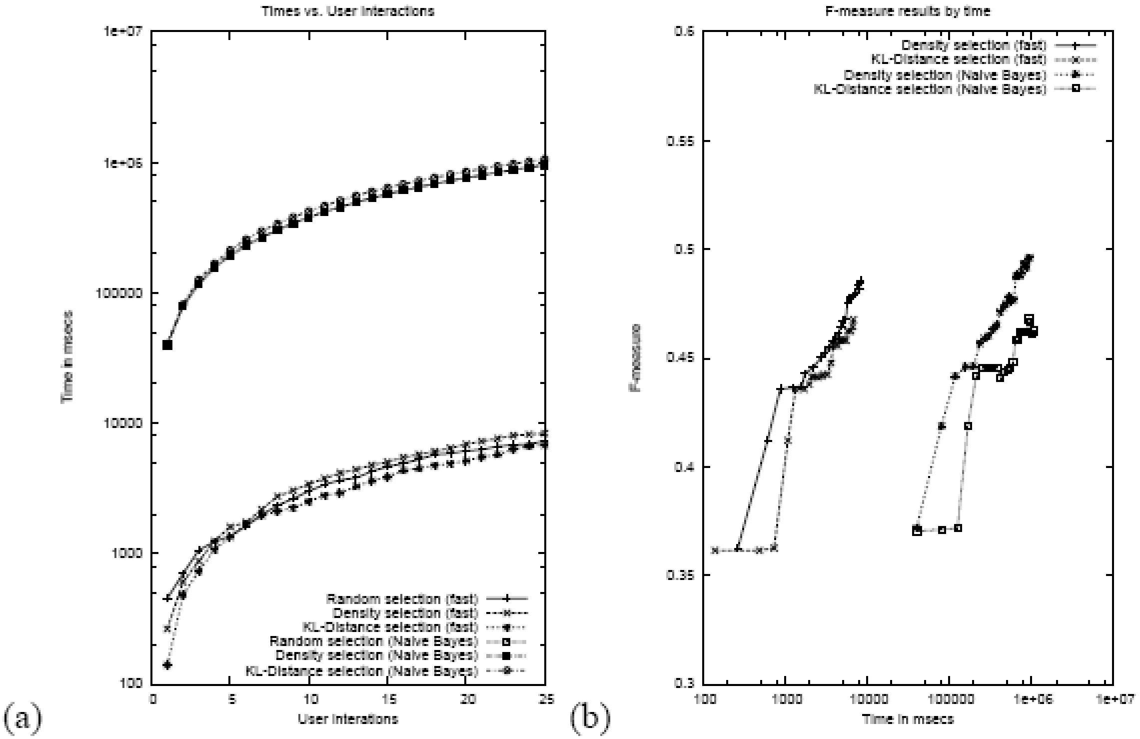
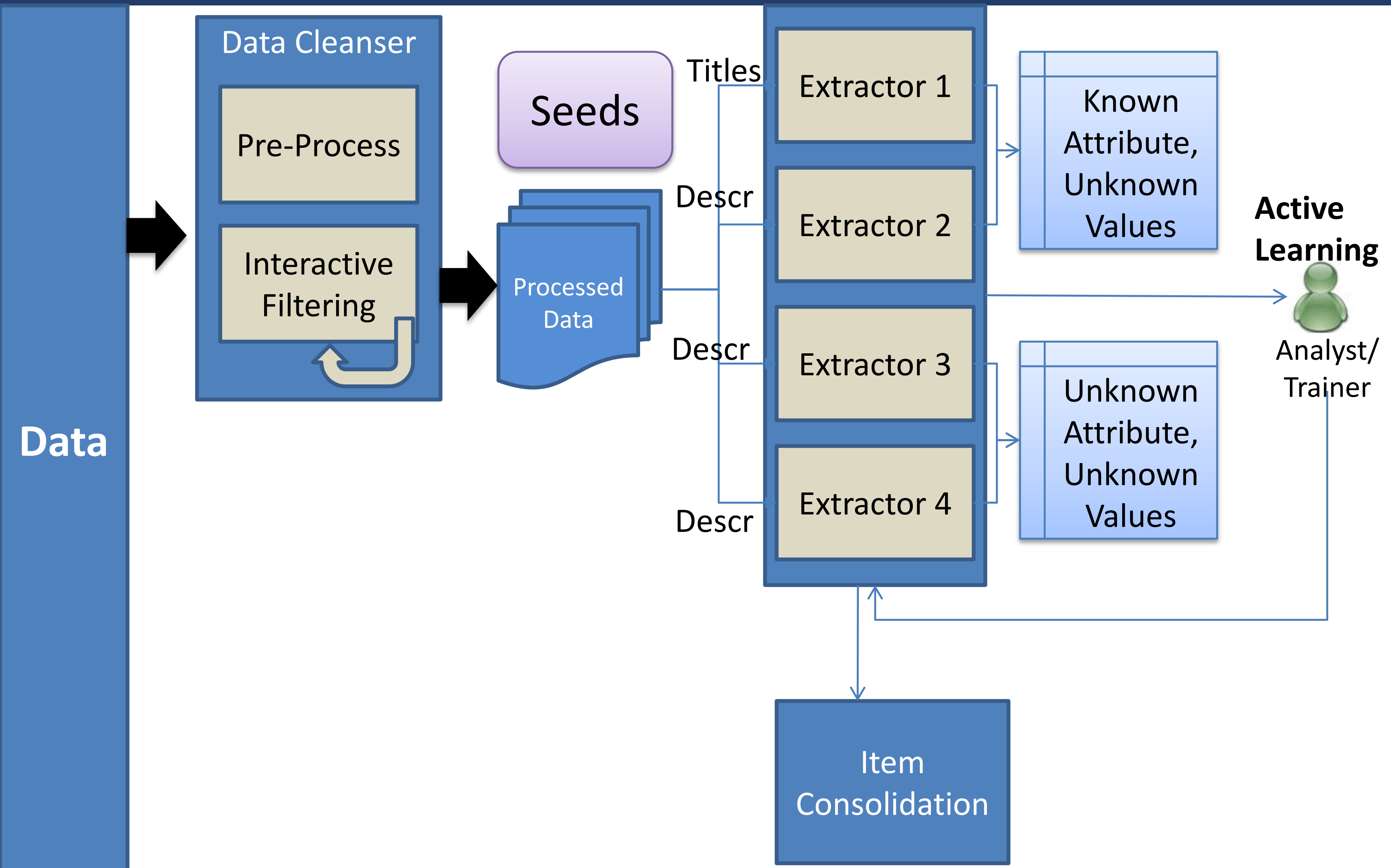


Fig. 2. Time comparison between Fast algorithm and coEM with naive Bayes. In (a), the y-value for k indicates the time the user needed to wait until the k^{th} user interaction.


Data
Data Cleanser
Pre-Process
Interactive Filtering
Seeds
Processed Data
Titles
Descr
Descr
Descr
Extractor 1
Extractor 2
Extractor 3
Extractor 4
Known Attribute, Unknown Values
Unknown Attribute, Unknown Values
Active Learning
Analyst/ Trainer
Item Consolidation

Forming attribute-value pairs

- Classification results in single words labeled as attributes, values, or neither, but many attributes are phrases
  - Merge adjacent words with the same label and with high correlation scores (chi^2, yule's Q, m.i.)

- Link attribute phrases and value phrases to form pairs
  - Syntactic dependencies (Minipar)
  - As fallback, also link phrases that have high correlation score or are adjacent
  - Add known attributes that are implicit but present in our KB
  - Assess whether unlinked attributes are "binary"

Examples

l Example	Attribute	Value
1 1/2-inch polycotton blend tape	polycotton blend tape	1 1/2-inch
1 roll underwrap	underwrap	1 roll
1 tape cutter	tape cutter	1
Extended Torsion bar	bar	Torsion
Synthetic leather upper	#material# upper	leather
Metal ghillies	#material# ghillies	Metal
adiWear tough rubber outsole	rubber outsole	adiWear tough
Imported	Imported	#true#
Dual-density padding with Kinetofoam	padding	Dual-density
Contains 2 BIOflex concentric circle magnet	BIOflex concentric circle magnet	2
93% nylon, 7% spandex	#material#	93% nylon 7% spandex
10-second start-up time delay	start-up time delay	10-second

Examples of extracted pairs for system run with co-EM

Interesting Results

- New Domain Attributes
  - Cyl (Cylinders)
  - Logo
  - Hub Bore
  - Lip
  - Warranty
  - Included
  - Donor Vehicle
  - Shown

Interesting Results

- New values discovered for Tire Type:
  - **Suv**
  - **Minivan**
  - **Snow**
  - **trailer**
  - **Offroad**
  - **boat**.

Context-dependent extraction

- Misleading terms
  - 'Other:' , skype:'
- Typos:
  - 'finnish' as an attribute
- POS-dependent
  - 'Bolt' (NP) = attribute (Bolt Pattern)
  - 'Bolt' (VP) = neither (bolts on to)
- 'Style'
  - *4 new 17x8 Morxchn Mustang Cobra R style deep lip wheels*
  - *We only guarantee it to fit the same Make Model Year and Body Style as its donor vehicle*

KL metric is useful in detecting confusable words

- interga
- quntity

Summary

- Businesses are not very good at capturing all the structure in their data
- Machine learning can help make that process faster, “better”, and cheaper
- The same techniques are applicable in the semantic web to help create, maintain , and update ontologies

Summary

- Traditional Enterprise Knowledge Management has interesting research challenges

- Next generation KM systems are likely to be personalized mash-ups with personalized knowledge models

- Opportunity for semi-automated semantic learning approaches