Enterprise Collaboration & Knowledge Management through Machine Learning and Semantic Technologies

Divna Djordjevic

Credits: Rayid Ghani, Chad Cumby, Mohit Kumar
Accenture Technology Labs

Accenture
Consulting & Services company
200,000 people in over 50 countries

R&D Groups in 4 Locations

Applied Research
Motivated by real business problems impacting most large organizations in the world
Focus Areas include:
- Machine Learning, Data Mining
- Software Engineering
- Collaboration & Knowledge Management
- Cloud/Distributed Computing
- Green Computing
- Biometrics
**ACTIVE**

**Increasing the productivity of knowledge workers**
- Pro-active
- Contextualized
- Easy and unobtrusive

**Accenture goals**
- Motivate, influence, and validate the research in ACTIVE
- Improve collaboration among knowledge workers for Accenture and its clients

**Technologies**
- Collaboration
- Web 2.0
- Machine Learning
- Semantic Technologies

**How?**
- Tacit and unshared knowledge
  - Transferable
  - Interoperable
  - Actionable
  - Enable tasks
  - Collaboration
  - Problem solving
Accenture ACTIVE Use Cases

Privacy Preserving Data Mining for Enterprise Data
Enterprise Tasks: 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

Non-task specific Knowledge Inquiry

Proposal Writing

Personal networks

Internal knowledge repository
Accenture Use Cases

1: Process and Context Sensitive Search & Browsing
- Targeted at validating the scalability of the components developed in ACTIVE
- Wide coverage of tasks and users, shallow support for any one task
- **Accenture Goal:** Improve the productivity of knowledge workers in a variety of tasks and business contexts

2: Collaborative Proposal Development Support
- Small-scale validation with teams focused on Proposal Development
- Deep support for a specific task
- **Accenture Goal:** Improve proposals eventually leading to more proposals being successful
Requirements Gathering Process

Requirements
- Prototype reviews & demonstration sessions
- Open ended interviews

- CIO Organization
- Updated Requirements and Use Cases
- Consultants (users of the system)
- Collaboration Technology & Business Leads
- Dedicated Proposal support teams

Accenture Technology Labs
Requirements Gathering Process

- Interviews & Surveys with 51 people
- Work Shadowed 10 People

- Interviews: 6
- Work Shadow: 10
- Surveys: 27

<table>
<thead>
<tr>
<th>Team Type</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dedicated Teams</td>
<td>6</td>
</tr>
<tr>
<td>Consultants</td>
<td>6</td>
</tr>
<tr>
<td>Collab. Tech. &amp; Business Leads</td>
<td>8</td>
</tr>
<tr>
<td>Internal IT Organization</td>
<td>3</td>
</tr>
</tbody>
</table>
## Requirements

- **Sample of requirements from deliverable:**
  - 10 requirements for use case 1
  - 25 requirements to use case 2

### Use Case 1
- Modeling user contexts
- Discovering new contexts
- Allowing users to specify contexts
- Modeling user tasks
- Discovering new tasks
- Allowing users to collectively tag a discovered set of actions as a task
- Specialization of the information presented to the user based on context and/or task
- Provide access to individual sections of documents
- Provide access to individual graphics categorized by type from documents
- Provide access to experts

### Use Case 2
- Allow creation of new Proposals by soliciting key information such as potential team members, proposal sections, section assignment to team members.
- Create Proposal outline.

- Create new Collaboration Proposal Workspace
- Create documents for individual team members with assigned sections
- Enable update of the section status for each section allocated to a team member

- Individual Workspace for Proposal Development
- Support for Office Applications (Word, Excel, PowerPoint, OneNote, Outlook, Communicator)
- Find similar proposals that have been developed earlier
- Provide graphics from documents that are relevant to the current context of the user
- Provide list of experts and possible collaborators that are relevant to the current context of the user
What needs to be done to validate ACTIVE technologies?

End-users → Validation

Prototypes

Content

Recommend

(Inferred or specified) Context

Knowledge Articulation & Sharing

Context Mining & Modeling

(Inferred or specified) Process

Process Mining & Modeling

Data Collection

Requirements Gathering and Refinement

Requirements Gathering and Refinement
Data Collection Activities

- ACTIVated the Accenture-developed SABLE application to log and monitor user actions
  - 700,000 records of user activity logs.
- Skills database with 170,000 Accenture employees
- 300,000 documents, document repository and accompanying metadata

- Augmented logs from SABLE with logs from Accenture official enterprise search engine
- 147,000 users, 134,000 documents and 7.2 million actions over 2 years
Prototype Development

Use case 1 Activated SABLE

- Content Analysis Components & Data Collection
- **Expert Search**
- Graphics Classification & Search
- Event Logging & external data collection
- **Context** Modeling, Context Mining, Context Discovery, Context Switching, Context Visualization

Use case 2 Collaborative Proposal Development Support

- Proposal Development Toolkit
  - Define the outline, select team members and create the Work space, Find content and experts to develop proposals
- **Collaborative Proposal Development Workspace** using SMW
Content Analysis Components

• Information needs for a Business task
  – *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
  – *Similar content* representations
  – ...
Expert Search: Entity Retrieval & Ranking

Query

“Large Interactive Displays”

Search

Docs

Extract Entities

Relevant Entities

Kelly Craig Jim ...

Compare and find closest matches

Ranked List of Relevant Entities

Kelly

Target Entity 1

Search

Docs

Jim Craig Kelly ...

Accenture Technology Labs
Expert Search: Online experiments

• Expert Search as an extension to enterprise search

• IM bot to find and interact with experts
Approach

- Image extraction from Office Documents
- Supervised learning structural, visual, and text features of the image
- Index image using words and image category
Overview of Approach

1. **Document Repository**
   - Graphics Extraction

2. **Feature Extraction**

3. **Graphics Consolidation**

4. **Reusable Graphics Repository**

5. **Create Training Data**
   - Feature Extraction
   - Graphics Taxonomy

6. **Training samples**
   - Classifier 1
   - Classifier 2
   - Classifier 3

7. **End-user Applications**

   - Active Learning / Relevance Feedback
   - User
Activated SABLE prototype:

• Context Modeling
  - Represented context as a vector of user attributes and content attributes

• Context Mining
  - Usage logs for 147,000 users, 134,000 documents, 7.2 million actions over 2 years
  - Cluster the events to create Accenture-wide contexts

• Context Discovery
  - Based on current user profile and search terms, discover the current context (and similar contexts)

• User-selected context
  - Allow user to select nearby contexts manually

• Contextual Information Delivery
  - Based on discovered (or selected) context, re-rank search results

LiveNetLife:

• Accenture developed "content preparatory" components

Super Contexts = Cluster All contexts over time

Context = User Role + Document Attributes
Activated SABLE
Prototype Development

Use case 1 Activated SABLE

• Content Analysis Components & Data Collection
  • Expert Search
  • Graphics Classification & Search
  • Event Logging & external data collection
• Context Modeling, Context Mining, Context Discovery, Context Switching, Context Visualization

Use case 2 Collaborative Proposal Development Support

• Proposal Development Toolkit
  • Define the outline, select team members and create the Work space, Find content and experts to develop proposals
• Collaborative Proposal Development Workspace using SMW
Proposal Development Process

Detailed Task Description

- Identify client and offering
- Identify Proposal Team
- Develop Win Themes
- Identify RFP sections
- Assign Sections to Team Members
- Create Proposal Workspace
- Distribute Document to Team Members
- Assist in finding content for each section
- Check for consistency, compliance, Win Theme, Integration
- Track Changes between versions
- Consolidate Sections periodically
- Update Workspace Status
- Final Consolidation & Review

Currently Implemented system for the "Develop Proposal" Task

Start point RFP/RFI
Deal Lead receives RFP/RFI
RFP/RFI with workspace creation
MS Word add-in
MS Word add-in that individuals use to help develop the proposal
Proposal Development Workspace (based on SMW)
Proposals, Credentials, Offerings, People

- Activities of a Deal Lead using Word-addons to get RFP sections and select team members
- Proposal Development Workspace
- Activates of a the proposal team members using Word (search, experts, contextual search, etc.) + Proposal Development Workspace
Prototype Development

- Back-end functionality
- Front-end functionality
- Functionality in year 1 prototype
- Functionality in year 2 prototype
- Functionality in fully functional prototype

Process-centric and Context Sensitive Search & Browsing

MS Word Front end

Accenture KX Repository
Workforce Monitoring Search & access activity

Accenture SABLE Usage Monitoring and Data Collection

Search & Click DB
Document Repository

MyCV Skills DB
TNT DB

Context Discovery

Context Identification

LiveNetLife

Expert Search

Graphics Search

Proposal Development Workspace

Proposal, Credentials Offerings, People

RFP

Taxonomy

ASM

Process-centric and Context Sensitive Search & Browsing
Use case 2: Collaborative Proposal Development

• Proposal Development Toolkit
  – Integrated Context Discovery & Visualization
  – Team Selection
  – Configuration & Creation of a new proposal development workspace (based on SMW)
  – Integrated Document, Expert & Graphics Search
  – Auto-Suggest based on content

• Proposal Development Workspace (SMW-based)
  – Customized for proposal development with templates and forms
  – Dynamic Data population from live Accenture sources
  – Content Annotation of uploaded documents in SMW
  – Import capabilities from SMW to MS Word for facts, sections, offerings, etc.
Accenture Proposal for SOA for Client X

Our Understanding of Client Needs

Consultancy Agreement

Services

Pricing

Risk Analysis

Reference cases
Collaborative Proposal Workspace

Accenture Technology Labs
What needs to be done to validate ACTIVE technologies?

Requirements Gathering and Refinement

Prototypes

Content

Known Context

Known Process

Knowledge Articulation & Sharing

Context Mining & Modeling

Process Mining & Modeling

Data Collection

Recommend

Validation

End-users
Current work

– Process Mining, Discovery & Visualization in SABLE
– Improved Process Editing functionality in SMW
– Consistency Checking with SMW
– Privacy Preserving Data Mining
Accenture ACTIVE Use Cases

Privacy Preserving Data Mining for Enterprise Data
Privacy preserving for protecting Enterprise Data

• Goals:
  – Allow data to be used/shared while following enterprise data privacy guidelines

• Formulation, Algorithms & Evaluation
  – Independent
  – Task-based

• Software & Integration with case study
Formulation

• Detection of concept as a classification problem

• Approach
  – Optimally perturb a document to maximize classification error of sensitive concept and minimize that of the utility concept

Maximize \( P(\text{sensitive class} \mid d) - P(\text{sensitive class} \mid \text{perturbed d}) \)

AND

Minimize \( P(\text{utility class} \mid d) - P(\text{utility class} \mid \text{perturbed d}) \)
Representation & Sanitization Operators

• Representation is dependent on kind of data
  – Bag of words (for now)

• Operators
  – Suppression
  – Generalization
  – Noise addition
Algorithms

- **K-confusability**

- **Batch Redaction**
  - Based on Feature Selection approaches in classification (Log-odds, Odds-ratio, Information Gain)

- **Document level redaction**
  - Linear program balancing the log-likelihood of the sensitive class against the log-likelihood of the utility class
Results: Guessing Game

• Goal: guess the newsgroup a message was posted in
• 50 people
• ~1100 documents
Human Game Results

Figure 3: Error rates human subject 20 Newsgroup experiment.

![Graph showing error rates for different values of k.

Table 2: Top 10 words to obscure alt.atheism document when $k = 2$ and $k = 5$.]

<table>
<thead>
<tr>
<th>k=2</th>
<th>k=5</th>
</tr>
</thead>
<tbody>
<tr>
<td>religion</td>
<td>faith</td>
</tr>
<tr>
<td>dogma</td>
<td>christian</td>
</tr>
<tr>
<td>system</td>
<td>dogma</td>
</tr>
<tr>
<td>encourages</td>
<td>system</td>
</tr>
<tr>
<td>toronto</td>
<td>encourages</td>
</tr>
<tr>
<td>humans</td>
<td>beliefs</td>
</tr>
<tr>
<td>beliefs</td>
<td>humans</td>
</tr>
<tr>
<td>genocide</td>
<td>secular</td>
</tr>
<tr>
<td>philosopher</td>
<td>philosopher</td>
</tr>
<tr>
<td>prison</td>
<td>christianity</td>
</tr>
</tbody>
</table>
In article <__@___.com> __@___.com (___ ___) writes:
> ... Of course, they are working on the ___ that ___
> overbloom with penetration into ___ membrane ___ with
> associated "___" ___ response can and does ___
> in a large ___ of people. If you reject this "___"
> ___", then I'd guess you'd view this ___ as one
> more wasteful and quixotic endeavor. Stay tuned.

I do not have enough ___ expertise to have much of an opinion
one way or another on hidden ___ _. I can
understand the ___ of those who see this associated with
various general ___ of ___, while there is a lack of ___
demonstration that this happens and causes such general ___.
(To understand this ___, one only needs to know of past
failures that ___ these characteristics with the notion of
hidden ___ ___. There have been quite a few, and the
___ of all thought that the ___ were overly skeptical.)

On the other hand, I am happy to read that some people are
sufficiently ___ in this possibility, spurred by
suggestive _____, to ___ it further. The
doubters may be ___. (It has happened before.)

I realize that ___ ignorance in the face of ignorance may
not endear me to those who are so sure they know one way or
another. (And, indeed, perhaps some of them do know -- I am the
one who is currently ignorant.) But I find this the most honest
___, and so I am happy with it.
Integration with Accenture Case Study

SABLE ENGINE AND DATABASE

BatchReducer.exe

ContextBuilder.exe

Persons Documents Events STORE

AccentureContexter

Persons Documents INDEX

Contexts

Accenture2SP.exe

Events

New Events

SOAP

Read

Metadata & Text

Metadata & Text

Bottom Up Contexts
Initial results for privacy preserving context mining

Error for different Reduction Amounts

Evaluation of the discovered contexts on anonymized Accenture data sets using the Context Evaluation Measure.
Integration with the Proposal Development Toolkit

- Automatic Document Anonymization Module (ADAM)
- Suggests words/phrases that need to be removed to obfuscate sensitive information
  - Customized currently for Clients
Summary

• Algorithms: Machine Learning framework for protecting sensitive information in text documents

• Experiments: Public data + Accenture Data covering both human and automated attack scenarios

• Software: Document Anonymization Module + MS Word add-in
Accenture ACTIVE Use Cases

Privacy Preserving Data Mining for Enterprise Data

Applied Example: Process–Specific and Context Sensitive Knowledge Work
Health Insurance Claim Process

• Inefficiencies in the healthcare process result in large monetary losses affecting corporation and consumers
  – $186 billion over-spent in US every year (McKinsey study)
  – 131 percent increase in insurance premiums over past 10 years

• Claim payment errors drive a significant portion of these inefficiencies
  – Increased administrative costs and service issues of health plans
  – Overpayment of Claims - direct loss
  – Underpayment of Claims – loss in interest payment for insurer, loss in revenue for provider
Random Audits for Quality Control

Extremely Low Hit Rates
Long audit times due to fully manual audits
Estimating performance on Unlabeled Data

- Mix test data with Unlabelled data
- Rank the entire set and measure the recall of the labeled test set
- 40% of known rework claims are found in top 10 percentile data

*Mohit Kumar, Rayid Ghani, and Zhu-Song Mei (KDD-2010).
Live evaluation (with auditors)

• Insurance Company 1
  – Audit time: 20 min – 1 hour
  – Gave a sample of 100 claims to auditor
    • Precision: 65%

• Pilot deployment (4 weeks) : Insurance Company 2
  – Audit time: 4 min – 10 min
  – Total claims audited: 1000
    • Precision: 29-55% based on audit strategy
System User/Audit Interface

©Accenture 2009. All Rights Reserved.
Interactive Classification Setting

- Classifier trained from labeled data
- Human (user/expert) in the loop using the results but also providing feedback at a cost

- Goal: Maximize the productivity of the human
Observations from pilot deployment

- “Context switching” costs were high
- Lack of intuitive explanations

(Expert/Labeling) Cost Models
- Variable cost but as a function of previous labels or examples

Goal: Maximize exploitation while minimizing cost
More Like This strategy

Intuitions
- Raw classifier ranking has high context switching cost but reasonable accuracy
- Post-classification clustering improves context switching cost but hurts accuracy
- Online version (more-like-this) improves both

*R. Ghani and M. Kumar, ICML 2010, Workshop Budgeted Learning
Live System Deployment

- Number of claims audited:
  - Baseline system: 200
  - More-Like-This: 307
- 90% relative improvement over baseline
- 27% reduction in audit time over baseline

~$10 Million savings /year for a typical insurance company
Summary

• Enable Knowledge workers (experts) to be more efficient and effective in performing task
• Allow the system to adapt over time
• Cost and benefit of any information that passes between expert and data mining system
  – Exploit the context of the user for the process at hand