SAND2017-10775C

Text Analytics in Building Personalized Information Retrieve Environment

Text Analytics Forum, 2017

Pengchu Zhang
John Herzer
Jessica Shaffer-Gant
Sandia National Laboratories

HJA1
HJA2

Motivations:

- Deliver Information to Employees based on:
  - Who they are:
    - *Current search engines retrieve information with built-in algorithms solely based on the query;*
    - *In an organization, members of the work force have different needs for information even though they use the same/similar queries, e.g.;*
      - Managers/Leaders want to know the progress of on-going projects
      - Developers search for current/historical technologies
      - Financial analysts want to know the health of the budget
  - What they may need:
    - *Users' information needs depend on what they are working on in an organization, not only on their backgrounds and job titles*
      - Employees with job title "computer science" may work in various areas:
        - » Web development; Modeling and simulation; Software development…

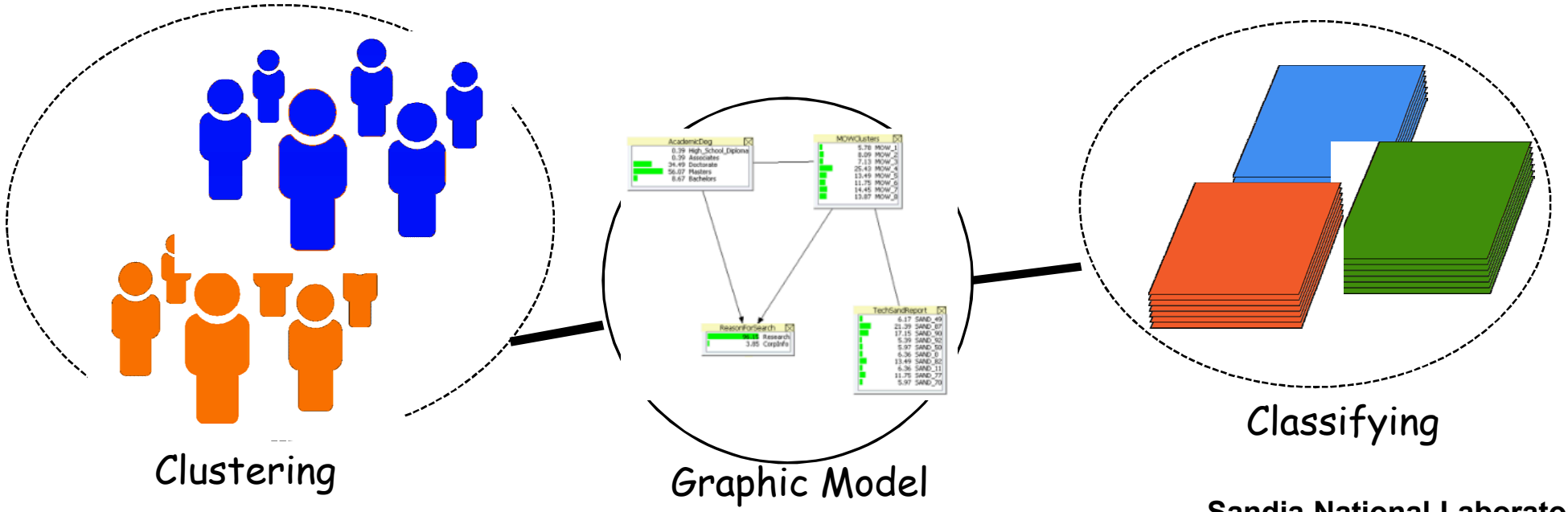
Slide 2

HJA1 Herzer, John A, 10/2/2017

HJA2 Who are
Herzer, John A, 10/2/2017

Our Solution: SPIRE

- The Sandia Personalized Information Retrieval Environment
- Private users with most relevant content based on who they are
- To accomplish this we need to cluster/classify customers, documents and build the probabilistic based predictive model



Profile MOWs with Personal Attributes

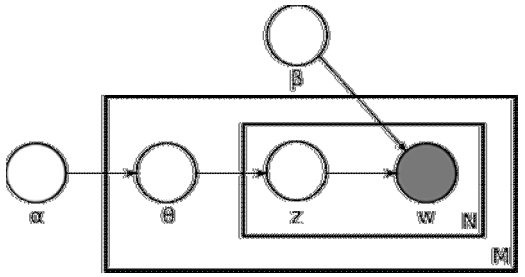
- Physical Attributes (easy to get but not very useful):
  - Education, years at Sandia, job title…
  - Example: Sandia has about 1000 MOWs in job title of "Computer Science & Dev" but they are working in very different areas
- Documents generated by MOWs (very useful):
  - Job assignments over the years;
  - Publications
  - Proposals, abstracts, presentations…

Unsupervised learning

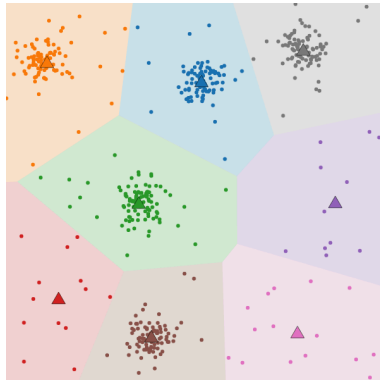
Clustering Users Based on Their Similarities

1. Job description
2. Publications, internal documents…

Users' data



LDA



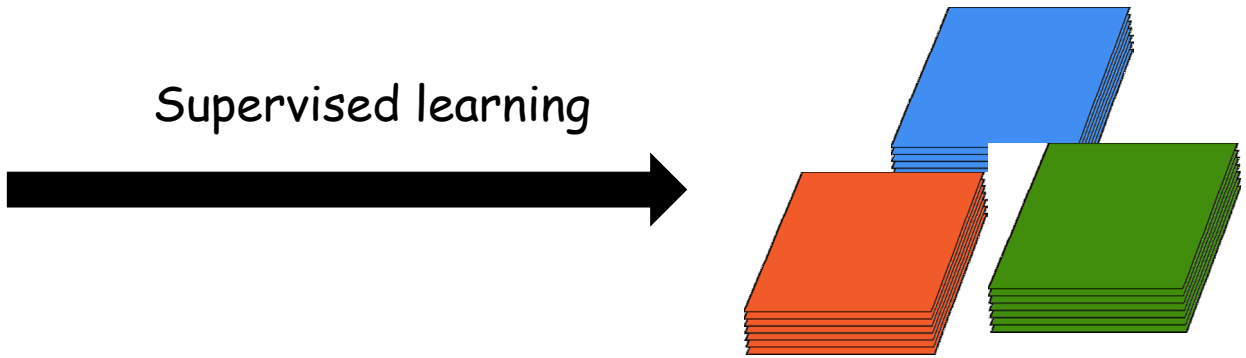
Kmeans

Clustering Algorithms

User Clusters

Document Classification

- There have been many models for classification:
  - Binary models such as sentiment classification: negative or positive
  - Statistical based:
    - *Need a lot of manually selected features as inputs, very expensive*
    - *Hard to scale up*



Build SAND Report Classifiers with Deep Learning

1. Classify documents into proper classes
2. Recognize the document class in various formats
3. Recognize the distribution of possible classes in a document

New document | Trained Classifiers | Distribution of Topics in a Document

Procedures

- Collected ~100,000 SAND reports over last 50 years
- Data cleaned and indexed with Apache/Lucene
- Built "Taxonomy" for Sandia Category Guide (SCG)
- Selected the highly ranked documents with SCG taxonomy terms/phrases as the training sets
- Tokenized the terms in documents with a 200 dimension numerical vector
- Built a Convolutional Neural Network (CNN)
  - 3 one dimension, 5 step convolutional layers with 128 feature maps
  - Three layers of maxpooling
  - ReLu and Softmax as the activation functions
  - Loss function = categorical_crossentropy
- Trained the network with various hyperparameters

Build "Taxonomy" based on Sandia Category Guide (SCG)

Category (Material Sciences)

Subcategories:

- Ceramics
- plastics
- seals and Adhesives

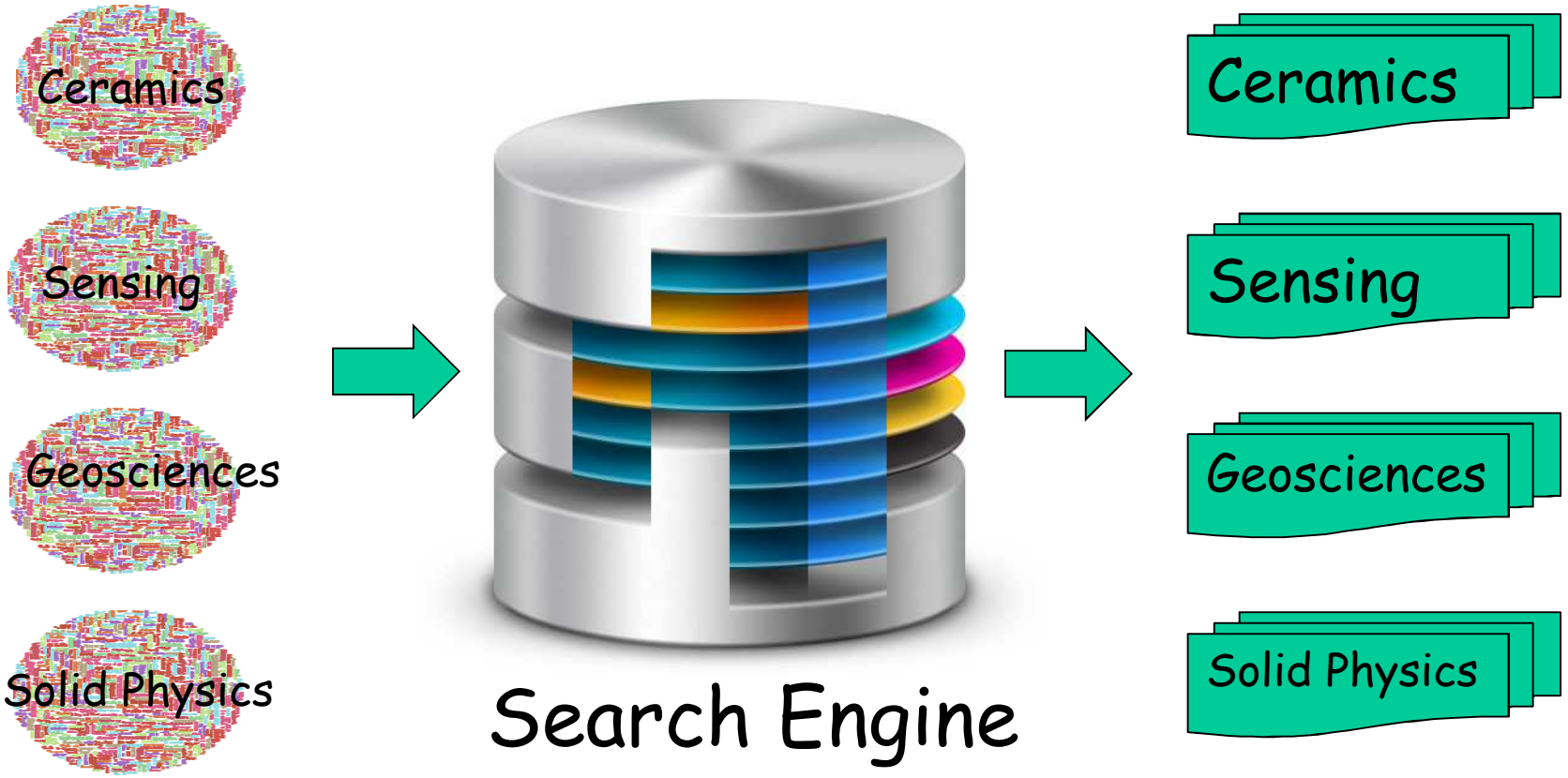
Obtain the terms/phrases for subcategories from:

- Documents created by Sandia authors
- Wikipedia
- Word2Vec built on Sandia's documents
- Internal Organization webpages

Taxonomy Example for "Material Sciences/Composite Materials:

"carbon composite" "carbon fiber" "ceramal" "ceramic composite"
"ceramic matrix composite" "CFRP" "Chobham armour" "clad metals" "CMC"
"composite material" "concrete-plastic composite" "fiber reinforced composite"
"fiber reinforced polymer" "fiberglass" "formica" "FRP" "glasdpolyester"
"graphite" "GRP" "laminate" "Mallite" "metal composite" "metal matrix composite"…

Collect Sets of Labeled Documents for Training

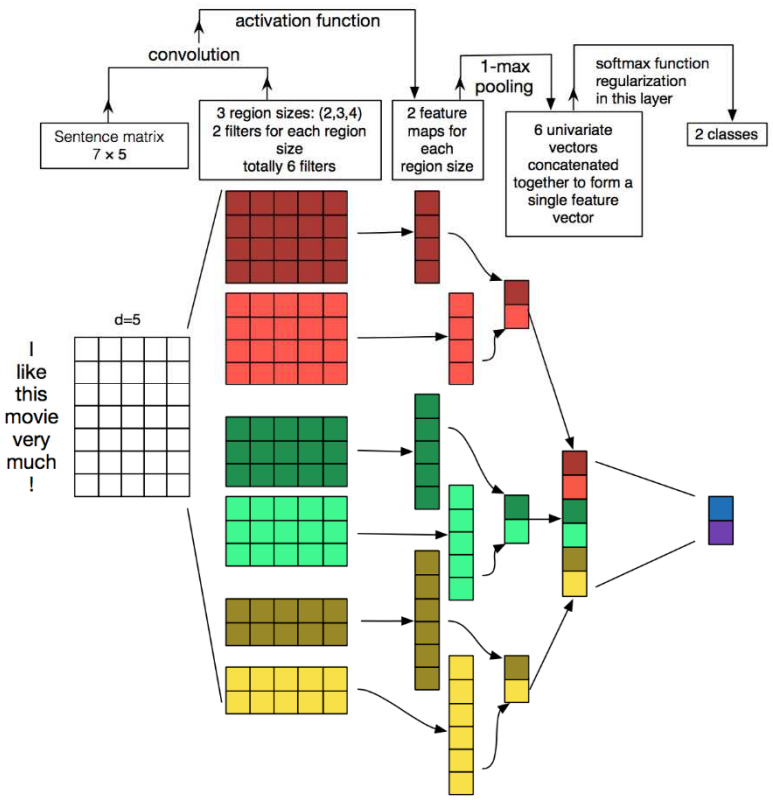


Terms/Phrases from Taxonomy

Labeled and Ranked Documents

Convolutional Neural Network for Text Classification

http://www.wildml.com/2015/11/understanding-convolutional-neural-networks-for-nlp/




```
print('Training model.')

# train a 1D convnet with global maxpooling
sequence_input = Input(shape=(MAX_SEQUENCE_LENGTH,),
dtype='int32')
embedded_sequences = embedding_layer(sequence_input)
x = Conv1D(128, 5, activation='relu')(embedded_sequences)
x = MaxPooling1D(5)(x)
x = Dropout(0.5)(x)
x = Conv1D(128, 5, activation='relu')(x)
x = Dropout(0.5)(x)
x = MaxPooling1D(5)(x)
x = Conv1D(128, 5, activation='relu')(x)
x = Dropout(0.5)(x)
x = MaxPooling1D(35)(x)
x = Flatten()(x)
x = Dense(128, activation='relu')(x)
preds = Dense(len(labels_index), activation='softmax')(x)
model = Model(sequence_input, preds)
model.compile(loss='categorical_crossentropy',optimizer='rmsprop',
metrics=['acc'])
```

Modified from Keras example

Examples of Classifiers for Classification

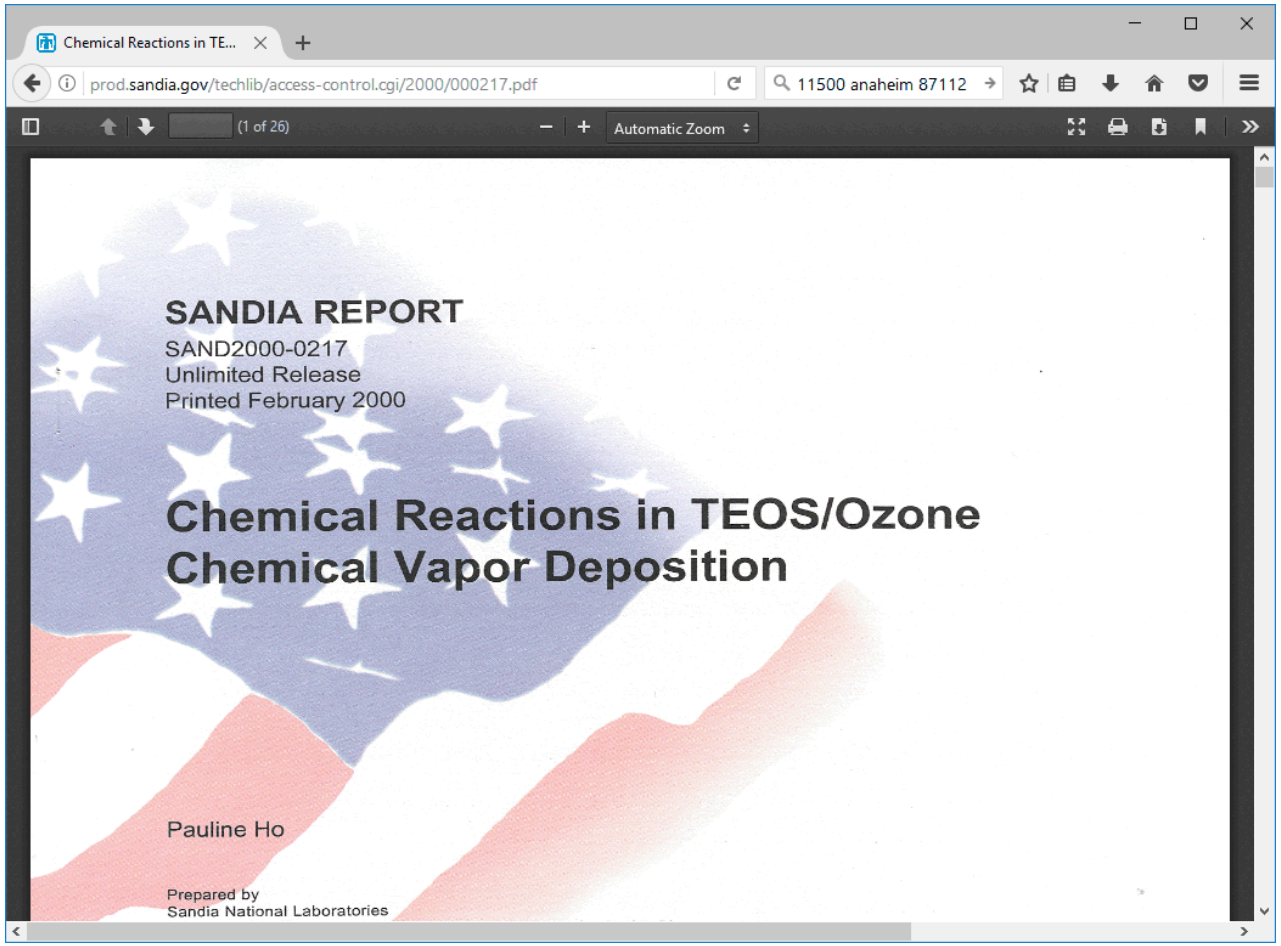
```
prediction = model.predict(data[146:150])
K=2
for p in range(0, prediction.shape[0]):

    a=np.array(prediction[p])
    b=np.argpartition(a, -K)[-K:]
    np.set_printoptions(precision=3)
    print(b, np.take(a, b)*100,'%', '\t', titles[p+146], '\t',labels_name[b[0]],'\t', labels_name[b[1]])
```

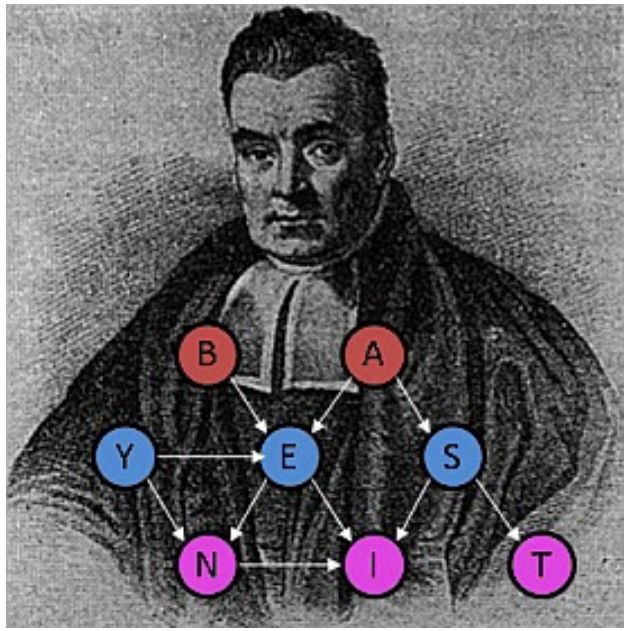
[37 1] [6.892 79.014] % SAND2000-0217.txt thermodynamics atmospheric sciences
[30 14] [6.755 91.043] % SAND2000-0218.txt particle physics electronics and electrical engineering
[16 31] [6.668 60.088] % SAND2000-0221.txt fluid mechanics plasma physics
[9 29] [29.797 52.055] % SAND2000-0222C.txt computer architecture optics

Example of Classifying a Document

[37 1] [6.892 79.014] % SAND2000-0217.txt thermodynamics atmospheric sciences



Building Graphic Model for Prediction



Prior probability

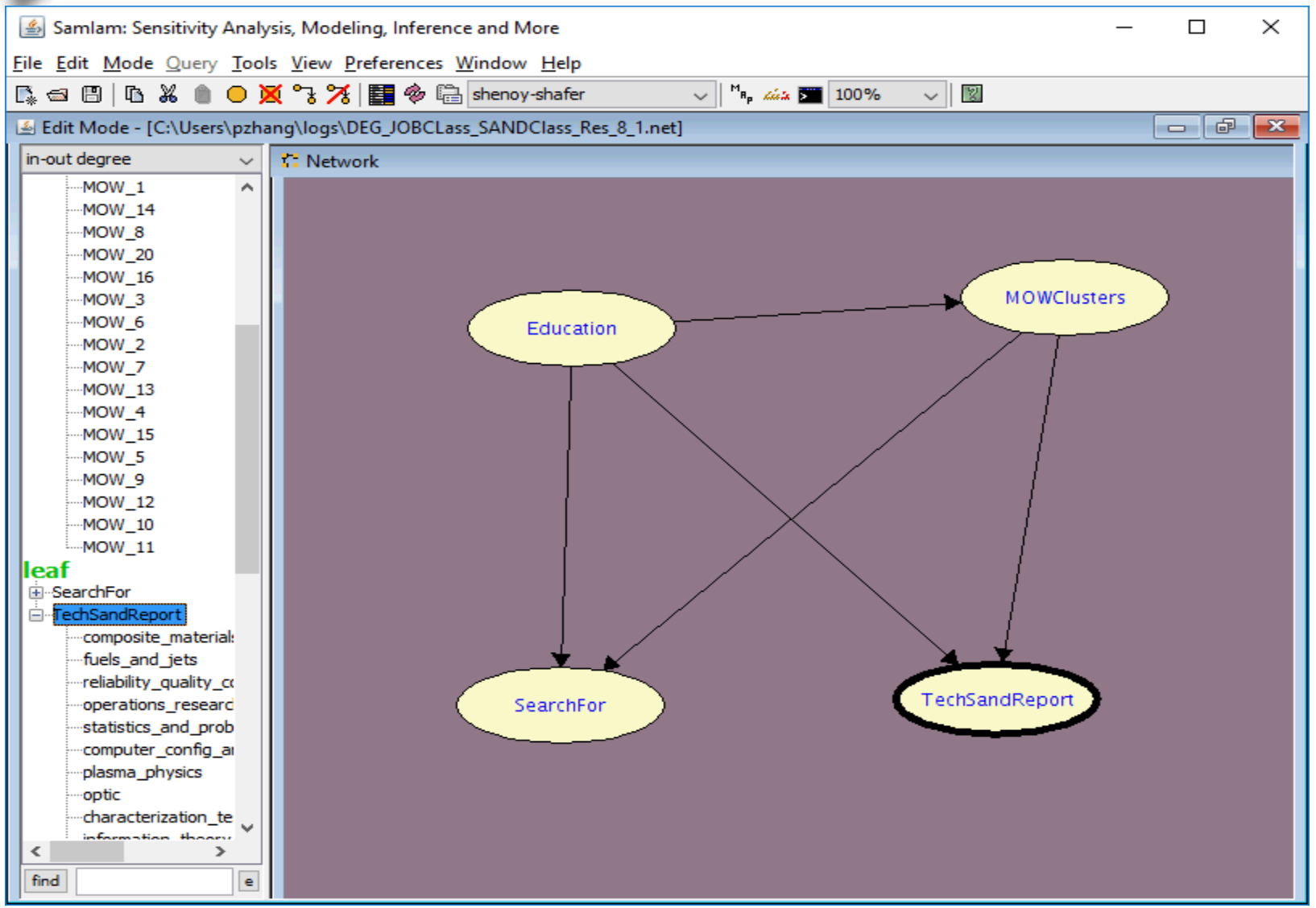
Likelihood of the evidence of 'E'
If the hypothesis 'x' is true

$$P(x|E) = \frac{P(x) * P(E|x)}{P(E)}$$

Posterior probability of 'x'
given the evidence 'E'

Prior probability that
the evidence itself is true

Graphic Model

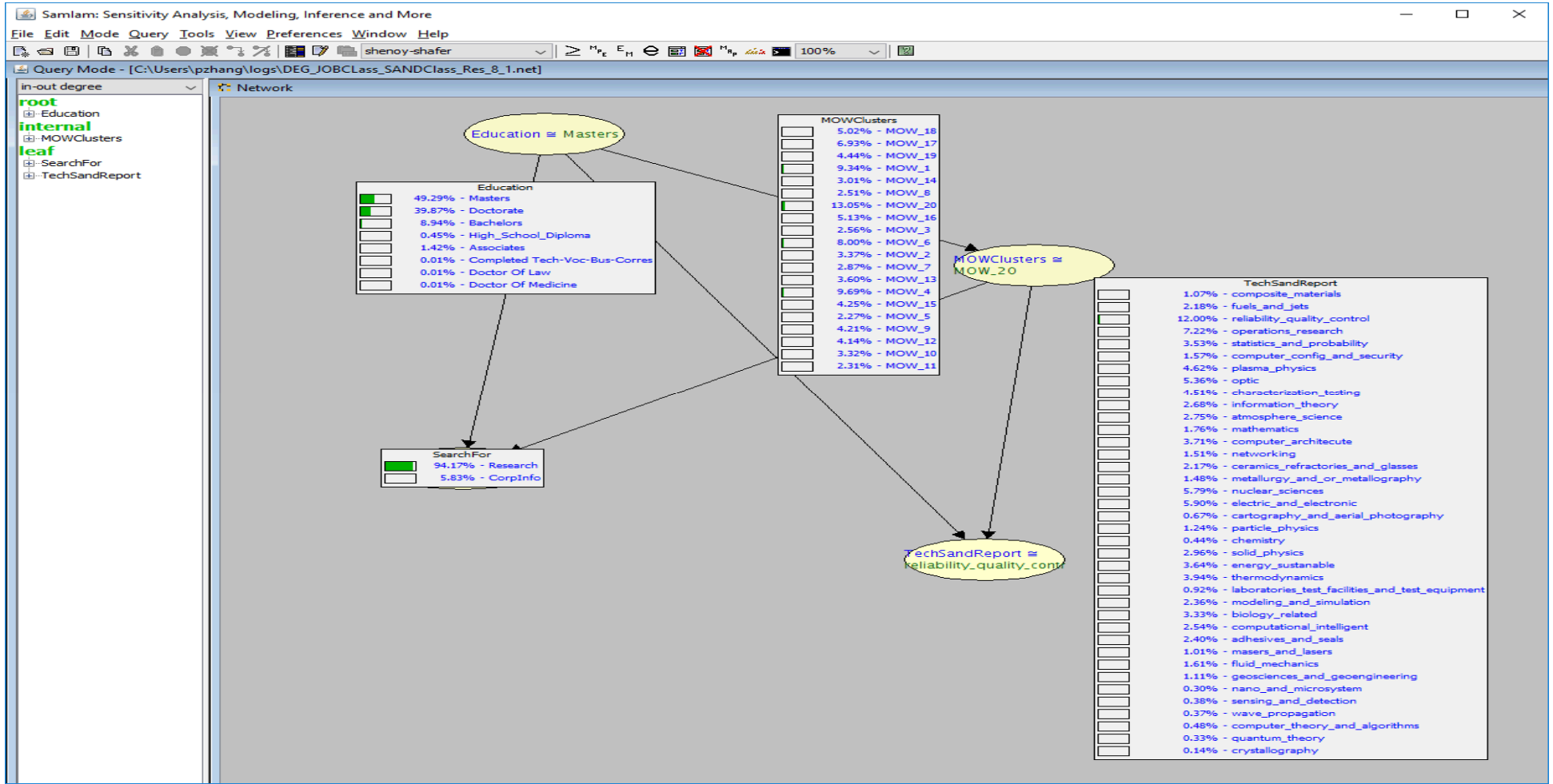


Parameters

- Clusters of Employees
- Academic Degrees
- Purpose for Search
- Classes of Tech Reports

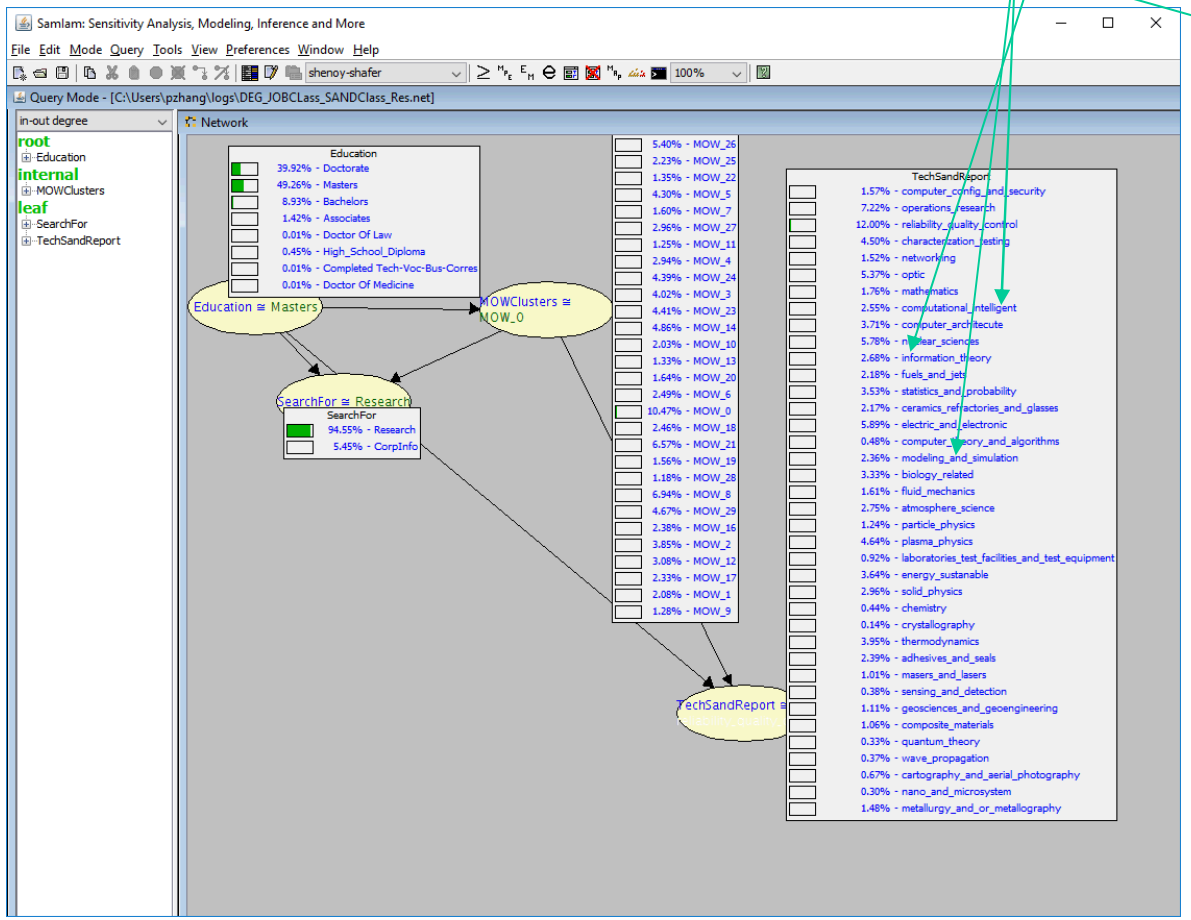
Conditional Probabilities of Parameters

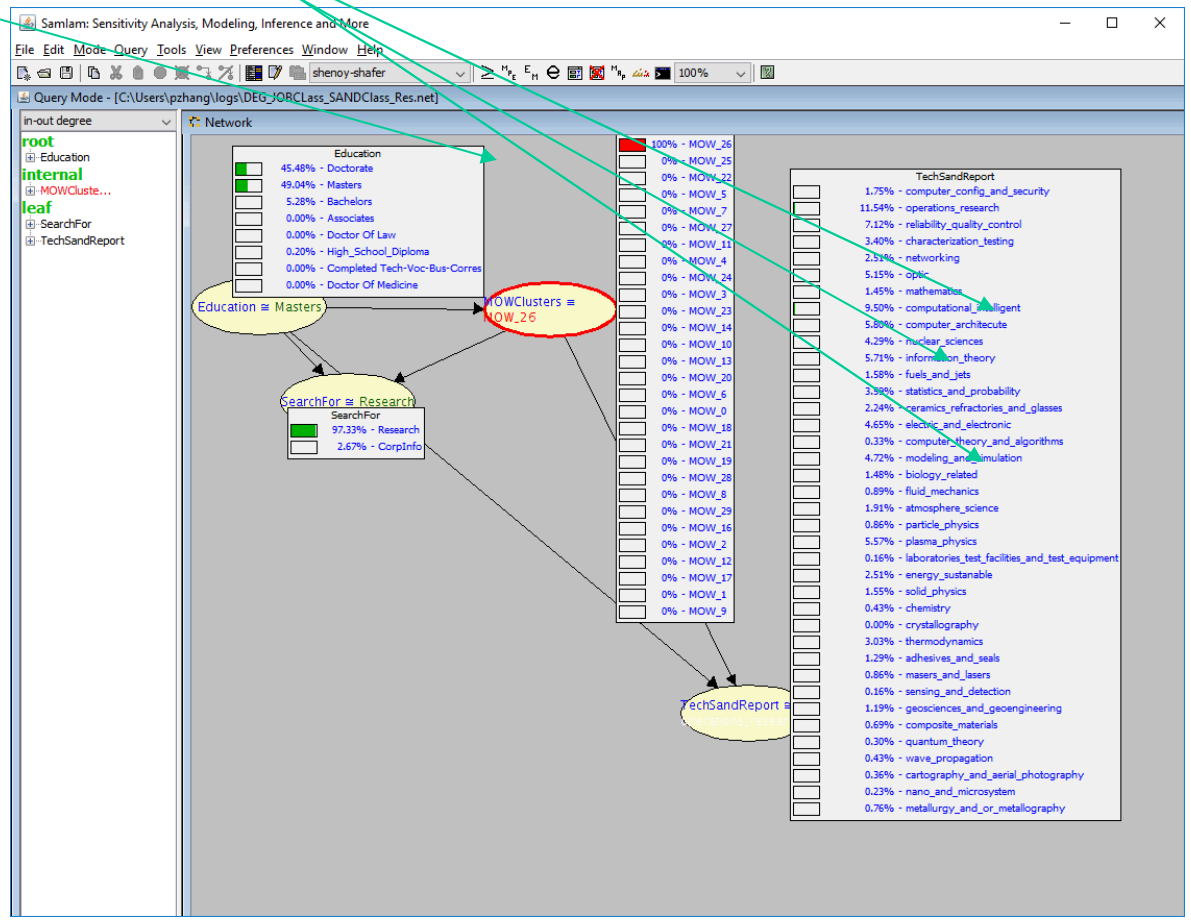
1. The behaviors of users on Solr-based Enterprise Search Engine
2. Attributes of Clusters of Users
3. Build the conditional probability tables and create the Bayesian Network



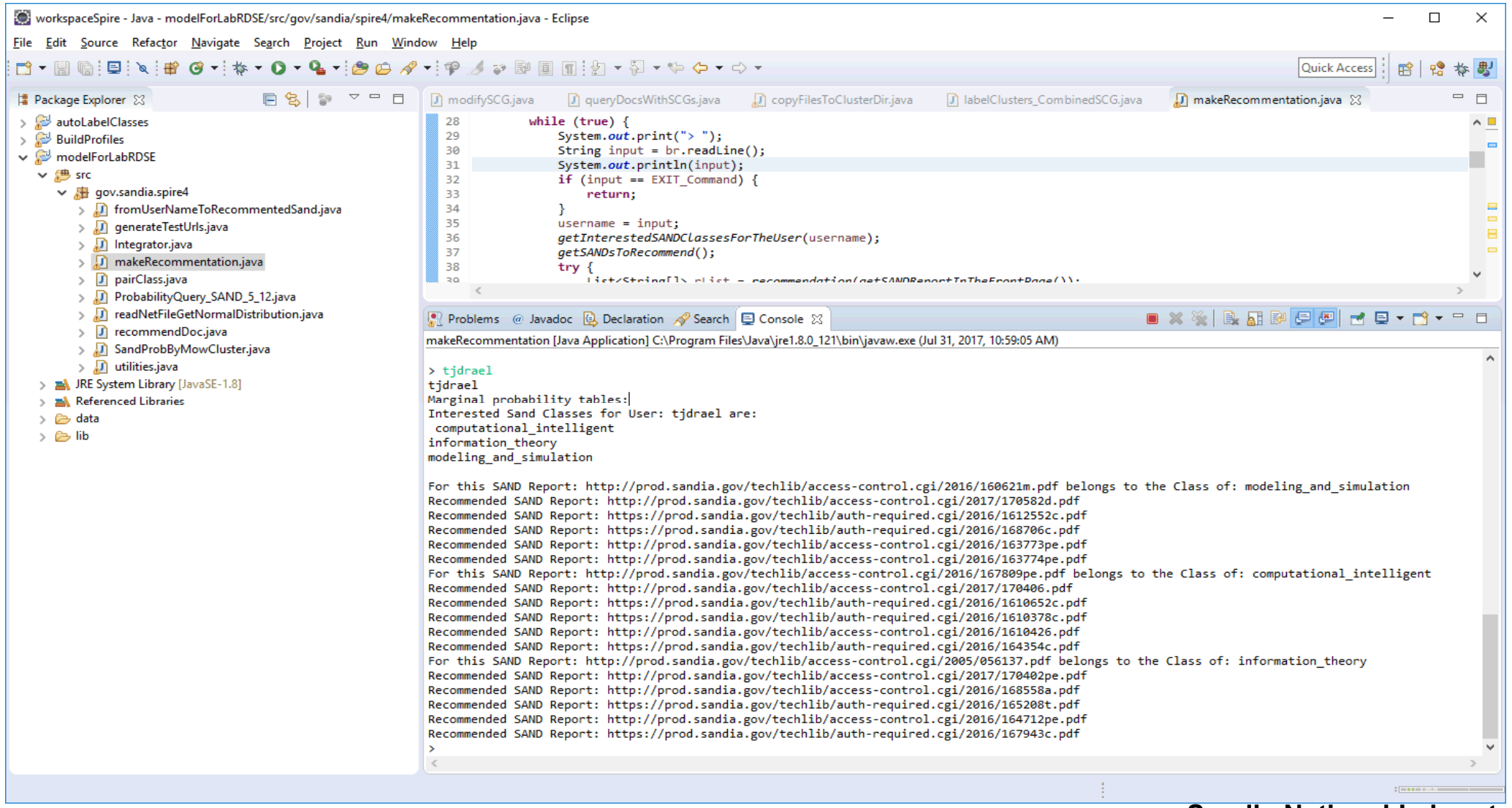
Predict Information Needed for a Given MOW

Computational Intelligent from 2.55 to 9.5%
Modeling and Simulation from 2.36 to 4.72%
Information Theory from 2.69 to 5.71%





Example of Recommendation based on the Predictive Model



USE Case

User: xxxxx

Interested Classes by the user: Energy Sustainable, Material Sciences, Information Theory

Query: battery safety testing

Search Engine Returns:

Developing Battery Safety and Abuse Testing for Stationary Battery Applications (in Energy Sustainable Class)
https://prod.sandia.gov/techlib/auth-required.cgi/2016/168008c.pdf

SPIRE model recommendations (in Energy Sustainable Class)

Developing an Energy Storage Project – A Technical Perspective
https://prod.sandia.gov/techlib/auth-required.cgi/2017/170203c.pdf

Battery Safety Testing: 2016 Energy Storage Annual Merit Review
https://prod.sandia.gov/techlib/auth-required.cgi/2017/170932c.pdf

Energy Storage – The Future
https://prod.sandia.gov/techlib/auth-required.cgi/2017/170289c.pdf

Monocarba-closo-polyborate electrolytes for all solid state Li-ion batteries
https://prod.sandia.gov/techlib/auth-required.cgi/2017/170810c.pdf

In situ characterization of charge rate dependent stress and structure changes in V2O5 cathode prepared by atomic layer deposition
https://prod.sandia.gov/techlib/auth-required.cgi/2017/170811j.pdf

Achievements and Further Efforts

- Text analytics provides the foundation to build up an intelligent information retrieval environment:
  - Understanding the similarities of employees through clustering
  - Collecting labeled documents for Machine Learning guided by Taxonomy
  - Classifying text documents with trained classifiers
- Probabilistic graphic model associates users and documents:
  - The model reasons and infers information needed by users
  - The model will be used to predict users' information needs
- More work needed to build a better taxonomy to select labeled documents for learning
  - This is always the issue for supervised learnings