SAND2016-3458C

Enhancing search results relevance with machine learning

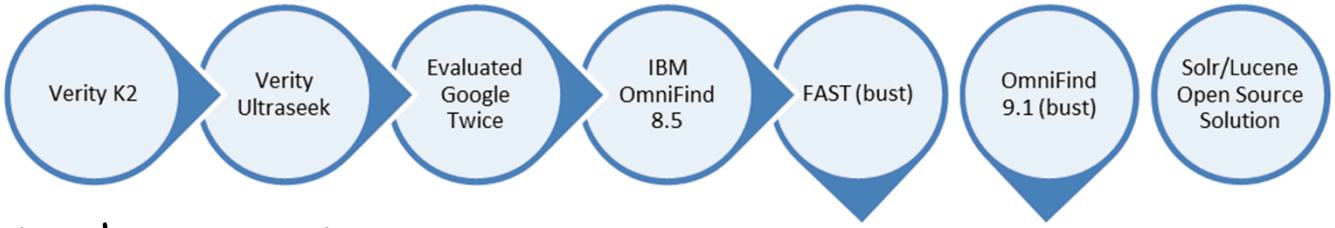
Pengchu Zhang
John Herzer

May 2016

Sandia National Laboratories

The search challenge

Corporate search at Sandia



Recent enhancements

- Migrated to open source search engine
- Developed custom crawlers
- Built federation to a selected set of information sources
- Created an "Applications That Listen" framework for delivering pre-computed answers
- Added a system to track and boost popular links in search

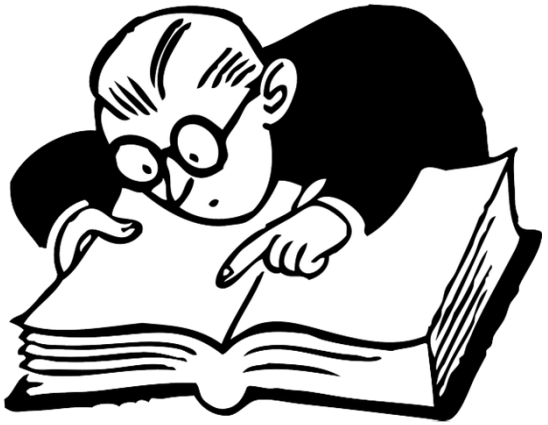
Customers still aren't always delighted with search results

The Holy Grail: Conceptual search

- Go beyond keyword search to actually search on the concept behind the customer's query
- Problems due to the imprecise nature of language
  - Synonymy
    - *Query is for "dog" but desired document contains "canine"*
  - Polysemy
    - *Does query for "lead" refer to "leading a team" or the chemical element "Pb"*
  - Stemming
    - *Searching for "strike" vs "striking" vs "struck"*
- How do we implement conceptual search?

Acquiring a corporate dictionary

- Commercial / open source dictionaries don't have words unique to our organization
- Efforts to build a corporate ontology/taxonomy/dictionary tend to fizzle
- What we need is a way to build the dictionary from our own corpus in an automated way
- Word2Vec is an unsupervised machine learning approach that lets us identify related words from the corpus

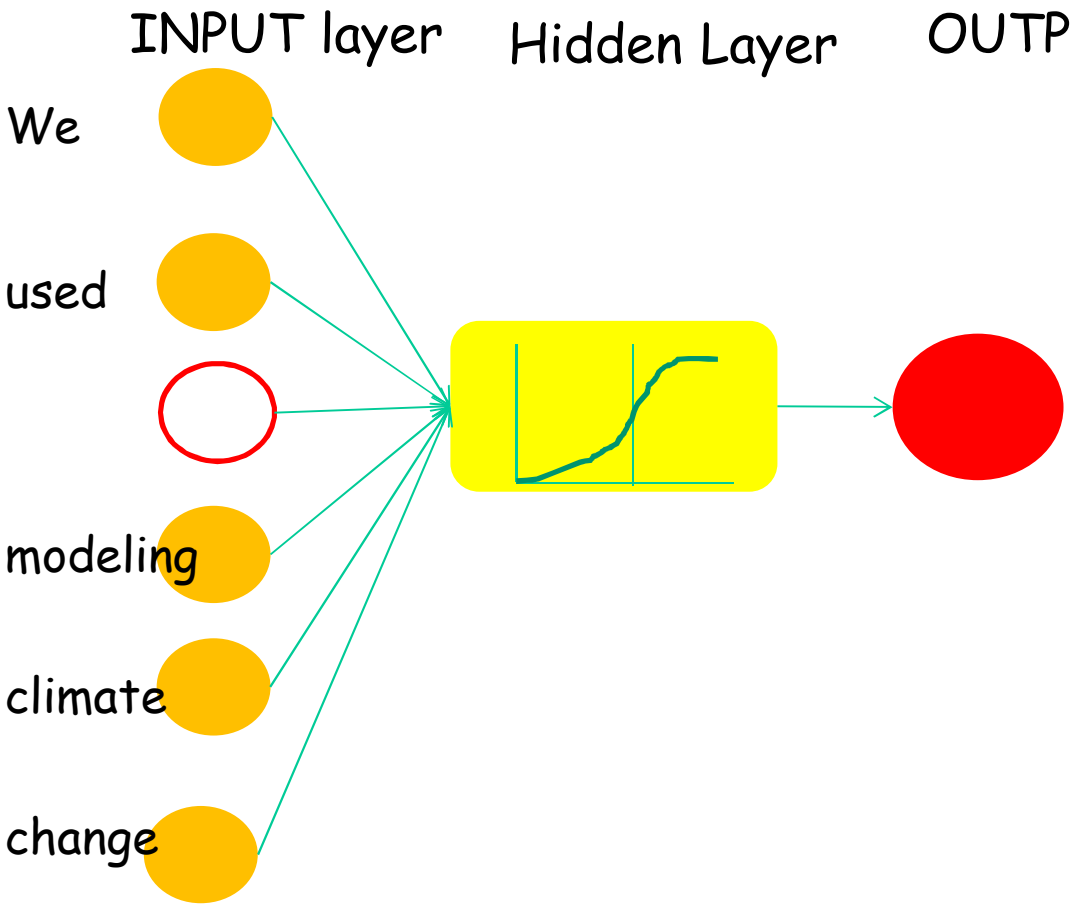


To Improve Search Results with Word2Vec

- Query with a single term or phrase
  - "retirement" Search engine will return documents that contain the term and rank the documents based on the frequency of "retirement";
- Word2Vec expands the query term into a set of RELATED terms or phrases

  "retirement
  Pension
  Savings
  Eligible
  pension_fund
  income_plan
  Pension_plans"
  Search engine will return documents that contain all or some of the terms or phrases and rank the documents based on the frequencies of the set of terms or phrases, the set of terms/phrases represents as "Concept"
- How to expand a set of RELATED terms from a single query term?

Concept of Neural Network Language Model



Term	P
Computer	0.9
Algorithm	0.8
Software	0.8
Brain	0.7
iPhone	0.4
Ghost	0.3
Dog	0.2

Problems of Word Representation in Traditional Language Models

- One-Hot Representations
  - Simple way to encode discrete concepts, such as words
    - *"we used computer modeling climate change"*
      - We = [1 0 0 0 0 0]
      - used = [0 1 0 0 0 0]
      - computer = [0 0 1 0 0 0]
      - modeling = [0 0 0 1 0 0]
      - climate = [0 0 0 0 1 0]
      - change = [0 0 0 0 0 1]
  - A one-hot encoding makes no assumption about word similarity
    - *All words are equally different from each other*
  - This representation is very high in dimensions
    - *The dimensionality is the size of the vocabulary*
    - *A typical vocabulary size is 100,000*

Word2Vec in Natural Language Applications

Collobert, R., *et al.* Natural language processing (almost) from scratch. J. Mach. Learn. Res. 12, 2493–2537 (2011).

Socher, R., Lin, C. C-Y., Manning, C. & Ng, A. Y. Parsing natural scenes and natural language with recursive neural networks. In *Proc. International Conference on Machine Learning* 129–136 (2011).

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. & Dean, J. Distributed representations of words and phrases and their compositionality. In *Proc. Advances in Neural Information Processing Systems 26* 3111–3119 (2013).

Sutskever, I. Vinyals, O. & Le. Q. V. Sequence to sequence learning with neural networks. In *Proc. Advances in Neural Information Processing Systems 27* 3104–3112 (2014).

Cho, K. *et al.* Learning phrase representations using RNN encoder-decoder for statistical machine translation. In *Proc. Conference on Empirical Methods in Natural Language Processing* 1724–1734 (2014).

Bahdanau, D., Cho, K. & Bengio, Y. Neural machine translation by jointly learning to align and translate. In *Proc. International Conference on Learning Representations* http://arxiv.org/abs/1409.0473 (2015).

Lecun, Y, Bengio, Y and Hinton G. Deep Learning. Nature 521, 436-444 (28 May 2015)

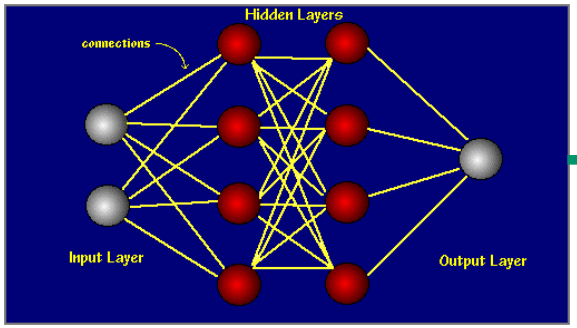
Representing Words as Vectors and Adjusting the Vector based on How Words are Used in Writing and Talking

Word2Vec (Google 2013)

The dog (cat) is walking in the room

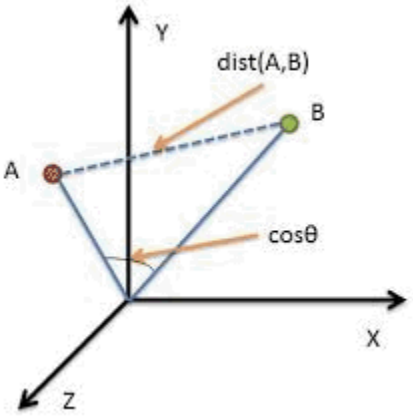
Initial vector

(0.12, 0.23, 0.22)
(0.32, 0.27, 0.94)
(0.18, 0.88, 0.45)
(0.23, 0.92, 0.23)
(0.77, 0.25, 0.11)
(0.12, 0.23, 0.22)
(0.41, 0.13, 0.29)



New vector

(0.12, 0.23, 0.22)
(0.62, 0.99, 0.14)
(0.18, 0.88, 0.45)
(0.23, 0.92, 0.23)
(0.77, 0.25, 0.11)
(0.12, 0.23, 0.22)
(0.41, 0.13, 0.29)

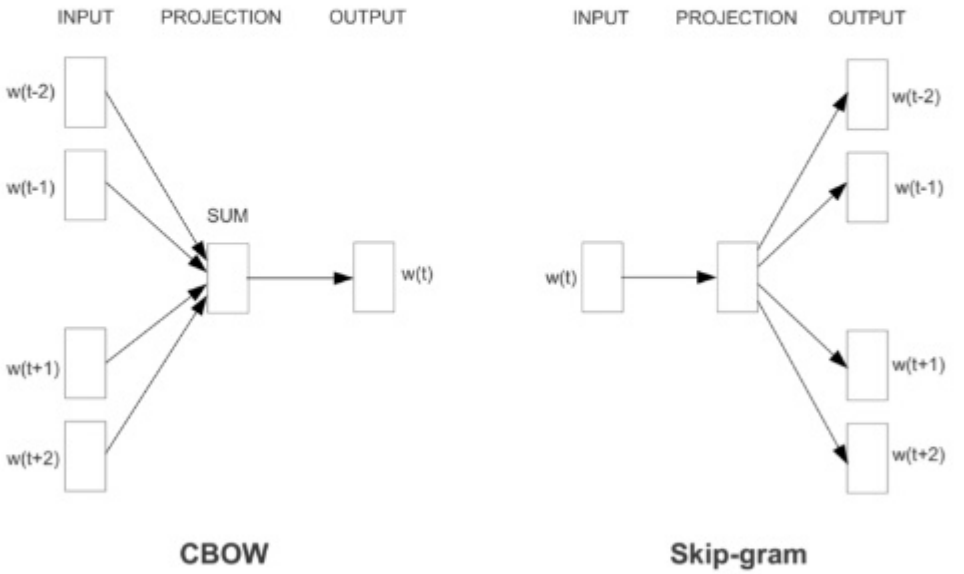


$$\text{similarity} = \cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}}$$

$$d(p,q) = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + \cdots + (p_i - q_i)^2 + \cdots + (p_n - q_n)^2}.$$

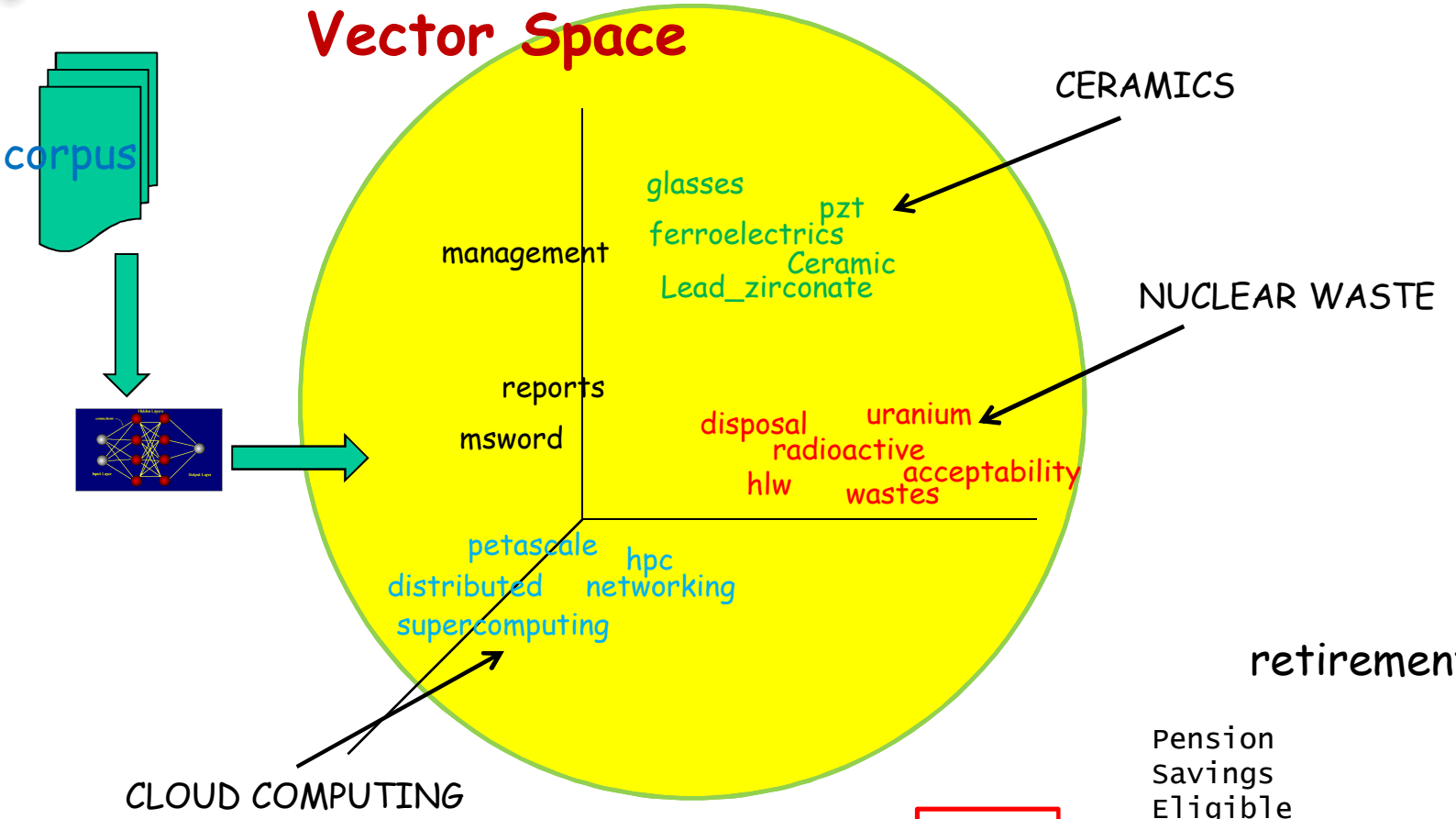
Two Approaches to Build Word2Vec Models

Word2Vec



Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jerey Dean. Distributed Representations of Words and Phrases and their Compositionality. NIPS, 2013.

Conceptual Vector Space distribution of terms by Word2Vec



retirement

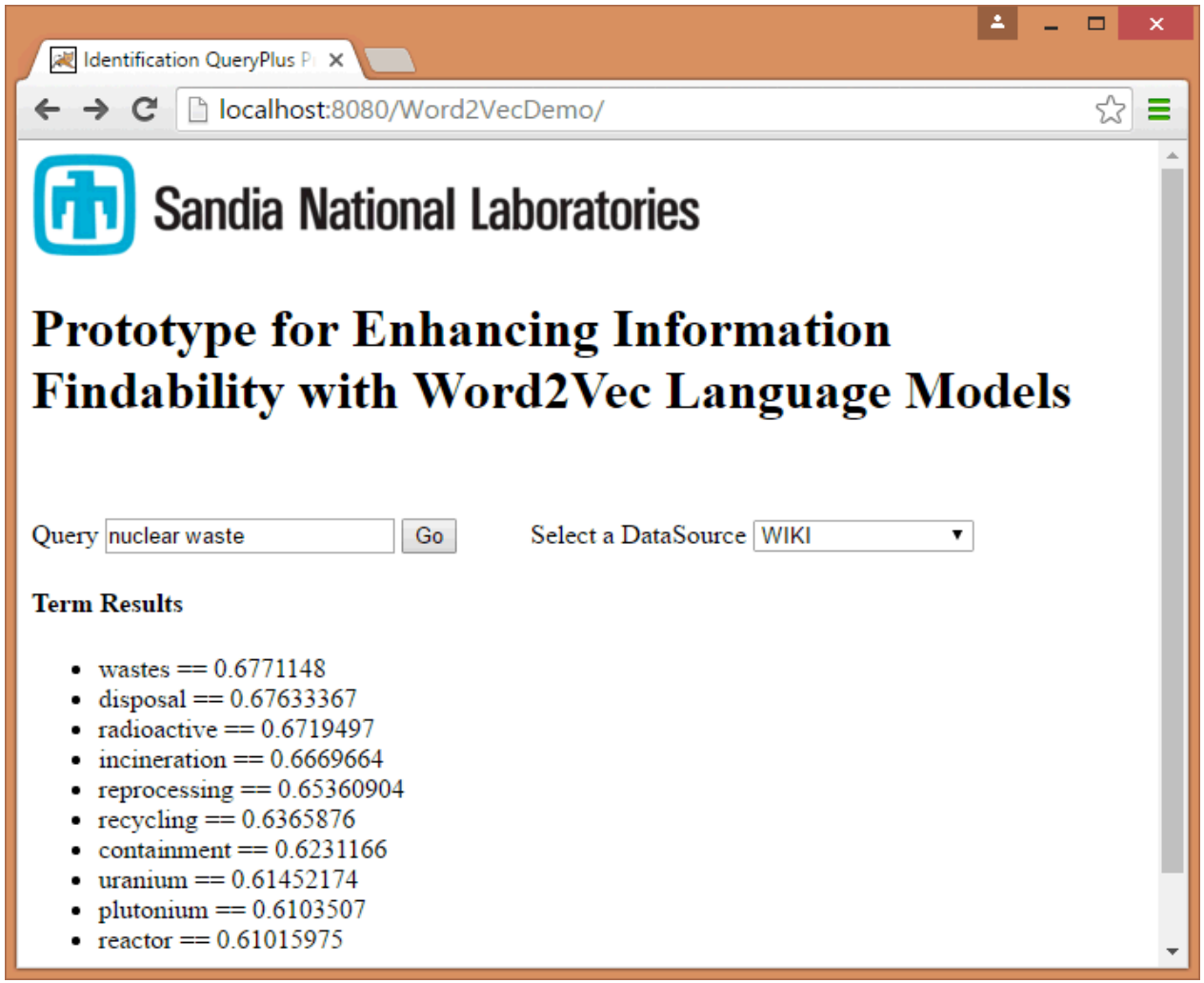
Vector distance

Related terms

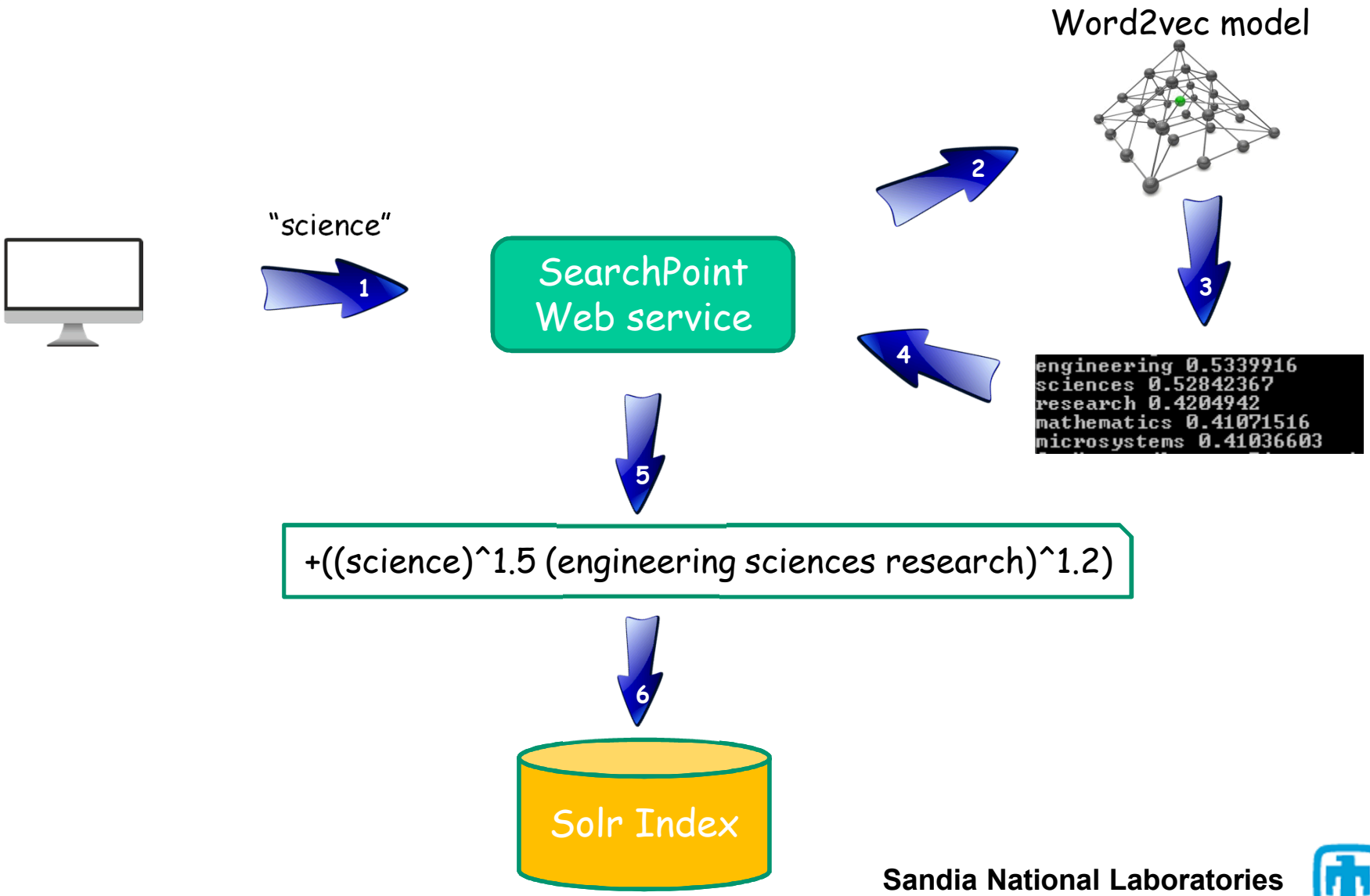
Pension	0.525072
Savings	0.495344
Eligible	0.469864
pension_fund	0.46272
income_plan	0.449314
pension_plans	0.435645

$$d(p,q) = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + \cdots + (p_i - q_i)^2 + \cdots + (p_n - q_n)^2}.$$

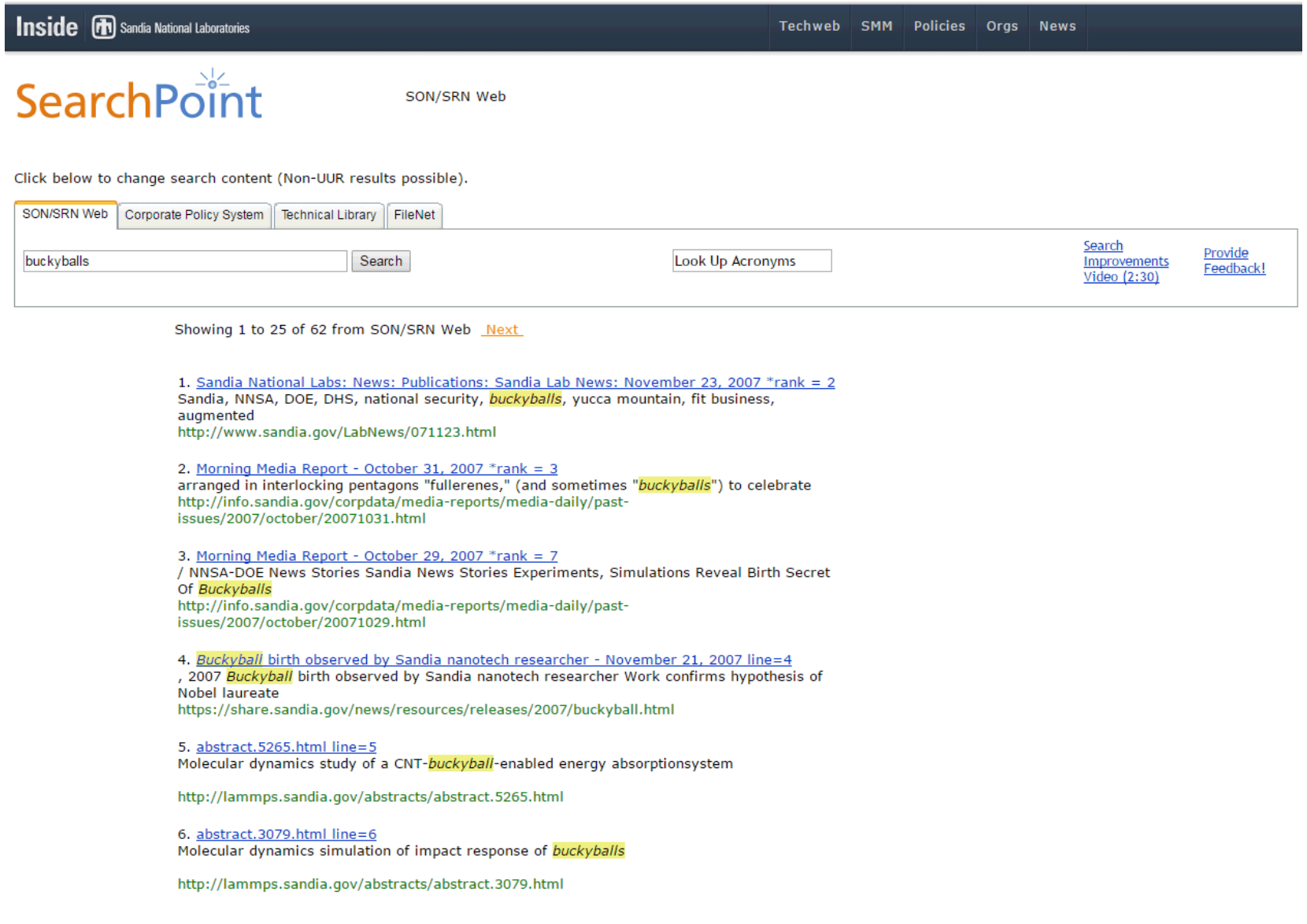
Example of Term Expansion



Enhancing queries with Word2Vec



Results without query expansion



Inside Sandia National Laboratories | Techweb | SMM | Policies | Orgs | News

SearchPoint SON/SRN Web

Click below to change search content (Non-UUR results possible).

SON/SRN Web | Corporate Policy System | Technical Library | FileNet

buckyballs Search Look Up Acronyms Search Improvements Video (2:30) Provide Feedback!

Showing 1 to 25 of 62 from SON/SRN Web Next

1. Sandia National Labs: News: Publications: Sandia Lab News: November 23, 2007 *rank = 2
Sandia, NNSA, DOE, DHS, national security, *buckyballs*, yucca mountain, fit business, augmented
http://www.sandia.gov/LabNews/071123.html

2. Morning Media Report - October 31, 2007 *rank = 3
arranged in interlocking pentagons "fullerenes," (and sometimes "*buckyballs*") to celebrate
http://info.sandia.gov/corpdata/media-reports/media-daily/past-issues/2007/october/20071031.html

3. Morning Media Report - October 29, 2007 *rank = 7
/ NNSA-DOE News Stories Sandia News Stories Experiments, Simulations Reveal Birth Secret Of *Buckyballs*
http://info.sandia.gov/corpdata/media-reports/media-daily/past-issues/2007/october/20071029.html

4. *Buckyball* birth observed by Sandia nanotech researcher - November 21, 2007 line=4
, 2007 *Buckyball* birth observed by Sandia nanotech researcher Work confirms hypothesis of Nobel laureate
https://share.sandia.gov/news/resources/releases/2007/buckyball.html

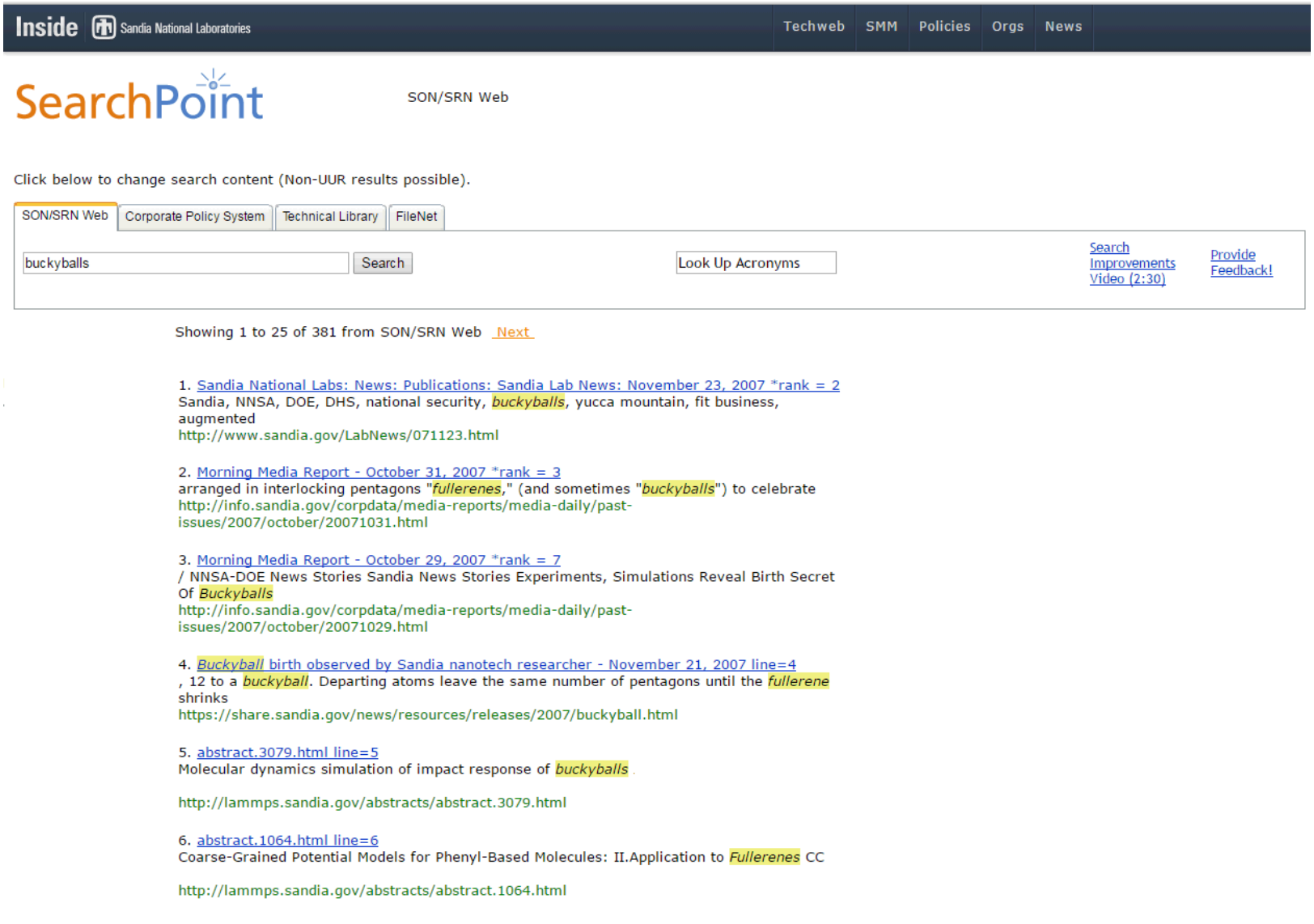
5. abstract.5265.html line=5
Molecular dynamics study of a CNT-*buckyball*-enabled energy absorptionsystem

http://lammps.sandia.gov/abstracts/abstract.5265.html

6. abstract.3079.html line=6
Molecular dynamics simulation of impact response of *buckyballs*

http://lammps.sandia.gov/abstracts/abstract.3079.html

Results with query expansion



Inside Sandia National Laboratories | Techweb | SMM | Policies | Orgs | News

SearchPoint SON/SRN Web

Click below to change search content (Non-UUR results possible).

SON/SRN Web | Corporate Policy System | Technical Library | FileNet

buckyballs Search Look Up Acronyms Search Improvements Video (2:30) Provide Feedback!

Showing 1 to 25 of 381 from SON/SRN Web Next

1. Sandia National Labs: News: Publications: Sandia Lab News: November 23, 2007 *rank = 2
Sandia, NNSA, DOE, DHS, national security, *buckyballs*, yucca mountain, fit business, augmented
http://www.sandia.gov/LabNews/071123.html

2. Morning Media Report - October 31, 2007 *rank = 3
arranged in interlocking pentagons "*fullerenes*," (and sometimes "*buckyballs*") to celebrate
http://info.sandia.gov/corpdata/media-reports/media-daily/past-issues/2007/october/20071031.html

3. Morning Media Report - October 29, 2007 *rank = 7
/ NNSA-DOE News Stories Sandia News Stories Experiments, Simulations Reveal Birth Secret Of *Buckyballs*
http://info.sandia.gov/corpdata/media-reports/media-daily/past-issues/2007/october/20071029.html

4. *Buckyball* birth observed by Sandia nanotech researcher - November 21, 2007 line=4
, 12 to a *buckyball*. Departing atoms leave the same number of pentagons until the *fullerene* shrinks
https://share.sandia.gov/news/resources/releases/2007/buckyball.html

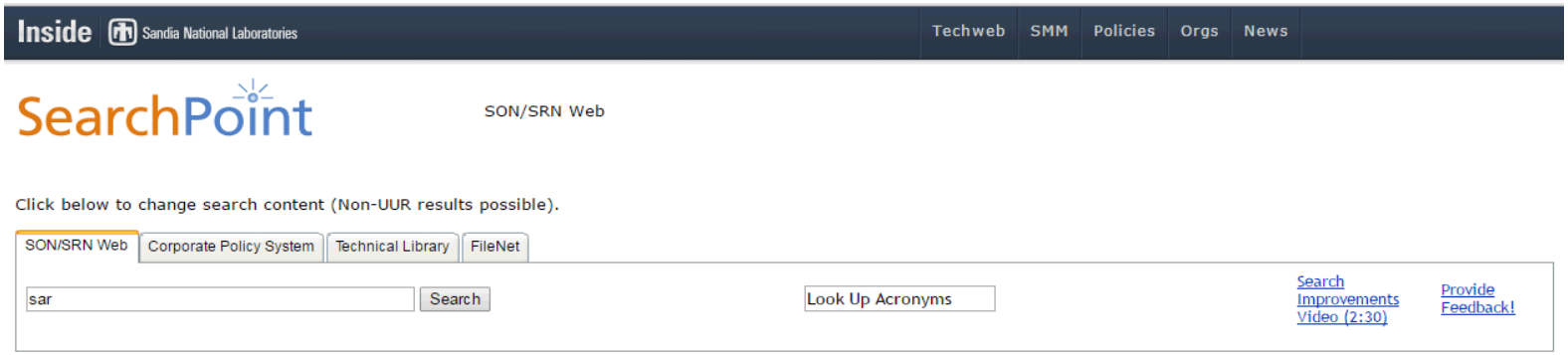
5. abstract.3079.html line=5
Molecular dynamics simulation of impact response of *buckyballs*

http://lammps.sandia.gov/abstracts/abstract.3079.html

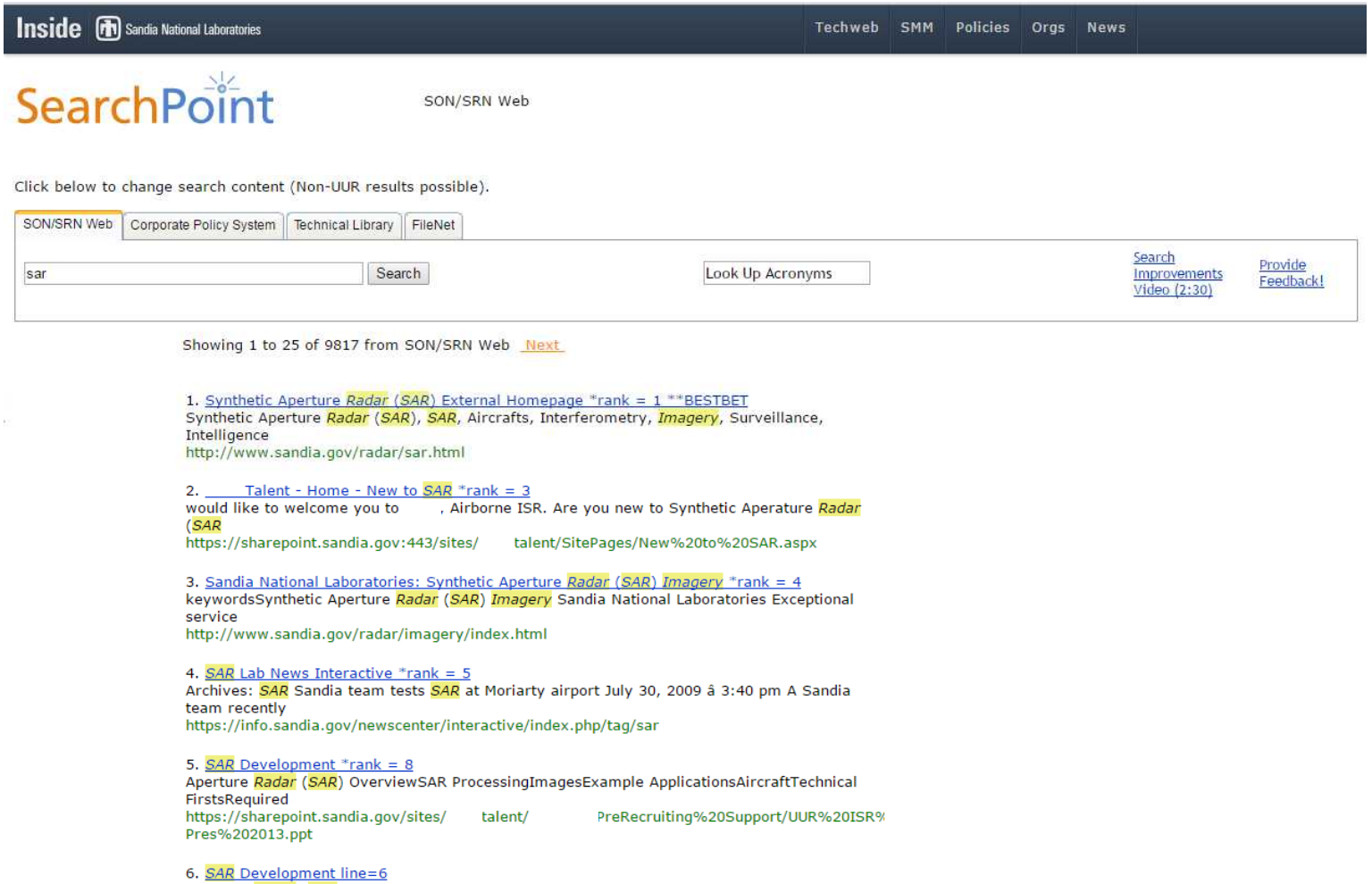
6. abstract.1064.html line=6
Coarse-Grained Potential Models for Phenyl-Based Molecules: II.Application to *Fullerenes* CC

http://lammps.sandia.gov/abstracts/abstract.1064.html

Acronyms without query expansion



Acronyms with query expansion



Performance considerations

- SearchPoint application is J2EE on Weblogic
- File sizes
  - solrTerms.ser 479Mb
  - solrPhrases.ser 825 Mb
  - SearchPointNext.ear 509 Mb
- Vectors loaded during Servlet init
  - Java memory usage: -Xms256m -Xmx2g
  - 69 seconds to load terms
- Time to access terms for query: < 150ms

Future work

- Tackle the query disambiguation problem (polysemy)
  - Predict most likely meaning of query term based on corpus usage
  - Predict personalized meaning from customer's prior usage
- Identify phrases within the query
  - Can eliminate many non-relevant results
  - Given an expertise query "server admin java"

    "server admin" or java will provide better results than server or admin or java