Do The Math

Team

1. Deshpande, Amol (amold1@umbc.edu)
2. Peshave, Akshay (peshave1@umbc.edu)

Data set

We will be using submissions to the ACM and AMS portals and their user-defined index tags/terms defined by a subset of the Computing Classification System(CCS 2012) and Math Subject Classification (MSC 2010) tag hierarchies respectively.

Project idea

We are attempting to implement a system whose:

- input is technical text belonging to the computing science domain .
- output is relevant topics or concepts in the mathematics domain.

Our data corpus is ACM and AMS publications which are technical terminology intensive and are tagged with index terms from the CCS and MCS topic hierarchies. Extracting keywords/topics for each of these documents will yield the mappings $\{topics_{ACM} \rightarrow tags_{CCS}\}$ and $\{topics_{AMS} \rightarrow tags_{MSC}\}$. The next logical step is to derive cross-domain mappings such as

1. Cross-domain generalization: $topics_{ACM} \rightarrow tags_{MSC}$
2. Cross-domain specialization: $tags_{ACM} \rightarrow topics_{MSC}$

At this stage we can train a model to accept computing science related topics and associate them with relevant math topics. This model forms the crux of the system we are attempting to build.

Software/Packages:

1. NLP based entity/concept identification extraction: Alchemy API
2. Topic modeling library for Python : Mallet (http://mallet.cs.umass.edu/)
3. Machine learning packages for Python : Shogun Toolbox (http://www.shogun-toolbox.org/) or PyML (http://pyml.sourceforge.net/)

Relevant papers

1. Arthur Asuncion, Max Welling, Padhraic Smyth, and Yee Whye Teh. 2009. On smoothing and inference for topic models. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence* (UAI '09). AUAI Press, Arlington, Virginia, United States, 27-34.
2. Steffen Bickel, Michael Brückner, and Tobias Scheffer. 2009. Discriminative Learning Under Covariate Shift. *J. Mach. Learn. Res.* 10 (December 2009), 2137-2155.

Project Plan (for mid-term review)

Deliverables	Deshpande, Amol	Peshave, Akshay
Training set	*Computing domain*	*Mathematics domain*
Elementary prediction model	*Computing text* $\rightarrow$ $tags_{CCS}$	*Math text* $\rightarrow$ $tags_{MSC}$
Possible approaches for cross-domain inference	✔	✔