Thread-level Analysis over Technical User Forum Data

Li Wang, Su Nam Kim and Timothy Baldwin

NICTA VRL
Department of Computer Science and Software Engineering
University of Melbourne
VIC 3010 Australia

December 9, 2010

Introduction

Motivation

- 'Information sharing' in social media
- Valuable information is being generated
- The information is not easily accessible
- A typical example: 'online forums'
- Little research in this domain

Example Thread

HTML Input Code - CNET Coding & scripting

User A Post 1	**HTML Input Code** ...Please can someone tell me how to create an input box that asks the user to enter their ID, and then allows them to press go. It will then redirect to the page ...
User B Post 2	**Re: html input code** Part 1: create a form with a text field. See ... Part 2: give it a Javascript action
User C Post 3	**asp.net c\# video** I've prepared for you video.link click ...
User A Post 4	**Thank You!** Thanks a lot for that ... I have Microsoft Visual Studio 6, what program should I do this in? Lastly, how do I actually include this in my site? ...
User D Post 5	**A little more help** ... You would simply do it this way: ... You could also just ... An example of this is ...

Source: http://forums.cnet.com/7723-6615_102-324299.html

Example Thread

HTML Input Code - CNET Coding & scripting

User A Post 1	**HTML Input Code** ...Please can someone tell me how to create an input box that asks the user to enter their ID, and then allows them to press go. It will then redirect to the page ...
User B Post 2	**Re: html input code** Part 1: create a form with a text field. See ... Part 2: give it a Javascript action
User C Post 3	**asp.net c\# video** I've prepared for you video.link click ...
User A Post 4	**Thank You!** Thanks a lot for that ... I have Microsoft Visual Studio 6, what program should I do this in? Lastly, how do I actually include this in my site? ...
User D Post 5	**A little more help** ... You would simply do it this way: ... You could also just ... An example of this is ...

External Link

External Video

500 words in total

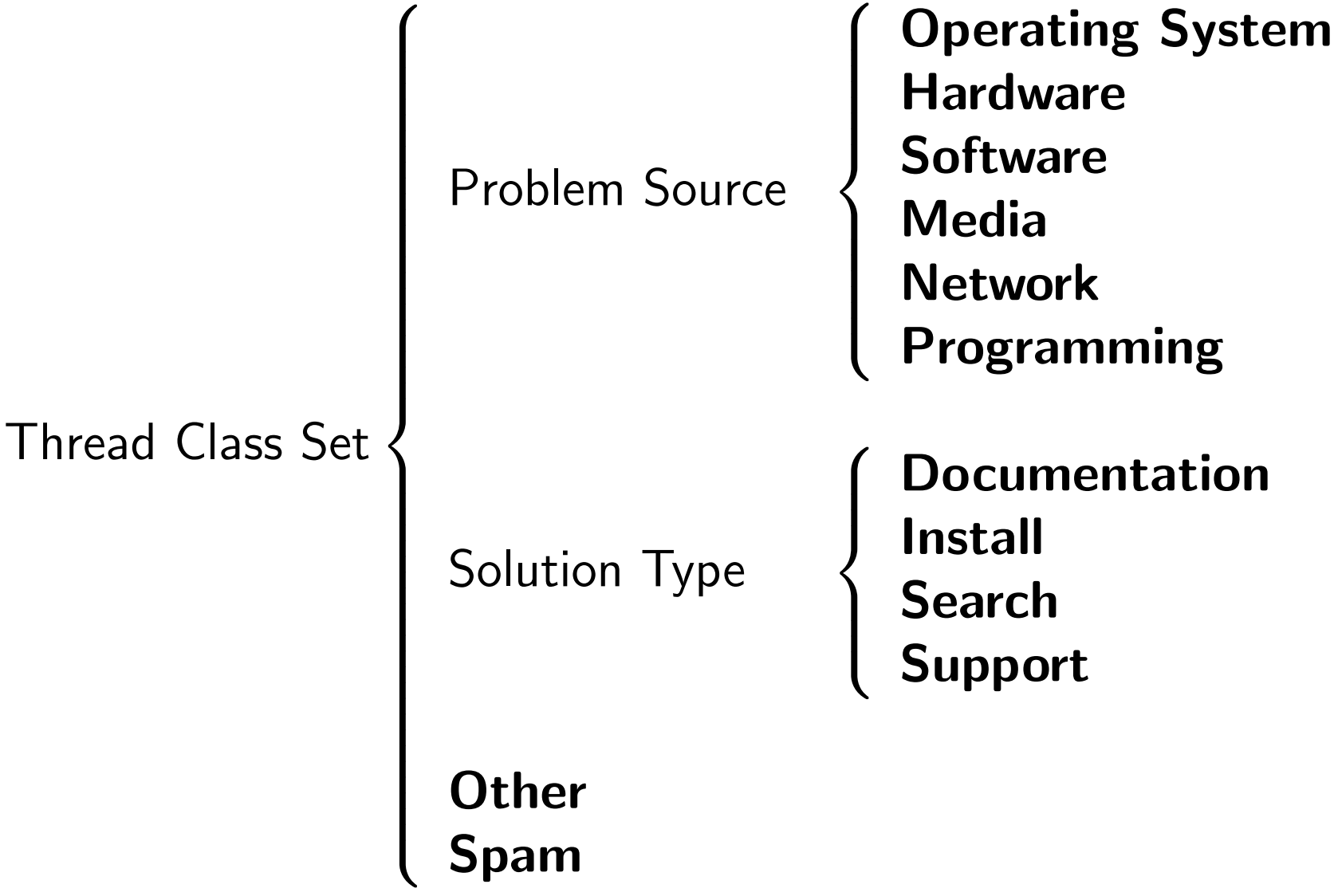
Source: http://forums.cnet.com/7723-6615_102-324299.html

Aim and Approach in a Nutshell

- The aim of the research
  - help users to more easily access existing information in online forums which relate to their questions
- The approach
  - automatically identify the topics of threads via text mining troubleshooting-oriented, computer-related technical user forum data (Baldwin et al., 2010)
- Contribution
  - designing a modular thread-level class set
  - constructing and publishing an annotated dataset
  - performing preliminary thread-level experiments over the dataset

Class Definition

Class Set Structure



Problem Source

- **Operating system:** Operating system
- **Hardware:** Core computer components, including core external components (e.g. a keyboard)
- **Software:** Software-related issues, including applications and programming tools
- **Media:** Non-standard external components or peripheral devices (e.g. a printer)
- **Network:** Network issues (e.g. connection speed, and installing a physical network)
- **Programming:** Coding and design issues relating to programming

Solution Type

- **Documentation:** How to use a certain function, select a computer/component, or perform a task
- **Install:** How to install a component
- **Search:** Search for a particular computer or component (e.g. a software package)
- **Support:** How to fix a problem with a computer or component

Miscellaneous

- **Other:** Troubleshooting-related, but the problem source is not included in the problem source set
- **Spam:** The thread is not troubleshooting-related

Annotation Class Set

	Annotation class set (26 classes)	
Combination of Problem Source and Solution Type classes	OS-Documentation	OS-Install
	OS-Search	OS-Support
	HW-Documentation	HW-Install
	HW-Search	HW-Support
	SW-Documentation	SW-Install
	SW-Search	SW-Support
	Media-Documentation	Media-Install
	Media-Search	Media-Support
	Network-Documentation	Network-Install
	Network-Search	Network-Support
	Programming-Documentation	Programming-Install
	Programming-Search	Programming-Support
Miscellaneous classes	Other	Spam

Example Thread

HTML Input Code - CNET Coding & scripting

User A Post 1	**HTML Input Code** ...Please can someone tell me how to create an input box that asks the user to enter their ID, and then allows them to press go. It will then redirect to the page ...
User B Post 2	**Re: html input code** Part 1: create a form with a text field. See ... Part 2: give it a Javascript action
User C Post 3	**asp.net c\# video** I've prepared for you video.link click ...
User A Post 4	**Thank You!** Thanks a lot for that ... I have Microsoft Visual Studio 6, what program should I do this in? Lastly, how do I actually include this in my site? ...
User D Post 5	**A little more help** ... You would simply do it this way: ... You could also just ... An example of this is ...

(Problem Source)
Programming

+

(Solution Type)
Documentation

||

(Thread Topic)
Programming-Documentation

Data, Methodology and Results

Data Collection

- 1000 threads were crawled from CNET forums and preprocessed.
- 150 threads were used for a pilot annotation, and reached a κ value of 0.43.
- 327 threads were annotated, and reached a κ value of 0.74.
- Most confusion is from Hardware vs. Media, and Documentation vs. Support.

Source: http://forums.cnet.com/

Experimental Methodology

- Preprocessing
  - punctuation removal
  - case-folding
  - lemmatisation
  - stopping
- Feature representation
  - bag-of-words (BoW): concatenating preprocessed tokens of all posts in a thread to form a single meta-document
- Learners
  - Support Vector Machines (SVM)
  - multinominal Naïve Bayes (NB)
  - majority-class baseline ($\textsc{ZeroR}$)

References: Tsuruoka et al., 2005, Hsu and Lin, 2006, McCallum, 2002

Experimental Methodology

- Class set representation:
  - all 26 multiclasses (AllClass)
  - only the Problem Source class sub-set with the Other class and Spam class (Problem)
  - only the Solution Type class sub-set with the Other class and Spam class (Solution)
- Evaluation:
  - based on stratified 10-fold cross-validation
  - macro-averaged precision ($\mathcal{P}_M$), recall ($\mathcal{R}_M$), F-score ($\mathcal{F}_M$)
  - micro-averaged precision ($\mathcal{P}_\mu$), recall ($\mathcal{R}_\mu$), F-score ($\mathcal{F}_\mu$)
  - mainly micro-averaged statistics
- Statistical significance test
  - randomised estimation with $p < 0.05$.

Reference: Yeh, 2000

Experiments over Three Class Sets

- The performance of different learners over AllClass, Problem and Solution

Class Space	Learner	$\mathcal{P}_M$	$\mathcal{R}_M$	$\mathcal{F}_M$	$\mathcal{P}_\mu/\mathcal{R}_\mu/\mathcal{F}_\mu$
AllClass	ZeroR	.006	.018	.009	.038
	SVM	.268	**.248**	**.246**	**.382**
	NB	**.306**	.211	.182	.333
Problem	ZeroR	.038	.142	.060	.266
	SVM	.564	**.485**	**.500**	.661
	NB	**.574**	.483	.481	**.691**
Solution	ZeroR	.122	.168	.140	.304
	SVM	.500	**.387**	**.413**	**.575**
	NB	**.513**	.270	.246	.520

Class Composition

- Results for class composition of the separate predictions from the PROBLEM and SOLUTION classifiers

PROBLEM	SOLUTION	ALLCLASS Results			
Learner	Learner	$\mathcal{P}_M$	$\mathcal{R}_M$	$\mathcal{F}_M$	$\mathcal{P}_\mu/\mathcal{R}_\mu/\mathcal{F}_\mu$
SVM	SVM	.345	**.313**	.314	.434
NB	SVM	**.379**	.310	**.316**	**.443**
SVM	NB	.278	.259	.229	.398
NB	NB	.268	.247	.206	.398

  - The best $\mathcal{F}_\mu$ (0.443) from class composition is significantly better than the best $\mathcal{F}_\mu$ (0.382) from multiclass classification approaches.

- Findings: class composition is effective in boosting overall classification performance.

Summary

- In this paper, we present:
  - a modular task formulation
  - a novel dataset
  - results from preliminary classification experiments
- Encouraging results from the class composition
- Possible future direction
  - feature engineering
  - text normalisation
  - hierarchical classification

Reference: Dekel et al., 2004, Tsochantaridis et al., 2005

References I

Timothy Baldwin, David Martinez, and Richard B. Penman. Automatic thread classification for Linux user forum information access. In *Proceedings of the 12th Australasian Document Computing Symposium (ADCS 2007)*, pages 72–79, Melbourne, Australia, 2007.

Timothy Baldwin, David Martinez, Richard Penman, Su Nam Kim, Marco Lui, Li Wang, and Andrew MacKinlay. Intelligent Linux information access by data mining: the ILIAD project. In *Proceedings of the NAACL 2010 Workshop on Computational Linguistics in a World of Social Media: #SocialMedia*, pages 15–16, Los Angeles, USA, 2010.

Ofer Dekel, Joseph Keshet, and Yoram Singer. Large margin hierarchical classification. In *Proceedings of the 21st International Conference on Machine Learning (ICML 2004)*, Banff, Canada, 2004.

Jonathan L. Elsas and Jaime G. Carbonell. It pays to be picky: An evaluation of thread retrieval in online forums. In *Proc. SIGIR'09*, pages 714–715, 2009.

Chih-Wei Hsu and Chih-Jen Lin. BSVM. `http://www.csie.ntu.edu.tw/~cjlin/bsvm/`, 2006.

Su Nam Kim, Li Wang, and Timothy Baldwin. Tagging and linking web forum posts. In *Proceedings of the 14th Conference on Computational Natural Language Learning (CoNLL-2010)*, pages 192–202, Uppsala, Sweden, 2010.

References II

Marco Lui and Timothy Baldwin. You are what you post: User-level features in threaded discourse. In *Proceedings of the 14th Australasian Document Computing Symposium (ADCS 2009)*, Sydney, Australia, 2009.

Marco Lui and Timothy Baldwin. Classifying user forum participants: Separating the gurus from the hacks, and other tales of the internet. In *Proceedings of the 2010 Australasian Language Technology Workshop (ALTW 2010)*, Melbourne, Australia, 2010.

Andrew Kachites McCallum. MALLET: A machine learning for language toolkit. `http://mallet.cs.umass.edu/`, 2002.

Jangwon Seo, W. Bruce Croft, and David A. Smith. Online community search using thread structure. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management (CIKM 2009)*, pages 1907–1910, Hong Kong, China, 2009.

Ioannis Tsochantaridis, Thorsten Joachims, Thomas Hofmann, and Yasemin Altun. Large margin methods for structured and interdependent output variables. *Journal of Machine Learning Research*, 6(Sep):1453–1484, 2005.

Yoshimasa Tsuruoka, Yuka Tateishi, Jin-Dong Kim, Tomoko Ohta, John McNaught, Sophia Ananiadou, and Jun'ichi Tsujii. Developing a robust part-of-speech tagger for biomedical text. In *Proceedings of the Advances in Informatics - 10th Panhellenic Conference on Informatics, LNCS 3746*, pages 382–392, Volos, Greece, 2005.

Alexander Yeh. More accurate tests for the statistical significance of result differences. In *Proceedings of the 18th International Conference on Computational Linguistics (COLING 2000)*, pages 947–953, Saarbrücken, Germany, 2000.

Questions?

Characteristics of online forum data

- Different from plain text documents
  - Complex structures
  - Posts are dynamic
  - Informal language is used
- Different from CQAs and FAQs
  - Broad and shallow vs. specific and in-depth
  - Longer history and more data
  - Multi-purpose
  - Asynchronous

CNET Forums and Sub-forums

Forum	Sub-forum		
Operating Systems	Windows 7	Windows Vista	Windows XP
	Windows 2000/NT	Windows ME	Windows 95/98
	Windows Mobile	Mac OS	Linux
Software	Audio & video	Browsers	CNET Download site
	E-mail, chat, & VoIP	Mac software	Office & productivity
	PC utilities	Photography & design	Spyware, viruses, & security
	Webware	Windows Live	
Hardware	Dell	Desktops	Laptops
	Mac hardware	Networking & wireless	PC hardware
	Peripherals	Storage	
Web Development	Coding & scripting	Web design & hosting	

Table: Data source forums and sub-forums

Class Distribution

Annotation class set (26 classes)	
OS-Documentation: 27	OS-Install: 9
OS-Search: 1	OS-Support: 28
HW-Documentation: 28	HW-Install: 5
HW-Search: 5	HW-Support: 23
SW-Documentation: 29	SW-Install: 3
SW-Search: 23	SW-Support: 29
Media-Documentation: 14	Media-Install: 8
Media-Search: 13	Media-Support: 15
Network-Documentation: 9	Network-Install: 9
Network-Search: 3	Network-Support: 18
Programming-Documentation: 7	Programming-Install: 0
Programming-Search: 0	Programming-Support: 1
Other: 8	Spam: 12

Semi-supervised Learning

- Semi-supervised Learning : SVMlin
  - Multi-switch linear Transductive L2-SVMs
  - Deterministic Annealing (DA) for Semi-supervised Linear L2-SVMs
  - no significant improvements

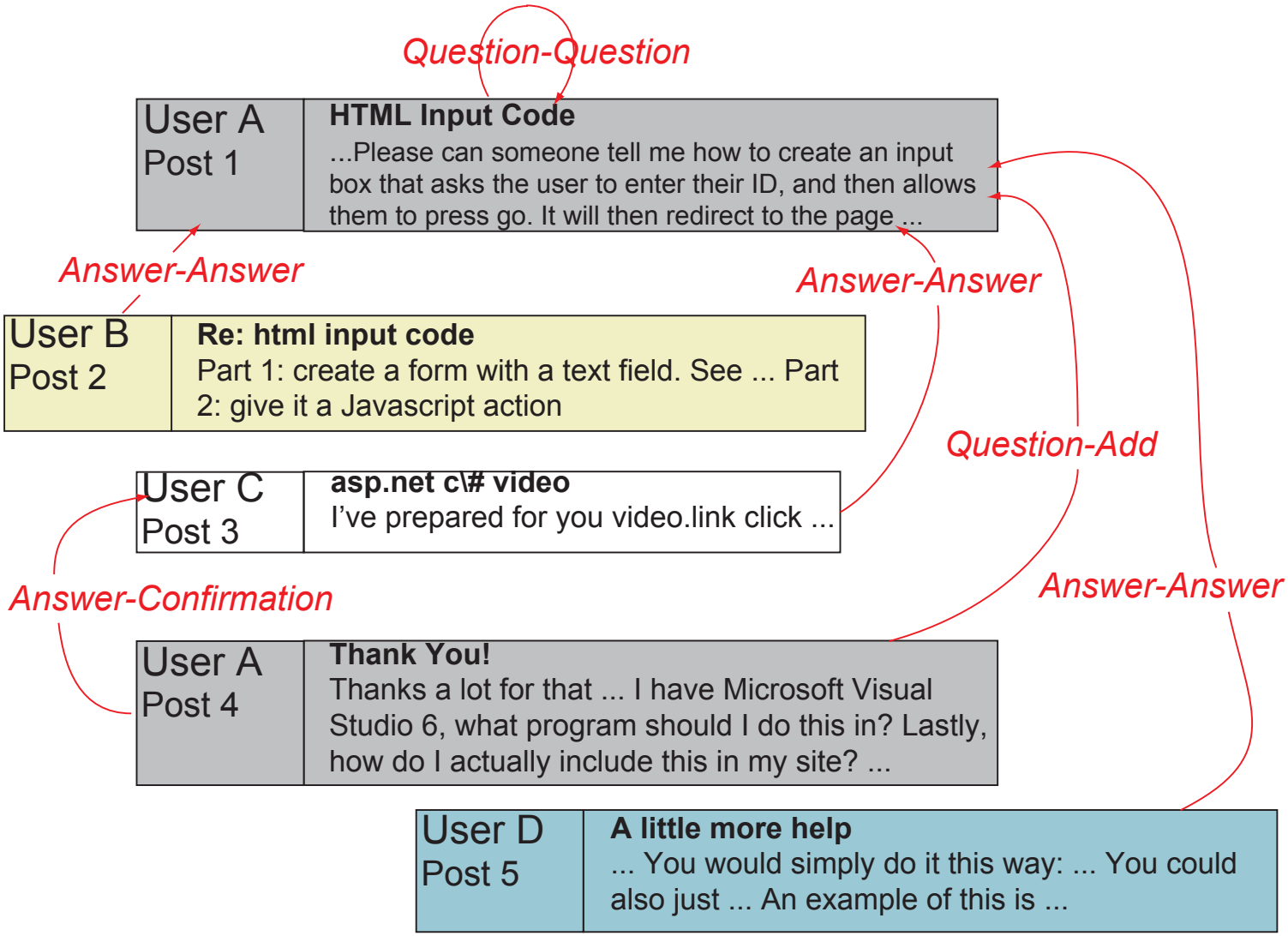
Source: http://vikas.sindhwani.org/svmlin.html

Thread Characteristic Classification

- Timothy Baldwin, David Martinez, and Richard B. Penman. Automatic thread classification for Linux user forum information access. In *Proceedings of the 12th Australasian Document Computing Symposium (ADCS 2007)*, pages 72–79, Melbourne, Australia, 2007.
- In the context of Linux web user forums
- Focus on classifying threads according to:
  - Task orientation
  - Completeness
  - Solvedness

Reference: Baldwin et al., 2007

Tagging and Linking Web Forum Posts



Reference: Kim et al., 2010

Classifying User Forum Participants

- User characteristic classification
  - Clarity
  - Proficiency
  - Positivity
  - Effort
- More about this research at 9:00 am, 10 December

Reference: Lui and Baldwin, 2010

User-level Features in Threaded Discourse

- Describe users based on their posts
- Based on existing techniques
- User-level features for post rating
  - Aggregate: aggregation over features describing individual posts
  - Network-Based: Author Network and Thread Network

Reference: Lui and Baldwin, 2009

An Evaluation of Thread Retrieval in Online Forums

- Treat the task as an information retrieval task
- Findings:
  - thread structure is important in thread ranking
  - selective models outperform inclusive models

Reference: Elsas and Carbonell, 2009

Thread Retrieval Using Thread Structure

- Treat the task as an information retrieval task
- Goals:
  - discover and annotate thread structures, based on interactions between community members
  - improve retrieval performance by exploiting the thread structure

Reference: Seo et al., 2009