Supplementary material for Subtle Topic Models and Discovering Subtly Manifested Software Concerns Automatically

Mrinal Kanti Das[†]   MRINAL@CSA.IISC.ERNET.IN
Suparna Bhattacharya   BSUPARNA@IN.IBM.COM
IBM Research-India

Chiranjib Bhattacharyya[†]   CHIRU@CSA.IISC.ERNET.IN
K. Gopinath[†]   GOPI@CSA.IISC.ERNET.IN

†Department of Computer Science and Automation,
Indian Institute of Science, Bangalore, India

1. Concern "Verify" in Berkeley-DB

Verify, is an important cross-cutting concern(CCC) in Berkeley-DB. According to the manual analysis in (Kastner et al., 2007), *verify* occurs individually as a main concern and also in 7 derivative concerns. We list them below. This shows the cross-cutting nature of *verify* and also significance in Berkeley-DB.

- Verifier
- Derivative_Latches_Statistics_Verifier
- Derivative_Latches_Verifier_INCompressor
- Derivative_LoggingInfo_Statistics_Verifier
- Derivative_Statistics_Verifier
- Derivative_Statistics_Verifier_DeleteOp
- Derivative_Statistics_Verifier_INCompressor
- Derivative_Verifier_INCompressor.

2. Comparison on average recall by varying subtlety

We plot the average recall (over all datasets) of the topics to check how performance of the models change by increasing the degree of subtlety (DoS), see Figure 1. We see that STM consistently performs better than all models in this range of DoS. FTM which is the best among the existing models perform quite steadily until DoS is less than 0.5, after that we see a slight declination. The ability of MG-LDA to detect local topics makes it better than HDP and the difference is more for higher DoS, where it also narrows down the gap with FTM.
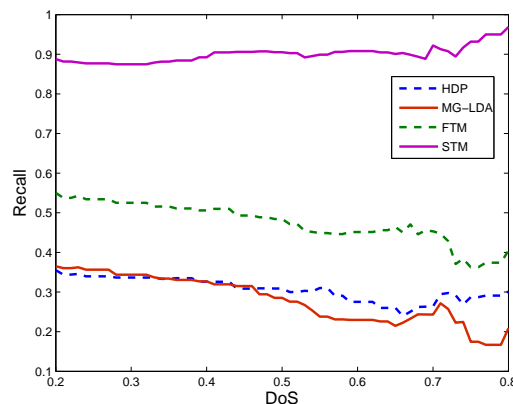

Figure 1. Plot of average recall by varying subtlety (DoS) to compare the models.

3. Detailed results on subtle topics

We list all the manually created gold-standard topics(keywords), their degree of subtlety (DoS) and performance of all the models on each individual topic. The result is given for all the datasets: BerkeleyDB(Table 1), JHotDraw(Table 2), NIPS-05(Table 3) and Obama-speech(Table 4).

References

Kastner, Christian, Apel, Sven, and Batory, Don. A Case Study Implementing Features Using AspectJ. In *Proceedings of the 11th International Software Product Line Conference(SPLC)*, pp. 223–232, 2007.

Table 1. Topics from BerkeleyDB with subtlety more than 0.2. Recall and coherence results are reported for those super-topics for all the models.

BerkeleyDB									
keywords $\mathcal{K}$	DoS	Recall				Topic Coherence			
		HDP	MG-LDA	FTM	STM	HDP	MG-LDA	FTM	STM
locker,locks,lock	0.23	0.67	0.67	1	1	-4.4	-48.72	-39.63	-19.14
evict,corrupted,entry	0.4	0.33	0.67	0.33	0.33	-10.26	-54.61	-79.46	-5.4
parent,node,child	0.44	0.67	0.33	1	1	-54.46	-8.26	-4.67	-20.43
stats,count	0.52	0.5	0.5	0.5	1	-75.42	-83.49	-83.45	-70.15
trace,level,info,config	0.52	0.25	0.25	1	1	-73.71	-22.92	-31.33	-11.09
transaction,checkpoint,recovery	0.53	0.33	0.67	0.67	1	-76.78	-79.21	-25.47	-13.83
verify,config,keys	0.68	0	0	0	1	-54.92	-6.13	-65.01	-31.76
trace,event,cursor,trigger	0.71	0.25	0.5	0.75	1	-54.61	-51	-36.03	-17.73
files,utilization,summary	0.72	0.67	1	1	1	-63.11	-32.25	-6.07	-6.54
checksum,errors,validate	0.74	0.33	1	0.67	1	-61.31	-23.1	-34.84	-16.97
fsync,sync,tracker	0.84	0.33	0.67	0.33	1	-4.32	-62.09	-35.2	-28.78
Average		0.42	0.59	0.68	0.94	-48.48	-42.89	-40.11	-21.99

Table 2. Topics from JHotDraw with subtlety more than 0.2. Recall and coherence results are reported for those super-topics for all the models.

JHotDraw									
keywords $\mathcal{K}$	DoS	Recall				Topic Coherence			
		HDP	MG-LDA	FTM	STM	HDP	MG-LDA	FTM	STM
track,slider,component	0.21	0.33	0	0.67	1	-79.5	0	-84.01	-20.33
pane,border,scroll	0.21	1	0.67	0.67	1	-42.74	-31.23	-35.22	-31.79
formatter,fill,opacity,color	0.22	0.25	0.25	0.25	1	-9.51	-24.67	-81.11	-17.35
affine,math,transformation	0.53	0.33	0.33	0.33	0.67	-77.38	-71.4	-62.47	-30.28
zoom,factor	0.58	0.5	0	0	1	-79.93	0	0	-51.07
collection,family,families	0.62	0.67	0.33	0.33	1	-77.31	-81.2	-45.75	-8.72
rendering	0.67	0	0	1	1	0	0	-53.94	-62.04
roundrect	0.7	0	0	0	1	0	0	0	-98.64
styles,style	0.7	0	0	1	1	0	0	-70.12	-65.37
nano,xmldom	0.79	0	0	0	1	0	0	0	-76.07
Average		0.31	0.16	0.42	0.97	-36.64	-52.13	-43.26	-46.17

Table 3. Topics from NIPS-05 with subtlety more than 0.2. Recall and coherence results are reported for those super-topics for all the models.

NIPS-05									
keywords $\mathcal{K}$	DoS	Recall				Topic Coherence			
		HDP	MG-LDA	FTM	STM	HDP	MG-LDA	FTM	STM
active,learning	0.19	0.5	1	1	1	-1.38	-40.48	-46.68	-27.81
training,classification,learning	0.21	0.33	0.67	1	1	-1.38	-52.67	-13.9	-22.44
online,algorithm	0.27	0.5	1	1	1	-10.55	-48.11	-32.72	-27.42
variational,learning,bayesian	0.32	0.33	0.67	1	0.67	-97.51	-78.31	-23.74	-34.88
kernel,function, functions,kernels	0.33	0.5	0.5	0.5	0.75	-46.2	-5.41	-22.87	-29.09
distribution,sample, samples,probability	0.38	0.75	0.5	0.75	0.5	-63.72	-48.8	-5.71	-76.38
risk,function,loss,error	0.44	0.75	0.5	0.75	0.75	-53.03	-54.31	-29.78	-16.98
convex,loss,function	0.46	0.33	1	0.67	1	-10.55	-5.41	-6.9	-35.54
bayesian,models, prior,likelihood	0.46	0.25	0.5	0.5	0.75	-64.16	-72.25	-71.67	-14.99
feature,features,score	0.48	0.33	0.67	0.67	1	-80.62	-64.66	-68.85	-21.66
feature,selection,lasso	0.54	0.33	0.33	0.67	0.67	-43.84	-5.21	-7.34	-33.54
statistical,test,tests	0.56	1	0.67	0.67	1	-38.7	-74.71	-74.09	-16
topic,model,topics,dirichlet	0.57	0.25	0.25	0.5	0.75	-1.38	-9.25	-57.37	-38.06
neurons,neuron, network,synaptic	0.58	0.5	0.5	0.75	0.75	-66.94	-56.9	-12.65	-18.44
convex,optimization, optimal,solution	0.65	0	0.25	0.25	0.75	0	-5.41	-63.5	-52.12
clustering,cluster, similarity,distance	0.69	0.25	0.25	0.5	0.25	-20.07	-5.41	-63.06	-42.45
node,nodes,propagation, belief,network	0.73	0.6	0.2	0.2	0.6	-45.92	-63.7	-90.57	-38.31
chip,processing, architecture,circuit	0.76	0.25	0.25	0.25	0.75	-70.74	-90	-83.87	-48.74
video,texture,resolution,image	0.79	0.5	0	0.5	0.75	-68.93	0	-34.2	-40.45
walk,walks,steering,robot	0.89	0.25	0.5	0.25	0.75	-25.68	-10.18	-12.81	-55.39
cochlear,cochlea	0.95	0	0	0	1	0	0	0	-44.35
Average		0.41	0.49	0.56	0.79	-40.23	-38.55	-37.38	-34.84

Table 4. Topics from Obama-speech with subtlety more than 0.2. Recall and coherence results are reported for those super-topics for all the models.

Obama-speech									
keywords $\mathcal{K}$	DoS	Recall				Topic Coherence			
		HDP	MG-LDA	FTM	STM	HDP	MG-LDA	FTM	STM
health,care,insurance	0.19	0	1	1	1	0	-27.36	-33.32	-9.17
economic,economy,									
crisis,recovery	0.35	0.25	0.5	0.75	0.75	-36.72	-10.25	-36.36	-22.09
energy,cars,fuel	0.5	0.67	0.67	1	1	-69.27	-49.83	-31.68	-34.66
cyber,security,internet	0.52	0	0.33	0.67	1	0	-74.51	-44.92	-54.97
former,president,clinton	0.52	0.33	0.33	0.33	1	-85.62	-94.21	-40.14	-23.82
deficit,cuts,budget	0.54	0	0.67	0.33	1	0	-35.67	-83.89	-10.22
afghanistan,afghan,									
qaeda,laden	0.57	0.25	0	0.25	0.75	-85.77	0	-24.58	-37.63
school,students,college,schools	0.64	0.25	0.5	0.25	1	-86.97	-60.6	-83.83	-31.16
black,white	0.65	1	0	1	1	-37.83	0	-62.19	-75.49
egypt	0.66	0	0	0	1	0	0	0	-83.44
libyan,libyans,libya	0.72	0	0	0.67	1	0	0	-67.88	-17.72
finance,financial,banks	0.74	0	0	0.33	0.67	0	0	-84.58	-55.44
nasa,space	0.85	1	0	1	1	-43.99	0	-41.63	-50.32
regulations,infrastructure,									
employees	0.88	0.33	0	0.33	1	-76.21	0	-80.18	-42.57
latin,americas	0.89	0	0	1	1	0	0	-55.16	-59.37
carnegie,mellon,technology	0.91	0	0	0.33	1	0	0	-80.05	-41.53
Average		0.26	0.27	0.58	0.95	-32.65	-22.03	-53.15	-40.6