Digg Dataset

Yu-Ru Lin (yu-ru.lin@asu.edu), February 2009

The dataset mentioned in our experiment is prepared by Yu-Ru Lin. The dataset is a subset of data scrapped from Digg by Munmun De Choudnury during January 2009.

The dataset includes Digg stories, users and their actions (submit, digg, comment and reply) with respect to the stories, as well as the explicit friendship (contact) relation among these users. To analyze users' topical interests, we also retrieve the topics of the stories and extract keywords from the stories' titles.

From this dataset, we select 5 facets (user, story, comment, keyword and topic) and build 6 relations among them. These facets and relations are summarized in Figure 1 and Table 1. Except for the contact relation, all relations have timestamps. We assume the contact relation is static and consider the other relations as dynamic.

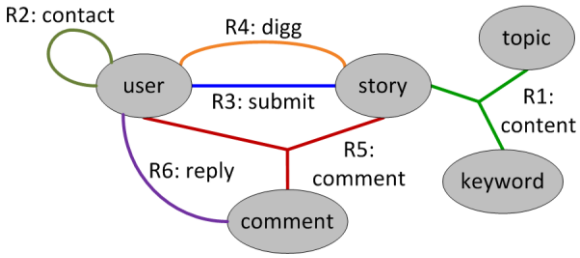

Figure 1: Facets and relations included in Digg dataset.

Relation	Tensor / incident facets	#Tuple
(R1) content	dynamic (story, keyword, topic)	151,77
(R2) contact	static (user, user)	56,440
(R3) submit	dynamic (user, story)	44,005
(R4) digg	dynamic (user, story)	1,157,5
(R5) comment	dynamic (user, story, comment)	241,80
(R6) reply	dynamic (user, comment)	94,551

Table 1: Summary of the relations in Digg dataset.

There are two formats of the dataset:

1. Text files (*.m, , in the folder "data/texts"): Each facet has a corresponding file <facet-name>_keys_<version>.m, and each relation has a corresponding file <relation_name> _<version>.m. Each line in a facet file corresponds to a unique entity, and the line number is the index of the entity (e.g. user_id) in our dataset. Each line in a relation file corresponds to a relation. Example line formats are given below:
   user_keys_v1.m: <Digg_user_id>
   story_keys_v1.m: <Digg_story_id>
   user_submit_story_v1.m: <user_id> <story_id> <timestamp> <month_id> <week_id> <3day_id>

user_submit_story_keyword_topic_v1.m: <user_id> <story_id> <keyword_id> <topic_id> <timestamp> <month_id> <week_id> <3day_id>

…

Note that the timestamp is the date vector in matlab where 1 corresponds to 1-Jan-0000 (ref. function "datenum" in matlab). The <month_id>,<week_id>, and <3day_id> are the slot index segmented per month, week and every three days starting from August 1, 2008. In the experiment we use the 3-day time slots.

2. Binary files (*.mat, in the folder "data/sptensors"): For experiment we have convert the text files into binary files which can be loaded as sparse tensor objects in matlab. The binary files provided here are in the 3-day time slots ranging from August 1 to 27, totally 9 slots. Each tensor sequence has corresponding 9 files <tensor_name>3d<3day_id>_<version>.mat, for example: sD3d7_v1.mat: the digg tensor of slot t=7 (size: 9583x44005) where the first mode is user and second mode is story.

Please email me if you have any question about the dataset.