LASTA: Large Scale Topic Assignment on Multiple Social Networks

Nemanja Spasojevic, Jinyun Yan, Adithya Rao, Prantik Bhattacharyya
{nemanja, jinyun, adithya, prantik}@klout.com

Introduction

In this study, we describe a scalable engineering system deployed in production that mines topical interests from five social networks and assigns over 10,000 topics to hundreds of millions of users on a daily basis. We extract and analyze features for topic inference that extend beyond authored text., and show that using a diverse set of features and cross-network information can lead to a better understanding of a user's interests. We focus primarily on assigning topics for a user that other users can socially recognize and acknowledge. This approach helps in building applications that are meaningful in the context of the social identity of a user.

Klout, Inc. is a social media platform that aggregates and analyzes data from social networks. A user on Klout can connect one or more of the above social profiles to form one unique profile. We present Klout's topic system called `LASTA', (Large Scale Topic Assignment), that focuses on inputs from four major social networking sites: Facebook (FB), Twitter (TW), GooglePlus (GP) and LinkedIn (LI), along with Wikipedia (WIKI). We evaluate LASTA's topic assignment system on an internal labeled corpus of 32,264 user-topic labels generated from real users.

Figure 1: Hierarchical Ontology Overview

Data Landscape

One of LASTA's goals is to understand different behaviors presented by users in different networks. The first figure below shows the distribution of the number of phrases In order to illustrate different user behavior and varied vocabulary choice across social networks, we examine the phrase overlap in messages created by a user who has connected multiple social networks to their Klout profile.

Facebook: Authored status updates, shared URL pages, commented and liked posts, text and tags associated with videos and pictures.

Twitter: Authored tweets, re-tweets, mentions and replies on other tweets, shared URL pages, subscribed, created and joined lists.

LinkedIn: Comments on posts, skills stated by the user and endorsed by connections.

Google+: Authored messages, re-shares, comments, shared URL pages and plus-ones.

Wikipedia: Wikipedia pages for well known personalities.

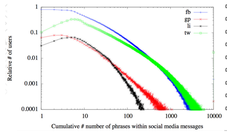


Fig. 1. Verbosity distribution across social networks

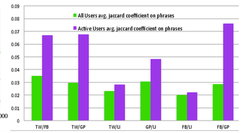
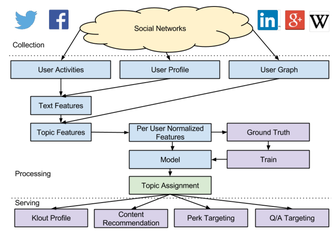


Fig. 2. Phrase overlap on social networks

Pipeline

Our backend system can be broken into two main components: data collection, and data processing. At the data collection stage Klout fetches the user's profile, activities and connection graphs from various social networks. This data is parsed and stored in normalized form. The data processing pipeline expresses topical interests for each user as a ranked list of topics. The inferred topic list is used for multiple applications including generating a unified user profile, content recommendation, targeting and question answering.



User Profile: A user may explicitly state some of his interests in his profile description on a social network.

User Activities: Text and URL information derived from user activities and messages is a valuable source for understanding topic associations for user.

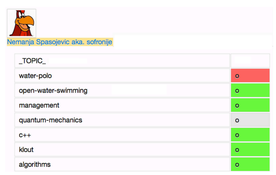
User Graph: We also collect the connection graph of a user within social networks. Such a connection graph has users as nodes and directed edges between pairs of users. This includes follower and following edges on TW, which are unidirectional relationships, and friend edges on FB, which are bidirectional relationships.

Bags of Topics: Bags-of-phrases are first extracted from inputs, by matching against a dictionary of approximately 2 million phrases. As some of these sources change daily, the dictionary dynamically updates itself to include the latest phrases in social media. Bags-of-phrases are then mapped to the topic ontology and are transformed into bags-of-topics, effectively reducing the dimensionality of the text from 2 million phrases to around 10,000 topics. The bags-of-topics thus generated have associated strengths for each topic in the bag.

Ground Truth: Our ground truth data is aimed at generating labels for socially recognizable user topics. In order to collect ground truth for building models, we designed a simple web app to collect ground truth data with labels for user-topic interests. In this experimental setup, the system pulls up a set of the participant's user graph first degree connections, and randomly assigns topics to the connections. The evaluator then gives positive or negative feedback, depending if the topic is good or bad match for his connection.

Table 3: Statistics on ground truth dataset

Statistics	Value
# of participants	43
# of evaluated users	766
# of (user, topic) labels	32,264
# of positive (user, topic) labels	17,208
# of negative (user, topic) labels	15,056



Feature Generation

Each feature is represented as a combination of three characteristics -- <network>_<data-source>_<attribution>. In particular, attribution denotes the relation of the input source to the user. It may be one of the following --

Generated: Originally generated or authored content by the user, including posts, tweets, comments and profiles.

Reacted: Content generated by another user (actor), but as a reaction to content originally authored by the user under consideration. This includes comments, re-tweets, and replies.

Credited: In this case the user has no direct association with the content from which the feature was derived.

Graph: The topics aggregated from a user's first degree connections are attributed to the user.

For each user, bags of topics are derived in the manner that encode the above information. Features are then generated for each user-topic pair by exploding the bags of topics, and creating feature vectors for each pair.

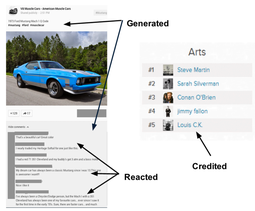


Table 4: Feature performance and coverage

Feature Source		P	R	C
Twitter				
GEN.	MSG TEXT 90 DAY	0.22	0.15	27.37
	URL 90 DAY	0.09	0.19	14.67
	URL META 90 DAY	0.33	0.14	11.63
REAC.	MSG TEXT 90 DAY	0.26	0.12	20.81
	URL META 90 DAY	0.36	0.11	10.66
CRED.	LIST	0.68	0.21	21.19
	URL META 90 DAY	0.37	0.10	4.81
	MSG TEXT 90 DAY	0.20	0.18	21.91
	MSG #TAG 90 DAY	0.43	0.11	13.70
GRAPH	FOLLOWERS	0.08	0.26	52.41
	FOLLOWING	0.10	0.31	52.77
Facebook				
GEN.	MSG TEXT 90 DAY	0.17	0.07	29.20
	URL 90 DAY	0.08	0.12	9.52
	URL META 90 DAY	0.21	0.06	7.08
REAC.	MSG TEXT 90 DAY	0.12	0.08	45.58
	URL 90 DAY	0.05	0.12	19.82
	URL META 90 DAY	0.14	0.06	14.81
CRED.	MSG TEXT 90 DAY	0.15	0.06	13.46
GRAPH.	FRIENDS	0.08	0.25	63.66
Google Plus				
GEN.	MSG TEXT 90 DAY	0.23	0.04	1.61
	URL 90 DAY	0.09	0.15	0.34
	URL META 90 DAY	0.25	0.07	0.23
REAC.	MSG TEXT 90 DAY	0.11	0.05	1.68
	URL 90 DAY	0.05	0.08	0.46
	URL META 90 DAY	0.02	0.03	0.34
CRED.	MSG TEXT 90 DAY	0.16	0.05	0.69
LinkedIn				
GEN.	SKILLS	0.53	0.20	19.17
	INDUSTRY	0.56	0.10	16.63
Wikipedia				
CRED.	WIKI PAGE	0.18	0.28	0.11

Evaluation and Results

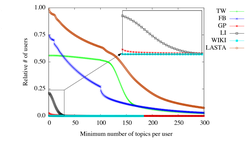
We cast the problem as a binary classification problem, in which the system must learn automatically to separate topics of interest from those that are not relevant to the user. We train our models using the feature vectors generated for the pairs against the labels from the labeled data. The models are trained using the logistic classifier, which learns a weight associated with each feature. The final bag of topics for a user are derived by applying these weights to the corresponding feature based bags of topics, and aggregating the strengths. In the paper, we discuss insights gained by comparing the performance of using all features versus using only subsets of features such as single networks, attributions and graph based features.

Table 8: LASTA topic assignment examples

User	Top 10 Topics
Marissa Mayer	yahoo, google, technology, business, twitter, social-media, flickr, design, marketing, seo, gmail
Lady Gaga	music, lady-gaga, celebrities, art, fashion, born-this-way, venus, entertainment, radio
Barack Obama	politics, affordable-care-act, healthcare, new-york-times, congress, chicago, twitter, washington, illinois

Cross-Network Analysis

We examine the distribution of topics in terms of number of topics assigned to users. for the same number of topics, LASTA always assigns topics to more users. Also, LASTA assigns more topics to each user compared to individual networks.



We examine similarities and differences between topical interests aggregated across users on different networks. We observe from the figure that users in each network have distinct topical interests. On FB and TW 'entertainment' is the most represented topic, whereas 'business' is the most represented one on LI, and 'technology' on GP. The left-most column shows the distribution of topics as assigned by LASTA.

Table 9: Super-topic percentage distribution across different networks

Super-topic	LASTA	TW	FB	LI	GP	WIKI
technology	23.972	19.706	11.559	33.420	22.822	8.247
entertainment	23.987	20.049	20.866	3.406	14.377	30.669
business	15.893	10.628	7.567	41.053	12.857	10.937
lifestyle	7.910	7.403	11.409	2.328	7.969	4.810
science-and-nature	4.431	3.705	3.604	1.266	4.682	3.208
arts-and-humanities	6.605	7.056	6.836	5.765	9.392	13.373
government-and-politics	3.547	4.763	4.388	2.182	3.534	5.261
sports-and-recreation	4.379	7.503	7.591	0.659	4.913	7.921
food-and-drink	2.671	7.228	11.863	0.819	7.255	2.142
health-and-wellness	1.976	3.894	5.150	1.691	4.083	1.867
fashion	1.439	2.645	2.945	0.732	2.776	2.203
education	1.443	2.375	3.485	3.369	2.170	4.058
news-and-media	0.966	1.722	0.899	2.597	1.060	4.366
travel-and-tourism	0.535	0.779	1.155	0.614	1.041	0.654
hobbies	0.246	0.543	0.683	0.100	1.070	0.285

Applications

LASTA is serving multiple personalized services at Klout:

User Targeting: Targeting influential users with messages and campaigns based on topics effectively propagates awareness in social networks.

Content Discovery: The topics deduced by LASTA provide utility to users in terms of serendipitous content discovery.

Question Answering: In a question answering scenario, a user in the system can ask a question, which can then be routed to specific users who may be able to answer the question, based on the topic of the question.

Conclusions and Future Work

LASTA assigns overs 10,000 topics to hundreds of millions of users spread across multiple social networks on a daily basis with a high accuracy. Future work to improve this system includes ontology improvements, other techniques for phrase-to-topic mapping and differentiating between topics of interest and topics of expertise. We hope that the engineering architecture and data transformation methodologies described here provide insights to build scalable and extendable topic mining systems.