LASTA: Large Scale Topic Assignment on Multiple Social Networks

Nemanja Spasojevic, Jinyun Yan, Adithya Rao, Prantik Bhattacharyya
Klout Inc.
77 Stillman Street
San Francisco, CA 94107
{nemanja, jinyun, adithya, prantik}@klout.com

ABSTRACT

Millions of people use social networks everyday to talk about a variety of subjects, publish opinions and share information. Understanding this data to infer user's topical interests is a challenging problem with applications in various data-powered products. In this paper, we present 'LASTA' (Large Scale Topic Assignment), a full production system used at Klout, Inc., which mines topical interests from five social networks and assigns over 10,000 topics to hundreds of millions of users on a daily basis. The system continuously collects streams of user data and is reactive to fresh information, updating topics for users as interests shift. LASTA generates over 50 distinct features derived from signals such as user generated posts and profiles, user reactions such as comments and retweets, user attributions such as lists, tags and endorsements, as well as signals based on social graph connections. We show that using this diverse set of features leads to a better representation of a user's topical interests as compared to using only generated text or only graph based features. We also show that using cross-network information for a user leads to a more complete and accurate understanding of the user's topics, as compared to using any single network. We evaluate LASTA's topic assignment system on an internal labeled corpus of 32,264 user-topic labels generated from real users.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Information filtering; Retrieval models

Keywords

Online Social Networks; Interest Mining; Topic Assignment; Large Scale; User Modeling; Distributed Systems

1. INTRODUCTION

Mining topical interests for users from social media is an interesting and important problem to solve, because the insights gained can be applied to many applications such as recommendation and targeting systems. Such systems can deliver accurate results tailored to each individual user, only if the user's interests are well understood. The task of interest mining from social media has many challenges that mainly lie in the characteristics of the data, viz. *size, noise* and *sparsity*. While the total volume of text generated on social media is huge, the size of each individual document tends to be very short. For example, posts on Twitter (tweets) are limited to 140 characters. Often the posts are also noisy due to abbreviations, grammatically inaccurate sentences, symbols such as emoticons and misspelled words [1]. Finally, because many users on social media are inactive, sporadically active or only tend to be passive consumers of content, the textual content available for topical inference is sparse for such users.

In this study, our contributions are as follows: 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. dWe extract and analyze features for topic inference that extend beyond authored text. We show that using a diverse set of features and cross-network information can lead to a better understanding of a user's interests. Compared to previous studies [2, 3, 4, 5] that attempt to mine all topics for a user, we focus primarily on assigning topics for a user that other users can socially recognize and acknowledge. For example, Warren Buffett is recognized for topics like 'Business', 'Finance' and 'Money', while his personal interests may include 'Cars' and 'Airplanes'. 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 like Twitter, Facebook, LinkedIn, Google Plus and Instagram, and other sources like Bing Search Engine and Wikipedia. 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).

To address the data challenges mentioned above, we consider the following approaches: We process information shared by users to get more context around individual user documents. To address data noise problems, we explode text into n-grams and map against an internal dictionary of approximately 2 million phrases to generate bags-of-phrases. We address data sparsity problems by extracting signals from

KDD'14, August 24-27, 2014, New York, New York, USA.

a user's reactions, such as comments or retweets on other user's posts. We also extract signals from posts in which a user is tagged or mentioned as well as from social graph connections, to increase data coverage for a given user.

We combine the signals mentioned above to generate over 50 distinct features. The set of features are categorized as following: *Generated, Reacted, Credited* and *Graph*. Features derived from user authored posts and profile information are categorized as *Generated*. *Reacted* features come from user reactions such as comments and retweets. *Credited* features are built from signals such as lists, tags and endorsements, while *Graph* features are based on social graph connections. We evaluate LASTA's topic assignment system on an internal labeled corpus of 32,264 user-topic labels generated from real users.

The underlying infrastructure is built on the Hadoop platform[1], an open-source implementation of MapReduce. The system uses HDFS as the file system and Hive[2] as the querying infrastructure. Each day, nearly a terabyte of raw data is ingested into our data warehouse. We hope that insights gained from our experience in building LASTA will prove valuable for the community to build future topic systems.

The rest of the paper is structured as follows. Section 2 discuses related work and formally introduces the problem statement. Section 3 describes system level details and data generation steps. Section 4 presents evaluation results and some interesting and useful findings. Section 5 discusses some of the application of LASTA implemented at Klout and conclude in section 6.

2. RELATED WORK & BACKGROUND

There are a variety of topic detection systems that have been proposed, and topic inference is a well studied area. However, the effectiveness of any given system is typically dependent on the specific domain or application under consideration. For example, modeling user interests is common practice for recommendation engines such as Amazon and Netflix, where the objective is to understand user interests in a particular domain such as products or movies. The user interests are often represented as latent vectors in recommender systems [6], and are derived from either explicit feedback, such as ratings, or implicit feedback such as clicks on products. Search engines also use topic inference to personalize results [7, 8], where user interests are learnt from click-history and browsing behaviors from search logs. Similarly, clicks on ads are used to model user interests in the domain of online display advertising [9].

In many topic inference settings, the individual documents have clean data and rich context. This may include text from scientific publications [10], or text derived from a large corpus of natural language. In such scenarios, modeling user interests as unseen latent vectors, such as Latent Semantic Analysis (LSA) [11] and Latent Dirichlet Allocation (LDA) [12] have been shown to provide good results.

More recently, there has been a lot of research on topic modeling for users in social networks. User generated tags have been used to model user interests [13]. In [14], a topic of interest is described by a cluster of frequently co-occurred tags. Zheng et. al. [15] present a model to infer user topics from Weibo. Guy et. al. [16] study interests and expertise in the context of enterprise social media users.

Twitter, in particular, has been the focus of many studies that aim to characterize topical interests for users. In [2], the author-topic model is proposed, and [3] describes an empirical study of the problem. [4] leverages a knowledge base to find entities in tweets, and a labeled LDA approach is presented in [5]. Twitter has also been studied as a platform for conversation between users [17, 18].

The problem we tackle here differs from the above work in three major aspects. First, in the context of short form social media messages, latent variable techniques such as LDA and LSA have a poorer performance as compared to using scientific publications or long-form text as the source. In some cases these techniques may identify topics for some users who have enough aggregated text, but they fail to do so for passive users who may not generate a lot of text themselves. Thus they cannot provide a scalable solution when identifying topics for millions of users. Second, while previous work has focused on single social networks for topic inference, as far as we are aware, this is the first attempt to incorporate multiple social profiles to form a single unique topic profile for a user. The context under which a single user creates or reacts to different messages in any given network is significantly different compared to the context in other networks. Finally, we specifically tackle the issue of identifying socially recognizable topics for a user, since this can have unique and interesting applications.

2.1 Problem Statement

At Klout, topics are represented as entries in an ontology tree, $\mathcal{T}$. The in-house ontology is manually curated and is bootstrapped using Freebase [19] and Wikipedia Concepts [20]. The ontology provides an explicit specification of topics and relationships among them and has a hierarchical tree structure as shown in Figure 1. It has three levels: super, sub and entity. The lowest level contains specific entities, including people, things and places and are regularly updated. We currently have close to 9,000 entities and includes proper nouns, popular terms in social media, and specific concepts. The sub level contains 700 sub-topics that are abstracted concepts and each corresponds to a cluster of entities. The super level is the top level abstraction and contains 15 super topics.

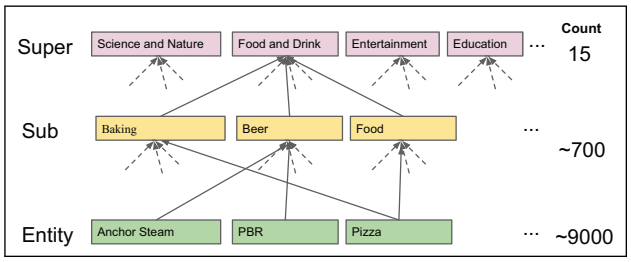

Figure 1: Hierarchical Ontology Overview

LASTA tokenizes text to generate n-grams and maps these against an internal dictionary to generate bags of phrases, $BP_u = \{BP_u^1, BP_u^2, ..., BP_u^s\}$. A user is associated with multiple bags-of-phrases where each bag-of-phrases is derived from a specific collected data source. Each of these bags-of-phrases is then mapped to a bag-of-topics, $BT_u = \{BT_u^1, BT_u^2, ..., BT_u^s\}$, from the above ontology. The mapping from a phrase to a topic ($p_j \rightarrow t_i$) is based on exact

[1] `http://wiki.apache.org/hadoop/`
[2] `http://hive.apache.org/`

match and rule-based synonym checks. Section 3.2.1 provides more details on how topics are extracted. For each user, we now obtain bags-of-topics where the k^{th} bag-of-topics is represented as follows.

$$BT_u^k = \{t_i : (p_j \rightarrow t_i),\ \forall\ p_j \in BP_u^k\}$$

Since multiple phrases may map to a topic, the strength of each topic in BT_u^k is the sum of the occurrences of corresponding phrases.

$$count(t_i|BT_u^k) = \sum_{p_j \rightarrow t_i} count(p_j|BP_u^k)$$

The strength for each topic in BT_u^k is normalized using min-max normalization.

$$s(t_i|BT_u^k) = \frac{count(t_i|BT_u^k)}{\max_{t_j \in BT_u^k} count(t_j|BT_u^k)}$$

The problem for LASTA can be stated formally as follows: Given a set of N users, $U = \{u_1, u_2, ..., u_N\}$ and a specific user u, we wish to generate a final bag-of-topics, $T_u = \{t_1, t_2, ..., t_m\}$ where $t_i \in \mathcal{T}$, and the strength of each topic represents the user's interest towards that topic. We discuss details on how we aggregate from BT_u to the final bag-of-topics, T_u in Section 4.

2.2 Data Landscape

Klout has millions of registered users. A registered user has to connect either their Facebook or Twitter account to create an account on Klout. After that, the user may connect other social network profiles, e. g. LinkedIn, Google Plus, Instagram, etc.

One of the primary challenges faced by LASTA is the size of text created by each user to infer correctly the topical interests. We present data in Table 1 on message character counts to illustrate the challenge.

Table 1: Message sizes across networks

Percentile	FB	TW	GP	LI
0.99	564.48	140.00	1578.62	577.00
0.95	200.60	134.00	521.95	277.20
0.90	128.20	118.22	323.16	175.00
0.80	77.56	89.00	204.00	132.00
0.70	59.72	68.88	133.00	112.00
0.60	50.53	54.18	92.00	93.00
0.50	43.95	43.62	61.00	75.35

Table 2: Percentage distribution of languages across networks

FB		TW		GP		LI	
en	**67.22**	**en**	**34.88**	**en**	**74.18**	**en**	**80.52**
pt	6.05	ja	12.33	es	5.61	es	6.40
it	5.78	id	11.52	it	3.00	fr	2.74
es	5.39	es	8.92	de	2.55	nl	2.23
id	2.00	ar	4.44	pt	2.41	it	1.97
rest	13.54	rest	27.88	rest	12.24	rest	6.12

LASTA focuses on topic detection in the English language, and we use off-the-shelf language detectors[3] and phrase parsers to detect English text. Because English is used only by a limited number of users on social networks, this creates another sparsity problem for non-English speaking users. In Table 2, we present details on language distribution as observed by LASTA.

Figure 2 shows the distribution of phrases used by users on each social network, on log-log scale, with base 10. The x axis is the number of distinct phrases, which corresponds to the vocabulary size by users. The y axis shows the number of users as a function of their vocabulary size in past 90 days. The distribution approximately obeys the inverse power law, particularly on GooglePlus.

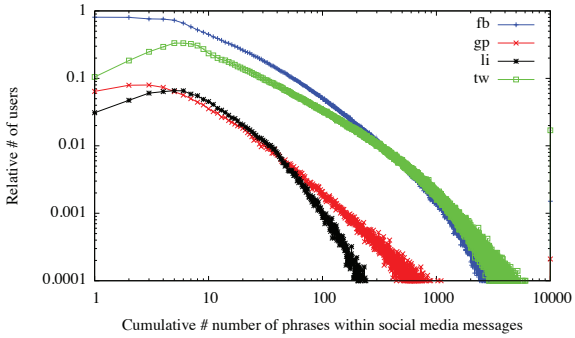

Figure 2: Registered user verbosity distribution across 90-day window

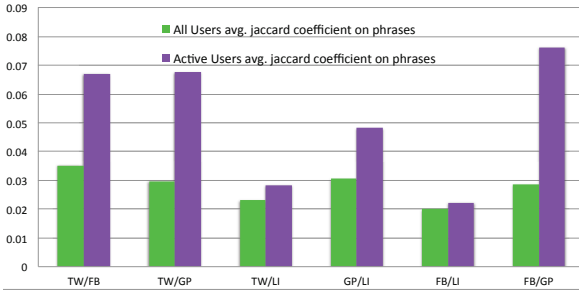

Figure 3: Phrase overlap across networks

One of LASTA's goals is to understand different behaviors presented by users in different networks. 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. We use jaccard coefficient to measure phrase overlap, $PO(u, (N_i, N_j))$ as follows.

$$PO(u, (N_i, N_j)) = \frac{|\{\text{phrase in } N_i\} \cap \{\text{phrase in } N_j\}|}{|\{\text{phrase in } N_i\} \cup \{\text{phrase in } N_j\}|}$$

where N_i, N_j are i-th and j-th social network, respectively. We then average over all users for each pair of social networks. Figure 3 shows the results. The phrase overlap value is very small on each pair; the highest overlap occurs between postings across Facebook and Google Plus and is approximately 0.075. To gain deeper insights into the overlap, we narrow down to active users only. A user is considered as active in a pair of social networks if he has generated at least 100 distinct phrases in each network in last 90 days. The overlap extent increases; however it is still small and

[3] `https://code.google.com/p/language-detection/`

less than 0.1. The highest overlap occurs between postings across TW and FB and is approximately 0.035. The low phrase overlap for a single user helps LASTA aggregate topical interests from multiple social media and produce a more complete set of user interests.

3. PIPELINE OVERVIEW

Our backend system can be broken into two main components: data collection, and data processing. When a user registers at Klout, he connects one or more social networks with his 'token', and grants permission to Klout to collect and analyze his data through the network APIs. 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. Figure 4 shows an overview of collection/processing pipeline.

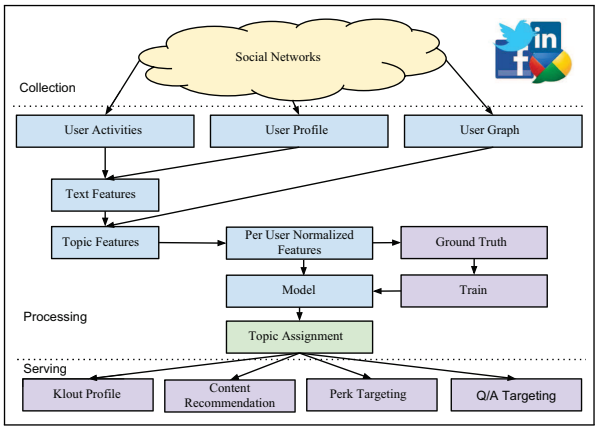

Figure 4: Data Collection and Processing Pipeline

3.1 Data Collection

LASTA focuses on collecting the following data types:

User Profile. A user may explicitly state some of his interests in his profile description on a social network. For example, the 160-character limited bio in a TW profile often contains information indicating the user's interests. On FB, users can edit their profiles to declare their interests in music, books, sports and other topics.

User Activities. Various activities on social networks provide valuable signals for topic assignment. On FB, we collect authored status updates, shared URL pages, commented and liked posts, text and tags associated with videos and pictures. On TW, we collect authored tweets, re-tweets and replies on other tweets, shared URL pages, subscribed, created and joined lists. On LI, we collect comments on posts, skills stated by the user and endorsed by connections. On GP, we collect authored messages, re-shares, comments, shared URL pages and plus-ones.

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. The social graph also contains a hidden interest graph. For instance, if a user follows "@NBA" then it is likely that he is interested in basketball. We leverage the user graph to discover the individual's interests.

For TW in particular, Klout also partners with Gnip to collect the the public data generated in the TW Mention Stream. This includes all tweets that include re-tweets, replies or a message that contains a "mention", where a user is referenced with '@' prefixed to his username. Finally, for well-known personalities we associate their Klout profile with their Wikipedia page.

3.2 Data Processing

To reiterate, the primary goal of our system is to build a comprehensive list of user interest topics at scale. The users under consideration include registered users who connect networks on Klout, and unregistered users whose public data is available via the TW stream. Overall we assign topics to hundreds of millions of unregistered users, and the number of registered users is in the order of millions.

We use the Hadoop MapReduce infrastructure to frequently bulk process the large amount of data collected. Topic inference is run daily as a bulk job, while machine learned models are built and improved in an offline manner regularly. The daily resource usage footprint of 'LASTA' for feature generation is: 55.42 CPU days, 6.66 PB reads, 2.33 PB writes, and for score generation is: 11.33 CPU days, 3.78 PB reads, 1.09 PB writes [4]. In particular we use Hive which is a warehousing solution used for querying and managing large datasets residing in distributed storage. Two of the main features of Hive are – a built-in data catalog, and SQL like syntax that gets translated to a series of MapReduce jobs at runtime. Having a data catalog makes problems tractable as the number of distinct feature types in the system grows. Performing complicated data transformations with multiple joins and secondary sorts in Hive is trivial and can be expressed as a single query, saving development time and effort which would otherwise be spent on writing multiple MapReduce steps. The Hive Query Language abstraction allows developers to mainly focus on data transformations, leading to quick prototyping and experimentation.

We also implemented independent Java utilities for entity extraction, text to bag-of-topics mapping and language detection with Hive UDF (User Defined Function) wrappers. We have open-sourced Brickhouse [5], a collection of these utility UDFs used for data aggregation and transformation.

One of the main advantages of our data processing pipeline is that new features can be easily added and removed. Having this flexibility allows us to support large number of features, some of which are network agnostic like those derived from message reactions or connection graphs, while others are more network specific like those derived from FB likes, TW lists, LI skills and so on. Overall in our production system we generate around 50 distinct types of features. We experimented with more than 100 features over the course of the project before settling on the 50 features with the most impact. In following section we will go in more detail about

[4]Uncompressed HDFS data reads/writes
[5]`https://github.com/klout/brickhouse`

bag-of-topics generation.

3.2.1 Bag of Topics Generation

Bags-of-phrases are first extracted from textual inputs, by matching against a dictionary of approximately 2 million phrases. Phrases are extracted as n-grams where n may vary from 1 to 10. The dictionary is updated daily using Freebase [19], Wikipedia Concepts [20], manual curation and top influential users' display names on Klout. 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. A larger discussion of the ontology is beyond the scope of this paper, but we note that the system is agnostic to the ontology used, and any other ontology can also be applied in this framework. We opt for exact match and rule based synonym mapping approaches here, to avoid incorrect phrase-topic associations and to minimize false positives at this step. Alternate approaches that cluster phrases to topics, or use latent variables to perform such mappings, can be addressed in future work.

The bags-of-topics thus generated have associated strengths for each topic in the bag. For most of the text based bags-of-topics we use the cumulative phrase frequency as the topic strength. For graph based bags-of-topics we use a slightly different approach, aggregating topic strengths from the user's first degree connections. Each bag-of-topics is associated with the corresponding user id, and is identified by a name representing the data from which the bag was derived. A feature vector is generated for each user-topic pair by exploding the bags-of-topics for a user, in order to formulate the problem as a binary classification problem for matching users to topics. We describe this procedure more formally in Section 4.1. The features are identified by the same name as the bag from which the topic under consideration originated. In the remainder of the paper, we will use feature names interchangeably to represent both the individual entry in a feature vector for a topic-user pair, as well as the corresponding bag-of-topics for a user.

3.2.2 Feature Generation

We use the following naming convention for feature names: `<network>_<source>_<attribution>`. Each feature is represented as a combination of three characteristics that annotate – (a) the social network in which feature originated, (b) the source data type, and (c) the attribution relation of a given feature to the user. We next provide a detailed description of these feature characteristics.

Network: The social network from which the data originated, which are abbreviated as TW, FB, GP, LI, WIKI.

Source: The feature source captures the input data source, and optionally the derivation method when the same source may be interpreted in different ways. Text and social graph based sources are the two major inputs from which features are generated.

Text based sources originate from text associated with messages, posts, profiles, lists, videos, photos, or URLs shared. In addition we also fetch shared URLs and extract text from the HTML, as well as the text from meta tags annotating the title, description and keywords of a URL. This enables LASTA to gain additional context about content with respect to a user.

User graph derived features are calculated by aggregating topical interest of a user's first degree social graph. The first degree user graph topics are bootstrapped using some individual features which have high coverage and precision, for example TW Lists. Since topics are assigned daily, subsequent graph features are generated using topic assignments from the previous day. For the graph based bags-of-topics, we associate raw strengths as:

$$s(t_i|BT_u^k) = \sum_{v \in G_u} s(t_i|BT_v^k)$$

where G_u is the social graph of the user u, and v is a first-order neighbor of u. These strengths are also normalized using min-max normalization as described previously. Examples of such graph sources include FRIENDS on FB, and FOLLOWING and FOLLOWERS on TW.

The Source may optionally also include the time window considered for generating the feature. Since users' interests on social media may vary over time, some inputs may be indicators of topical interests only temporarily, while others such as country of birth, or professional interests, may indeed be long term indicators of topics associated with a user. We therefore consider inputs in a 90 day window to capture the temporal nature of changing topical interests, and an all-time window for the more permanent inputs.

Attribution: Attribution denotes the relation of the input source to the user. It may be one of the following:

1. *Generated:* Originally generated or authored content by the user, including posts, tweets, and profiles. This also includes comments which are attributed as generated, to the person who authored the comment.
2. *Reacted:* Content generated by another user (actor), but as a reaction to content originally authored by the user under consideration. This includes comments, retweets, and replies.
3. *Credited:* In this case the user has no direct association with the content from which the feature was derived. Examples include text that is associated with the user because he was mentioned with tags, or added to lists and groups by other users.

The most obvious attribution is *Generated*, which is based on text that the user has authored himself. Traditionally, this has been the primary input used to infer topics, but in the context of social media, this may often be insufficient or inaccurate. Users typically talk about a variety of subjects casually, such as "I had a late lunch today", which does not necessarily indicate the user's interest in lunch or food. In addition, self-authored posts may cover only temporary or partial interests. For example, Bill Gates uses his Twitter account to primarily talk about topics like 'Philanthropy', 'Books', 'Malaria' and 'HIV infection'. While his work as a philanthropist is captured by textual input from tweets, it's essential that the system also assigns topics like 'Software industry' and 'Microsoft'. Thus generated inputs by users themselves may be inaccurate or insufficient to derive topical interests for users. To address these issues, we consider two other categories of text to derive topical signals.

The first is *Reacted* text, which considers messages included in comments or replies that were created by other 'actors', in reaction to an original message created by user. In this case we attribute the text of the comment or reply to the original message author and label it with the *Reacted* attribution. For some users the amount of text generated through reactions greatly exceeds the amount of original text, thus providing a lot more context and a much better signal for topic inference.

The second attribution that we consider is *Credited*. In this case the user is only indirectly involved with the signal under consideration, and neither generates, nor directly provokes the creation of the input with which he is associated. Instead, other users in the social network associate certain messages or content to the original user. Examples of such inputs are tweets in which a user is mentioned, or posts on FB where a user is tagged, or recommendations written by colleagues on LI, or a user being listed as a member of a TW list. These messages provide strong signals for topics associated with a user, because they indicate how other members of the social network perceive the user's topical interests. This attribution is important especially in the case of celebrities who may not be regular content creators themselves, but indirectly generate text via users who talk about and mention them.

The alert reader may have also noticed that the *Generated*, *Reacted* and *Credited* categories are analogous to the first person, second person and third person views used in language and grammar.

3.3 Ground Truth

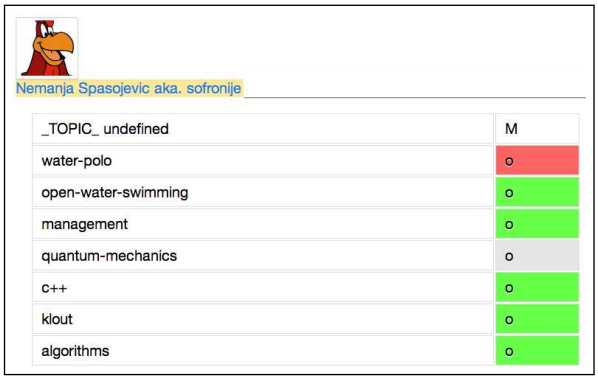

Figure 5: Ground Truth Collection Tool

In order to build models based on the features described above, we designed a simple web app to collect ground truth data with labels for user-topic interests.

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

We conducted controlled user studies where the evaluators in the experiment were registered Klout users. In this experimental setup, the system pulls up a set of the participant's user graph first degree connections who are also Klout users, or Twitter users whose data is available via Gnip stream. System randomly assigns topics to the users' first degree connections. The evaluator then gives positive or negative feedback, depending if the topic is good or bad match for his connection. If participants are uncertain about the relevance of the topic-user pair, they skip the evaluation for that pair. Table 3 shows the statistics of the dataset we collected through the user study. The screenshot of the tool is shown on Figure 5.

Our ground truth data is aimed at generating labels for socially recognizable user topics. A participant does not evaluate himself to ensure that personal biases are separated from the feedback. In our dataset, we found that out of all pairs of user-topic pairs that received more than one vote, only 27% have conflicting feedback. The conflicting votes contribute to only 2.2% of all the votes that were collected, suggesting that in most cases the association is clear.

4. ANALYSIS AND EVALUATION

In this section we describe our approach to solve the problem of predicting topics for a user, using supervised learning. We designed experiments and collected ground truth from real users, to acquire labeled data for matching users to topic. This labeled data is then used for training and evaluation, and the results are presented below.

4.1 Feature Analysis

As explained previously, multiple bags-of-topics are derived from different sources for each user. We explode these bags-of-topics, and for each topic-user pair (t_i, u), we build a feature vector $x_{i,u}$. The value of each feature in the vector is the topic strength of t_i given the bag-of-topics, BT_u^k:

$$x_{ik} = s(t_i|BT_u^k),$$

where BT_u^k is the k^{th} bag-of-topics for the user. We name the k^{th} feature with the same name as the bag BT_u^k. One of the primary contributions of this study is to analyze which features are indicative of a user's topical interests on social networks.

We find that textual input authored by users themselves accounts for at least one topic for only 58% of users on the labeled set. The remaining users either do not create enough text, or generate text that is not necessarily indicative of their topical interests. For such users we include reacted and credited signals in order to predict their topics, as described in the previous section.

We evaluate the performance of the topic prediction through traditional IR metrics:

Precision(P) measures the fraction of retrieved topics that are relevant to the user.

$$P = \frac{|\{\text{relevant topics}\} \cap \{\text{retrieved topics}\}|}{|\{\text{retrieved topics}\}|}$$

Recall(R) measures the fraction of relevant topics that are retrieved.

$$R = \frac{|\{\text{relevant topics}\} \cap \{\text{retrieved topics}\}|}{|\{\text{relevant topics}\}|}$$

Table 4 shows a selected list of important features along with their Precision (**P**) and Recall (**R**) values as evaluated on the labeled set. In this case, the predicted topics for a

user are the bag-of-topics associated with the feature. We also present the coverage (**C**) in terms of percentage of registered users who have the feature.

We notice that credited List based features on Twitter and generated LinkedIn features have the highest individual predictive quality in terms of precision. Generated URL features typically have higher recall than other features, suggesting that shared URLs are a strong signal of a users topical interests. We also find that the graph based features have the highest coverage and recall values, which highlights why these features can predict topics for users who are not very active themselves.

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

4.2 Training and Evaluation

Given the bags-of-topics generated for users, our objective is to accurately predict the topic preference for each user. Feature vectors are generated from exploded bags-of-topics for user-topic pairs as described above. When a certain topic occurs in multiple bags for a user, then the feature vector for that pair will include all these values x_j, and 0.0 values for features where it does not occur.

We now 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 experimented with several classification algorithms, including those reported to achieve good performance with text classification tasks, such as support vector machines, logistic classifiers, and stochastic gradient boosted trees. We found that the best and most stable performance among the techniques we tested was obtained with the logistic classifier. We predict the label by $\hat{y} = P(y|t_i, u) = \sigma(x_{i,u} \cdot \theta)$, where $\sigma(a) = \frac{1}{1+e^{-a}}$ is the sigmoid function. The label $y \in \{0, 1\}$ assigns 1 if the topic t_i is of interest to the user u, 0 otherwise.

We train our models using the feature vectors generated for the pairs against the labels from the labeled data. The final model applies weights w_k to get the final bag-of-topics, T_u. The topic strength for a specific topic $t_i \in T_u$ is:

$$s(t_i|u) = \sum_{BT_u^k \in BT_u} w_k s(t_i|BT_u^k)$$

4.3 Results

4.3.1 Classification Prediction Results

In addition to precision and recall, we also use the F1 Score, $F1 = \frac{2PR}{P+R}$ to measure performance as a tradeoff between precision and recall.

Table 5 presents the performance of topic prediction using k-fold cross validation on the labeled set, where $k = 10$ and the held out set is 20% of the data. Class 1 represents positive instances where the topic was correctly predicted, and class 0 represents negative ones, where the topic was correctly discarded. We consider the predictive power of different feature sets, and how they compare to the case when the full feature set is used. The "Feature Set" column indicates the feature subset used for the prediction.

We discuss insights gained by comparing the performance of using all features versus using only subsets of features:

Single Network Comparison: The precision when all features are used is higher than when we use only features from a single network like Twitter. This shows that increasing the information available for a user by using the user's presence on other networks improves the correctness of the predicted topics in both classes. While using features from only Facebook may yield a higher precision, the recall in this case is very low, and we are able to predict fewer topics for each user. These observations together imply that because of the nature of any given social network, a user may not reveal all his interests on any single network alone, making it necessary to use features from multiple networks.

Attribution Comparison: We also compare the performance when we use only features derived from user generated input, which includes text as well as shared URLs (GEN.), or use only features from the user's reacted and credited inputs (REAC. + CRED.). The generated set of features yield a high precision, but a low recall value. The reacted and credited features give a slightly lower precision, but slightly higher recall compared to the generated input. But using all inputs together yields a much higher recall value than using them separately. This shows that using only user generated text can predict much fewer topics for the user, as compared to using the generated, reacted and credited inputs together.

Graph Comparison: Finally we also compare how the graph based features (GRAPH) play a role in topic prediction. Excluding graph based features gives a higher precision but a

low recall value, and using only graph features provides a much higher recall value, with a slightly lower precision. This highlights the value of using graph features, because by the nature of the social networks, it is possible to predict topics for a user by considering the topics of the other users that he is connected to. But relying solely on graph based features gives some incorrect predictions, because of the possible noise introduced.

Thus we observe that using the complete set of features maintains a relatively high precision, while greatly improving the recall. The results show that including multiple networks, generated text input, reacted and credited signals, and graph based features together gives the best performance overall, as indicated by the F1-score in Table 5. LASTA also achieves a 92% precision k, where $k = 10$, on the full training set.

Table 5: Binary classification prediction for different feature sets

Feature Set	Class	P	R	F1	F1 Avg.
TW	1	0.703	0.484	0.573	0.613
	0	0.572	0.771	0.657	
FB	1	0.792	0.298	0.433	0.548
	0	0.538	0.912	0.677	
GEN.	1	0.799	0.335	0.472	0.568
	0	0.543	0.904	0.678	
REAC. + CRED.	1	0.733	0.373	0.495	0.572
	0	0.541	0.845	0.660	
ALL − GRAPH	1	0.792	0.411	0.541	0.610
	0	0.599	0.809	0.688	
GRAPH	1	0.739	0.501	0.597	0.633
	0	0.599	0.809	0.688	
LASTA	1	0.758	0.526	0.621	0.652
	0	0.599	0.809	0.688	

4.3.2 Curation Evaluation

We deployed LASTA in production at Klout, and displayed the top 10 predicted topics in ranked order on each user's profile. Users could then add, delete, or reorder the list, indicating agreement or disagreement with the predicted list. We also evaluate our system against this self-curated user data. We select the set of users who have made changes on their topic profiles, and evaluate the initially predicted list of topics against the final curated list for each user. Table 6 has the statistics of this dataset.

Table 6: Statistics of the curated dataset

Statistics	Value
# of users	19,505
# of (user, topic) pairs	196,481
Avg # of positive topics per user	7.37

We evaluate LASTA using the following metrics on the curated data:

Mean Average Precision (MAP) For a single user, average precision calculates the average of the precision of the top K topics. $AP@K = \frac{\sum_{i=1}^{K} P@i}{K^+}$, where K^+ is the number of positive samples. Here $P@i$ is the precision at cut-off i in the retrieved list. The mean average precision for N users at position K is the mean of the average precision for each user, i.e., $MAP@K = \frac{1}{N}\sum_{i=1}^{N} AP@K(i)$.

Normalized discounted cumulative gain (nDCG) Measures graded relevance of the list of topics, i.e., $DCG = \sum_{i}^{k} \frac{2^{r_i}-1}{\log_2(p_i+1)}$ where $r_i = 1$ if the topic has a positive label in the curated list, and p_i is the position of topic in the ranked list. Normalized DCG is the ratio of DCG by the model's ranking to the DCG by the ideal ranking: $nDCG = \frac{DCG}{IDCG}$.

We use these metrics to compare the output of LASTA against other approaches. In particular, we compare LASTA to approaches where the topics for a user are predicted using aggregated topic frequency (TF) from subsets of features. These subsets are those derived from generated textual input only; all generated inputs including URLs shared, LinkedIn Skills etc.; and all inputs that were generated, reacted and credited. Table 7 shows the results for ranking the top K topics of interest for each user, where $K = 10$.

Table 7: Ranking performance comparison on user curated data

Model with Features used	MAP@K	nDCG
TF - Message Text Generated	0.150	0.140
TF - All Generated	0.155	0.139
TF - All	0.160	0.141
LASTA - All	0.314	0.269

Note that the users who curate their own data are only a small fraction of users on Klout, who are self motivated to edit their topic list. Since most users do not edit their list, either because they are satisfied with it, or because they are not motivated enough to change it, we exclude such users from the dataset. On this dataset, LASTA significantly outperforms the other approaches in terms of both the MAP and nDCG metrics, showing that it does indeed produce a better set of ranked topics for a given user.

Table 8 shows some examples of topics assigned to some well known personalities according to LASTA.

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

4.4 Cross Network Analysis

In our dataset, around 13% users connect to a single social networks, 40% of users to two social networks, and less than 10% users connect all four social networks. Typically it is expected that a user does not connect all four networks, since most users are only active in one or two networks. But the advantage of using four networks is that the fraction of users using at least two out of the four is higher, leading

to more information about the user. The details of user behavior patterns in a cross-network setting is beyond the scope of the paper, but here we present some interesting topical insights across networks.

4.4.1 Super-topics comparison

As discussed previously in Section 2.2, we observe that phrases used by a user have low overlap across social networks. Here we examine similarities and differences between topical interests aggregated across users on different networks. To aid visualization, we roll up entities and subtopics to super-topics, reducing the topic dimension space from 10,000 to 15. We sum up the presence of user interests rolled up to super-topics in each individual social network, and plot this distribution. Table 9 shows the percentage breakdown of super-topics on each social network for the users on that network, and also the breakdown across all users according to LASTA.

We observe from the figure that users in each network have distinct topical interests. On FB and TW the super-topic "entertainment" is the most represented one, whereas "business" is the most represented super-topic on LI, and "technology" on GP. FB users are also more interested in topics related to "lifestyle" and "food-and-drink" compared to users on other networks, while a significant number of GP users show interest in "arts-and-humanities". For LI, apart from "technology" and "business", other topics are not highly represented, which is expected since it is a professional networking platform. The left-most column shows the distribution of topics as assigned by LASTA. The "business" row is an interesting one to observe. While this topic is not highly represented on TW, FB and GP, LASTA is able to assign "business" related topics to users, because it also takes into account signals from LI. This shows that using multiple networks can lead to not only a deeper understanding for each user, but also a better understanding across topics.

4.4.2 Topics distribution

While the previous section analyzes cross-network topic distributions qualitatively in terms of super-topics, here we examine the distribution quantitatively in terms of number of topics assigned to users. While assigning a very large number of topics is not necessarily the goal of the system, we analyze these distributions in order to perform cross-network comparison. In Figure 6, each plotted point represents the fraction of users who have at least x number of topics assigned to them.

We find that the number of topics assigned to users with TW and FB is much larger than that assigned using GP or LI. This is because GP and LI do not provide API access to graph data, and also have a smaller volume of textual input compared to TW and FB. We conclude from the graph that 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.

5. APPLICATIONS

LASTA is serving multiple personalized services in Klout. We briefly describe some applications next.

Targeting: Given that social media is a modern means of spreading awareness among people, many brands desire to target promotional messages and campaigns to social network users. As an example, a car company that wants to spread awareness about a new car model, may want to target certain incentives or "perks" related to the car to some users on social media. When users interested in cars are targeted with the perk, they may be motivated to talk about the car on their respective social networks, effectively generating word-of-mouth awareness about the new model. This approach of targeting users based on topics, that can provide value to companies and brands, has been successfully implemented at Klout with LASTA.

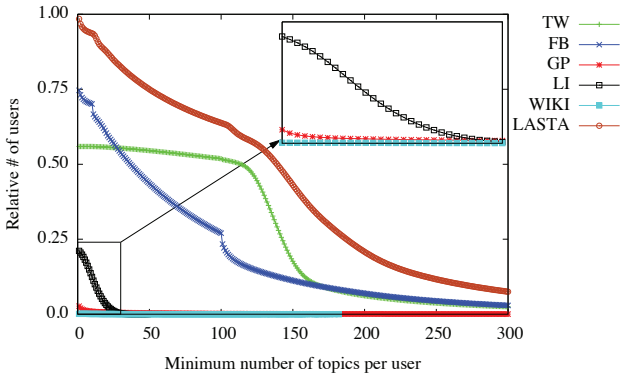

Figure 6: Distribution of registered users for minimum number of topics assigned across different networks.

Content Discovery: The topics deduced by LASTA provide utility to users in terms of serendipitous content discovery at Klout. This system aggregates online articles, categorized by topic, and ranks them based on relevancy to a user. The system can also identify topics that some members from the user's social graph may be interested in. A user can then be shown a customized feed of articles that he may either want to discover and read about himself, or may want to share with a wider audience on his social networks.

Question Answering: In a question answering scenario, a user in the system can ask a question pertaining to a certain topic, which can then be routed to specific users who may be able to answer the question. For example, a question such as "What is the best place to go fishing near San Francisco?", may be routed to users interested in fishing who live in San Francisco. Users to whom questions are routed are able to give credible answers to such questions, and the original asker may get multiple good answers. This system was implemented and used by users on Klout, and was again enabled by LASTA.

6. CONCLUSION AND FUTURE WORK

In this paper, we presented Klout's topic assignment system, LASTA. The system assigns overs 10,000 topics to hundreds of millions of users spread across multiple social networks on a daily basis with a high accuracy. We hope the engineering architecture and technology choice provides insights to build scalable and extendable topic mining systems.

LASTA provides the foundation to build other data-driven products that utilizes the user-topic relationships. In future, we want to build systems that will understand topical expertise among the millions of users.

7. ACKNOWLEDGMENTS

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

We thank Sarah Ellinger for her continuous work on curating the topic ontology. We are extremely grateful to our former colleagues Yize Li, Jerome Banks, Guangle Fan and Ding Zhou who contributed towards building LASTA. We thank Joe Fernandez for his valuable feedback throughout the project and also thank Prof. S. Muthu Muthukrishnan of Rutgers University for his inputs.

8. REFERENCES

[1] Alan Ritter, Mausam Clark, Sam, and Oren Etzioni. Named entity recognition in tweets: an experimental study. In *Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2011.

[2] Zhiheng Xu, Rong Lu, Liang Xiang, and Qing Yang. Discovering user interest on twitter with a modified author-topic model. In *Web Intelligence and Intelligent Agent Technology*, 2011.

[3] Liangjie Hong and Brian D Davison. Empirical study of topic modeling in twitter. In *Workshop on Social Media Analytics*, 2010.

[4] Matthew Michelson and Sofus A Macskassy. Discovering users' topics of interest on twitter: a first look. In *Workshop on Analytics for noisy unstructured text data*, 2010.

[5] Daniel Ramage, Susan T Dumais, and Daniel J Liebling. Characterizing microblogs with topic models. In *Weblogs and Social Media*, 2010.

[6] Yehuda Koren, Robert Bell, and Chris Volinsky. Matrix factorization techniques for recommender systems. In *Computer*. IEEE, 2009.

[7] Susan Gauch, Jason Chaffee, and Alexander Pretschner. Ontology-based personalized search and browsing. In *Web Intelligence and Agent Systems*, 2003.

[8] Feng Qiu and Junghoo Cho. Automatic identification of user interest for personalized search. In *World Wide Web*, 2006.

[9] Amr Ahmed, Yucheng Low, Mohamed Aly, Vanja Josifovski, and Alexander J Smola. Scalable distributed inference of dynamic user interests for behavioral targeting. In *Knowledge discovery and data mining*, 2011.

[10] David M Blei and John D Lafferty. A correlated topic model of science. *Annals of Applied Statistics*, 2007.

[11] Thomas K Landauer and Susan T Dumais. A solution to plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological review*, 1997.

[12] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 2003.

[13] Julia Stoyanovich, Sihem Amer-Yahia, Cameron Marlow, and Cong Yu. Leveraging tagging to model user interests in del. icio. us. In *AAAI Spring Symposium: Social Information Processing*, 2008.

[14] Xin Li, Lei Guo, and Yihong Eric Zhao. Tag-based social interest discovery. In *World Wide Web*, 2008.

[15] Zheng Yang, Jingfang Xu, and Xing Li. Data selection for user topic model in twitter-like service. In *Parallel and Distributed Systems*, 2011.

[16] Ido Guy, Uri Avraham, David Carmel, Sigalit Ur, Michal Jacovi, and Inbal Ronen. Mining expertise and interests from social media. In *World Wide Web*, 2013.

[17] C Honey and Susan C Herring. Beyond microblogging: Conversation and collaboration via twitter. In *System Sciences*, 2009.

[18] Danah Boyd, Scott Golder, and Gilad Lotan. Tweet, tweet, retweet: Conversational aspects of retweeting on twitter. In *System Sciences*, 2010.

[19] Google. Freebase data dumps. `https://developers.google.com/freebase/data`, 2012.

[20] Valentin I. Spitkovsky and Angel X. Chang. A cross-lingual dictionary for english wikipedia concepts. In *Language Resources and Evaluation*, 2012.