Characterizing Latent User Interests on Enterprise Networks

Ben Priest and **Kevin M. Carter**
MIT Lincoln Laboratory
244 Wood Street
Lexington, MA 02420, USA
{benjamin.priest, kevin.carter}@ll.mit.edu

Abstract

We present a methodology for promoting situational awareness of an enterprise network using only network artifacts discernible from network protocol logs. We utilized latent Dirichlet allocation (LDA) over two corpora, the first composed of search queries and the second composed of external domain names issued by enterprise users through the network proxy of a mid-sized enterprise network. We not only quantitatively demonstrate that the LDA topic modeling approach achieves superior fit to the data than do naïve models, but we also demonstrate that a topic modeling approach yields qualitative business analytic information.

Introduction

As an enterprise becomes more diversified, stakeholders find the tasks of characterizing the body of work taking place on the network increasingly difficult. Although enterprise stakeholders have access to categorical organizational data separating users into the internal hierarchy of the enterprise, we argue that such data proves insufficient to the task of characterizing the actual behavior of users. An understanding of how users utilize both internal and external network resources can inform network operators of how to best perform load-balancing and characterize the enterprise's body of work.

While the ground-truth usage behaviors that users execute on the enterprise network may be difficult or impossible to recover without the deployment of expensive and invasive surveillance tools, usage of the network leaves behind indelible traces that can be used in an attempt to reconstruct said behaviors. Some of these traces are left behind in network protocol logs. These protocol logs, such as those associated with the web proxy, DNS, domain controller, et cetera, contain references to objects with semantic content, such as IP addresses, port numbers, human-readable text, MAC addresses, website domains, user IDs, and so on. We will heretofore refer to these objects as network artifacts.

By mining protocol records and correlating network artifacts with a user ID, it is possible to reconstruct a trace of a user's usage of the network. However, this trace alone provides an incomplete picture of the behavior of both the individual and the group, and it is unclear how one could leverage such models to gain business intelligence into the inner workings of the enterprise.

We argue that a latent topic model, such as latent Dirichlet allocation (LDA) (Blei et al. 2003) is an appropriate means by which to characterize both group and individual behavior on an enterprise network. In this paper, we assume a generative topic model over human users, whereby users draw from a distribution of topics of interest, which are in turn distributions over network artifacts. For the purposes of this paper, we shall restrict ourselves to artifacts in the form of search queries issued to external search engines and visits to external web domains, due to their human readability.

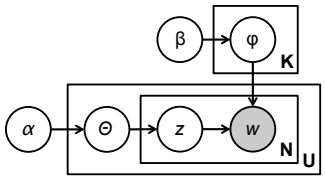
Treating the records of individual activity as documents in a topic model gives us the advantage that we may succinctly represent users as vectors in the topic space, which will be of much lower dimensionality than the space over all possible activities.

Related work

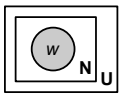
Traditional topic modeling models human-crafted documents as entities that possess topic distributions over natural language words (Blei et al. 2003). We model users as entities with topic distributions over network artifacts, effectively treating the history of activity for a user as a document. Particular instantiations of network artifacts, to which we will heretofore refer as "tokens", fill the role of words in these documents.

The novelty of this work lies in modeling enterprise users' network behavior through methods typically reserved for modeling text corpora and images (Blei et al. 2003) (Russell et al. 2006) (Sivic et al. 2005). The ability to obtain accurate models in this fashion is non-obvious, and to our knowledge has yet to be presented.

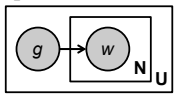
In our latent topic model we assume that a "document" consists of the bag of network artifact tokens attributed by a particular user during the collection period. This assumption distinguishes our work from prior work modeling network behavior using Bayesian topic models (Cramer and Carin 2011) by focusing on particular artifacts instead of the protocol in use. There have been prior works using latent topic models over search queries, particularly for the purposes of query clustering. However, these works treat each of these queries as individual observations (Peng, Wang, and Sun



(a) The standard LDA plate model



(b) The unigram baseline



(c) The unigram baseline conditioned on organization hierarchy

Figure 1: We considered a number of Bayesian graphical models as candidates for the generative process underlying the network artifact observations.

2012) (Aiello et al. 2011) or as individual documents themselves (Song et al. 2010). We are, on the contrary, concerned with the totality of a users' words and short phrases within queries, not necessarily the individual queries themselves.

We do not consider in this paper other latent topic models, as we are not concerned with advocating LDA specifically; we are instead interested in demonstrating the viability of a latent topic modeling approach to the problem at hand.

Latent Dirichlet allocation

Latent Dirichlet allocation, first proposed in (Blei et al. 2003), is a generative Bayesian topic model where observations are conditioned on latent hidden variables with Dirichlet priors. LDA is widely used in the literature for topic modeling and document classification.

Assume that there are K topics, W unique artifact tokens and U users, where each user u is associated with a bag of N_u observations (tokens). Furthermore, each user u has a multinomial distribution θ_u over the K topics, and each topic z has a multinomial distribution ϕ_z over the W artifacts. These distributions have a Dirichlet prior with fixed hyperparameters α, β, respectively. A plate diagram for this model is shown in Figure 1a. For each token $w_n^{(u)}$ from user u, LDA asserts that a topic $z_n^{(u)}$ is drawn from u's topic distribution θ_n, and that $w_n^{(u)}$ is in turn sampled from $z_n^{(u)}$'s artifact distribution. Training an LDA model to data entails estimating both the user-topic $\mathbf{\Theta}$ and topic-artifact distributions from the observed users, given the fixed hyperparameters of the Dirichlet priors and the number of topics K. Exact inference is intractable, so we perform inference utilizing Gibbs sampling (Jordan 1999).

Documents may be represented by their learned K-dimensional topic vector. In practice, this vector is much smaller and more manageable than representing documents in the original word space, which is very high-dimensional.

Related generative models

The data considered in this paper, being unordered bags of network artifacts, presupposes treatment using a Bayesian graphical model. Considering other information, such as timing information, and training an appropriate generative model would be comparing apples to oranges. We will consider simple baseline models that do not presuppose complex topic machinery to serve as a counterpoint to LDA.

First, we consider a simple unigram baseline, where we assume that the words from each document are drawn from a single multinomial distribution. The plate diagram for the model is given in Figure 1b. The unigram model makes the generative assumption that individual users draw from the same distribution over tokens.

We also consider the possibility that users may behave differently based upon the organizational structure of the enterprise. To test this hypothesis, we obtain an organization mapping of users to groups, and train a unigram model for each group within the enterprise. The plate diagram for this model is given in Figure 1c. Here, g is the label of the organizational group to which the user is assigned. The generative assumption in this case is that organizational hierarchy completely determines individual behavior.

Demonstrating superior model fit using LDA over these baselines implies that latent topic models are not only an appropriate method of characterizing human network behavior but also expose behavioral patterns that are opaque to the organizational structure of the enterprise.

Tokenization and artifact discovery

We capture user history for a set period of time by mining user HTTP request lines. An example of such would be the HTTP request

``GET http://www.google.com/search?q=**supreme**+**court**..."

in which the underline portion of the URL is the collected domain, the portion tagged by "q=" represents the query terms, and the rest of the request line is omitted for space reasons.

After the collection stage, the dataset consists of collections of artifacts (either domains or search queries). We assign these artifacts to users by cross-referencing the IP address and time of the source proxy record with other network protocol records. Finally, we define the set W of acceptable tokens in the corpus by collecting all unique tokens from the user base and performing further filtering, depending upon the type of artifact being collected, as defined below.

At the end of the tokenization stage, each user document consists of a bag of tokens, consisting of all of the tokens (de-noised) mined from a user's search history.

Web domain filtering

Analysis of the web domain dataset reveals that content delivery networks (CDNs) (Hofmann and Beaumont 2005) such as Akamai (Nygren, Sitaraman, and Sun 2010) and ad servers are often among the most frequently requested domains on the internet, a fact that is not surprising given their rampant deployment. On average, more than half of the top ten weighted tokens in discovered topics were CDNs and ad networks when we ran LDA over an unfiltered artifact

```
input : u, a list of a user's search queries;
W, the token vocabulary;
N, the maximum phrase length;
output: bag of words u* composed of elements of W
for query ∈ u do
    for i ← max(N, Length(query)) to 1 do
        candidateTokens ←
        GetSubstringsOfLength(query, i);
        for token ∈ candidateTokens do
            if token ∈ W and token is not a strict textual
            substring of a member of AcceptedTokens then
                AddToken(u*, token);
                AddToken(AcceptedTokens, token);
            end
        end
    end
    Clear (AcceptedTokens);
end
return u*;
```

Algorithm 1: Query Tokenization

vocabulary W. However, such domains, particularly ad networks, are usually not informative concerning human behavior. The same ad network may serve arbitrarily many content sites over time, which have no guarantee of being related, conceptually or otherwise.

We get around this problem by utilizing a whitelist filter. We removed from the artifact vocabulary W domains that are used by more than half of the user documents or do not appear on the list of top million visited websites as determined by Alexa ("www.alexa.com"). In practice, this step filtered most of the ad servers from the vocabulary, although many prominent CDNs are still present on the list.

We further reduced the vocabulary by shortening domains to their second level, e.g. we stored "`www.google.com`" as "`google.com`". Thus, all of the domains used in this paper are in fact second-level domains.

Search query filtering

Many latent topic models assume that words are generated independently from one another and hence lose the short-range lexical significance of these observations (Blei et al. 2003). The same significance is applied to coocurrences of the individual words within the same document, discounting the textual distance separating them. This assumption is often invalid, as multi-word phrases can encapsulate different semantic content than their individual words. Hence, we consider multi-word phrases or "n-grams" to also be network artifacts.

Some latent topic models (Griffiths et al. 2005) (Griffiths, Tenenbaum, and Steyvers 2007) (Wallach 2005) modify or eliminate the bag-of-words assumption to account for multiword phrases modeling short-range word dependencies. However, search query word choice is governed by utility and the technical savvy of the end user, leading to little structure outside of the phrasing of atomic concepts, such as "supreme court". Hence, we use a novel heuristic approach to identifying multiword phrases by processing the data before aggregation into documents.

When dealing with a corpus of search queries, we select the vocabulary W by counting all strings of concurrent words of size 1 to n within each query for each user, for some integer n. We then select the n-grams that occur at least m times across the corpus, for some threshold m. We remove from W unigram tokens that are composed of stop words or are shorter than 3 characters.

Filtering the user documents is not as simple as deleting tokens not in W, however, since it is not clear where the token boundaries occur, as n-gram tokens may encapsulate smaller tokens. We filter user bags following Algorithm 1, which hierarchically parses queries for n-grams, attributing them as tokens to the user only when they are both members of W and not a strict substring of a previously accepted token from the same query. This is a greedy left-to-right approach that is biased toward attributing longer tokens, and avoids counting the same textual word multiple times. The algorithm runs in $O(QN^3)$ time, where Q is the total number of search query tokens, and N is the maximum phrase length. In practice, N will be very small, so the algorithm effectively runs more efficiently than training LDA.

Experiment

We implement LDA using the Matlab Topic Toolbox (Griffiths and Steyvers 2004) on commodity hardware. We set the hyperparameters $\beta = 200/W \approx 0.05$ and $\alpha = 0.25$, which biases the method towards sparse topics.

We collected web proxy logs aggregating the activity of 3715 users of a mid-sized enterprise network from September 1, 2011 to May 15, 2012, anonymizing the logs to protect user privacy. This enterprise is composed of several divisions that focus on technical research and development in various fields, such as Aerospace and Air Traffic Control, and divisions that provide business support, such as Human Resources, Travel, and Information Technology. We composed documents for each user using the procedure described above to derive a dictionary of 11146 web domain tokens and 34621 search query tokens. We then eliminated the users with zero token assignments from our analysis, resulting in corpora of 3519 and 3490 users, respectively.

Figure 2 illustrates the cumulative distribution functions (CDFs) of the number of both the total number of domain and search query tokens used by individual users, as well as the CDFs for the number of unique domain and search query tokens used by individuals. The distribution indicates that active users issue a far greater number of domains than search query tokens, while also indicating that users repeat search query tokens very infrequently. Meanwhile, users revisit domains with great regularity, which makes sense considering that we are operating at the level of second level domains. Users are likely to frequent the same news sites, such as `cnn.com`, as well as technical sites such as `stackoverflow.com`.

The distribution of n-gram tokens, for $n = 1, 2, 3$, is displayed in Table 1. The numbers here indicate that, as expected, there are relatively few multiword phrases that make it through the tokenization process.

We take a basic quantitative approach at measuring model fit by computing the model perplexity of the trained models

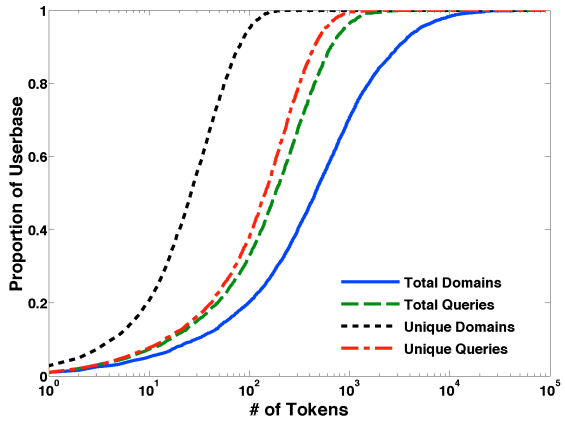


Figure 2: CDFs of the number of unique queries issued by individual users.

$n =$	1	2	3
# of Tokens	22365	11079	1177

Table 1: Number of n-gram tokens in the final dictionary

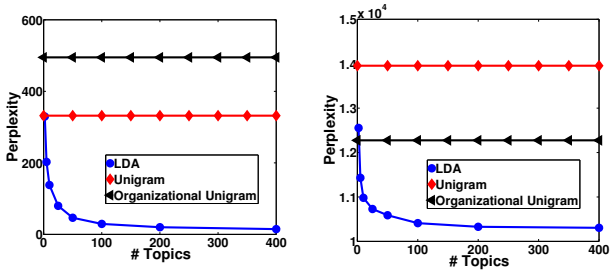
against a held out test set (Blei et al. 2003).

$$perplexity = \exp\left\{-\frac{\sum_{u=1}^{U}\log p(\mathbf{w}^{(u)})}{\sum_{u=1}^{U} N_u}\right\}, \qquad (1)$$

where $p(\mathbf{w}^{(u)})$ is the likelihood of observing the collection of N_u tokens associated with user u. In each case, we performed 10-fold cross validation to test model fit. We utilized Laplace smoothing to account for out-of-vocabulary words encountered in the test set.

The domain name and search query perplexity results for both of the unigram models described in above, as well as the LDA model computed over varying number of topics, are shown in Figures 3a and 3b, respectively. Note that the LDA model perplexity is vastly superior to both unigram models in each case. This result is consistent with previous approaches and is consistent with our assertion that a body of search queries is representable as a topic-model document.

Over search queries in Figure 3b, the organizational unigram model performs better than the flat unigram model, suggesting that enterprise organization does influence user search behavior. However, the flat unigram model outperforms the organizational unigram over web domains in Figure 3a. Figure 2 explains this phenomenon. It turns out that the user distributions over the vocabulary are much sparser for the web domain dataset than the query dataset. Thus, the likelihood of out-of-vocabulary words occurring in the test set is much higher when the training is done over a subset of the total training set. This leads to the increased perplexity observed in the organizational unigram, which trains the distribution for each organization group based solely upon the behavior of members of that group.



(a) Domain name perplexity (b) Search query perplexity

Figure 3: Model perplexities computed over a varying number of topics

Web Domains

Development: sourceforge.net, stackoverflow.com, netbeans.org, github.com, uml.edu, getfirebug.com, java.com, eclipse.org
Android: phandroid.com, androidforums.com, pelican.com, wikitravel.com, hamptonroads.com, androidandme.com, androidcommunity.com, rpi.edu
Webcomics: penny-arcade.com, trenchescomic.com, thehairpin.com, megatokyo.com, questionablecontent.net, xkcd.com, pvponline.com, blip.tv
News: dailycaller.com, washingtonexaminer.com, freebeacon,com, rt.com, frontpagemag.com, imgfarm,com, rpxnow.com, gravatar.com
Physics: spiedigitallibrary.org, spie.org, jezebel.com, arxiv.org, optics.org, physicsworld.com; phys.org; aapt.org

Search Queries

Unix: linux, centos, perl, fedora, command, solaris, redhat, rpm
Mobiles: android, galaxy nexus, verizon, samsung, droid, google, mobile, phone
Stocks: stock, aapl, dow jones, dina, apple stock, smith, fidelity, fund
Matlab: matlab, data, plot, size, image, time, figure, matlab plot
Aviation: aircraft, fas, flight, tav, radar, aviation, airport, air

Table 2: Example topics illustrating LDA output for both web domains and search queries.

Analysis

The relative ease with which search queries and web domains may be interpreted by a human observer makes it fruitful to further pursue a qualitative analysis of the topic models. Running LDA on the full corpus, with 100 topics, we obtain an output of the most frequent topics of interest and the words most commonly associated with them. We manually assessed each of the output topics and found the vast majority of them to contain straightforward and obviously related tokens. In Table 2, we selected 5 topics and list their top eight tokens, in order of frequency of attribution, for both of the models trained on web domains and search queries. We gave each topic a **post-hoc** subject which we use to describe the general theme of the topic.

One may notice that the search query topics are dominated by unigrams, even though our preprocessing of the data into tokens allowed for up to 3-grams. Table 1 indicates that there are simply far more unigrams than either bi- or tri-grams in the dataset. Moreover, search is an inherently order invariant process. For example, searching for "google android" and "android google" return the same results. As such, the frequency of the individual terms will be higher than that of the bi-gram; in fact neither bi-gram

occurs across enough users to survive the pre-filtering in Algorithm 1 into the corpus of tokens. However, a bi-gram like "galaxy nexus" did occur with sufficient frequency to filter up through to the top of the same topic.

Observing tokenized n-grams offers significant strength to inference. In the "matlab" topic, for example, we observe "matlab", "plot" and "matlab plot" as frequent tokens, qualitatively confirming the intuitive cohesion of said topic. Several similar instances are observed when expanding topics past the top eight tokens. As a comparison, we performed LDA over strictly unigrams derived from the same observation period and yielded a topic containing the top 6 words of {*new, england, york, years, weather, hampshire*}. One can see that several unrelated terms such as "new england", "new york", "new years", and "new hampshire" form an uninformative topic of phrases that start with "new". This issue illustrates the need for the tokenization approach described in the tokenization and artifact discovery section and codified in Algorithm 1.

Group behavior

We noted that, for search queries, the model generated by the organizational unigram baseline approach yields a lower perplexity score than the naïve unigram approach, but is still more perplexing than the LDA model. Consistent with this observation is the assertion that while enterprise structure has some impact on human behavior, it does not completely determine it. Hence, it is worthwhile to investigate how the enterprise structure is reflected in the resulting LDA model. Table 3 summarizes the topic distributions for two technical divisions and one support division. For each division, we have included the top five topics, presented in the same style as Table 2. We include only these divisions for reasons of space; the results provided are consistent with the divisions not displayed.

A qualitative analysis of the topic clusters around the disparate divisions reveals some interesting observations about their locations in the topic space. First, note that the most prevalent topics of the technical divisions appear to focused around their areas of research. For instance, the Aerospace topics include terms semantically tied to radio frequency, radar, and space systems, while the Materials Engineering division's top topics includes mostly terms tied to semiconductors, lasers, and laboratory equipment and safety.. These observations and others indicate an explicit relationship between the charter of these technical divisions and the actual searching behavior of their employees.

Additionally, note that many of the divisions contain references to software development tools. For instance, it appears that many employees of the Aerospace and IT divisions tend to search for, and presumably use, databases, as well as OS tools, in the case of IT. Although not presented in the figure, the topic distributions of other groups also exposed their development biases. Some groups favor java and scala (Air Traffic Control) while others favor python and perl (Communications and Networking). These observations are examples of the level of insight into the collective operations of employees at the enterprise gained by applying a topic modeling framework to aggregated search query records.

Materials Engineering
Semiconductors: temperature, gold, quantum physics, physics, silicon, index, wave, semiconductor
Lasers: laser, optical, fiber optics, newport, thorlabs, photonics, lasers wavelength, array
Electronics: microwave, chess, amplifier, agilent, power, club, waveguide, ghz
Lab Safety: msds, cas, health, rating, google.com, acid, chemical, sodium home, chloride
Wires: connector, electronics, digikey, pcb, cable, connectors, datasheet, wire
Aerospace
Space: satellite, space, nasa, telescope, observatory, earth, radar, launch
Optics: camera, lens, infrared, imaging, ccd, sensor, nikon, labview
Databases: oracle, mysql, database, sql, php, apache, freebsd, server
Power: energy, power, battery, solar, cost, model, data, generator
Radar: radar, antenna, frequency, noise, phase, bandwidth, signal, filter
Information Technology
Training: sap, excel, online, training, jobs, work, you, purchase
OSs: windows, download, firefox, microsoft, outlook, install, mac os, set
Troubleshooting: error, using, file, list, access, delete, test, server
Databases: oracle, mysql, database, sql, php, apache, freebsd, server
Web: google, security, site, splunk, web, boston ma, form, international

Table 3: The 5 most common topics with respect to each of a few divisions at the enterprise.

We performed a similar analysis across divisions at the enterprise for web domains. However, the resultant distributions of the topics involved significant overlap, to the point that the analysis is of limited usefulness. Figure 4 plots the distribution of the same three divisions presented in Table 3, this time over the topic space constructed by web domains.. While the Materials Engineering division is separable, the other two are representative of the rest of the divisions, with near uniform distributions over the topic space. The two discriminating topics for this division are the **Physics** topic and one that we labeled **Electronics Purchasing**, which was not included for confidentiality reasons.

It is worth noting that technical divisions did more heavily weight technical topics, such as the **Development** and **Physics** topics from Table 2, while such topics were absent from support divisions such as HR and Travel. The lack of interpretability from the web domain topics is not entirely unexpected, as Figure 2 demonstrates that the users tend to operate over a significantly smaller vocabulary space. We would likely need to operate at a lower level of abstraction, at third- or higher-level domains to achieve the needed granularity for the grouped topic distributions to separate.

The fact that we can accurately reconstruct the charter of the technical divisions from the prevailing topics of their users and that we can infer network usage information about the support divisions indicates that the approach described above can allow a human-in-the-loop to gather otherwise opaque intelligence about their enterprise.

Conclusions

In this paper we have demonstrated the ability to characterize enterprise user behavior using network artifacts, dif-

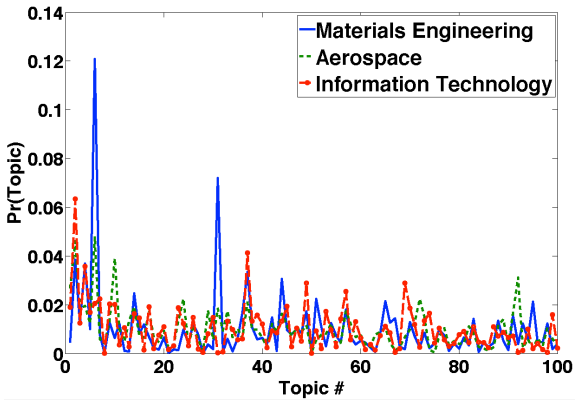


Figure 4: Distribution of three divisions over the topic space.

fering from traditional applications of topic modeling. We commented on the challenges and procedures involved in processing such logs, and were able to illustrate the coherent and informative output of LDA on a corpus of user search query terms. The results not only indicate quantitatively that our topic model exhibits a better fit to data than reasonable baselines, but also provide qualitative evidence that the resultant models are sensible to a human reader. In so doing, we gain all of the inferential power of a Bayesian topic modeling approach over our user information and may therefor gather further information about enterprise behavior and network usage.

Though there are clear privacy concerns surrounding the collection of web activity logs, automating the process of characterizing topics of interest across the entire network can potentially reduce the degree to which human operators examine individual logs; topic distributions shift the focus from individual behavior to broad patterns that might necessitate policy changes.

Further work entails utilizing topic models of the type described above to perform deep analysis of business functionality. One such mode of analysis would be to leverage machine learning and data mining tools on LDA models to perform community detection among enterprise employees for the purpose of identifying working groups that may be opaque to ground truth enterprise structure. Another future task is to perform a similar analysis including different types of network artifacts that may be opaque to human scrutiny, such as the IP addresses of internal services parsed from Kerberos requests.

Acknowledgements

The authors would like to thank Rajmonda Caceres of MIT Lincoln Laboratory and Kevin Gold[1] for their valuable comments and suggestions contributing to this work. This work is sponsored by the Department of Defense under Air Force Contract FA8721-05-C-0002. Opinions, interpretations, conclusions and recommendations are those of the authors and are not necessarily endorsed by the United States Government.

[1]Now at Google, 3 Cambridge Center, Cambridge, MA

References

Aiello, L. M.; Donato, D.; Ozertem, U.; and Menczer, F. 2011. Behavior-driven clustering of queries into topics. In *Proceedings of the 20th ACM International Conference on Information and Knowledge Management*, CIKM '11, 1373–1382. New York, NY, USA: ACM.

Blei, D. M.; Ng, A. Y.; Jordan, M. I.; and Lafferty, J. 2003. Latent Dirichlet allocation. *Journal of Machine Learning Research* 3:993–1022.

Cramer, C., and Carin, L. 2011. Bayesian topic models for describing computer network behaviors. In *ICASSP*, 1888–1891.

Griffiths, T., and Steyvers, M. 2004. Finding scientific topics. In *Proceedings of the National Academy of Sciences, 101 (suppl. 1)*, 5228–5235.

Griffiths, T. L.; Steyvers, M.; Blei, D. M.; and Tenenbaum, J. B. 2005. Integrating topics and syntax. In *In Advances in Neural Information Processing Systems 17*, 537–544. MIT Press.

Griffiths, T. L.; Tenenbaum, J. B.; and Steyvers, M. 2007. Topics in semantic representation. *Psychological Review* 114:211–244.

Hofmann, M., and Beaumont, L. R. 2005. *Content Networking: Architecture, Protocols, and Practice*. San Francisco, CA: Morgan Kaufmann.

Jordan, M., ed. 1999. *Learning in Graphical Models*. Cambridge, MA: MIT Press.

Nygren, E.; Sitaraman, R. K.; and Sun, J. 2010. The Akamai network: a platform for high-performance Internet applications. *SIGOPS Oper. Syst. Rev.* 44(3):2–19.

Peng, B.; Wang, Y.; and Sun, J.-T. 2012. Mining mobile users' activities based on search query text and context. In *Proceedings of the 16th Pacific-Asia conference on Advances in Knowledge Discovery and Data Mining - Volume Part II*, PAKDD'12, 109–120. Berlin, Heidelberg: Springer-Verlag.

Russell, B. C.; Freeman, W. T.; Efros, A. A.; Sivic, J.; and Zisserman, A. 2006. Using multiple segmentations to discover objects and their extent in image collections. In *Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Volume 2*, CVPR '06, 1605–1614. Washington, DC, USA: IEEE Computer Society.

Sivic, J.; Russell, B. C.; Efros, A. A.; Zisserman, A.; and Freeman, W. T. 2005. Discovering object categories in image collections. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

Song, W.; Zhang, Y.; Liu, T.; and Li, S. 2010. Bridging topic modeling and personalized search. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, COLING '10, 1167–1175. Stroudsburg, PA, USA: Association for Computational Linguistics.

Wallach, H. M. 2005. Topic modeling: beyond bag-of-words. In *NIPS 2005 Workshop on Bayesian Methods for Natural Language Processing*.