An Approach to Analyze Web Privacy Policy Documents

Parvathi Chundi
Computer Science Department, University of Nebraska – Omaha
6001 Dodge St, Omaha, NE, 68182.
pchundi@mail.unomaha.edu

Pranav M. Subramaniam *
Millard North High School
1010 S. 144th St, Omaha, NE 68154
submarine3.14159@gmail.com

ABSTRACT

Privacy policies at websites are often difficult to comprehend. Our project developed a system to help users understand privacy policies at websites. Policy document paragraphs were processed using a novel combination of Latent Dirichlet Allocation and complete linkage clustering to discover the underlying *themes*. Paragraphs from the document set are grouped into themes to highlight latent, related information across policies. Our system analyzed policies at 46 websites and discovered that these can be modeled using a handful of themes with the paragraphs in the themes exhibiting 60 – 80% coherence in their content. Our approach found novel themes related to password sharing, information concerning young children, etc. which are not present in current privacy tools. The proposed method is general and can be applied to find themes from policy documents of other domains, such as financial and health care.

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Miscellaneous; D.2.8 [**Software Engineering**]: Metrics—*complexity measures, performance measures*

General Terms

text mining

Keywords

latent Dirichlet allocation, text mining, privacy policies

1. INTRODUCTION

A recent study estimates that it would cost around 365 billion dollars per year in lost productivity if users were to read all privacy policies of all websites they visit [11]. Systems such as the *Nutritional Label for privacy* [6] and the Platform for Privacy Preferences [5] underscore the importance of presenting the privacy policy of a website in a format that can be comprehended by the visitors to the website. However, privacy policies of websites often contain content that may be hard to understand and evolve very quickly. Therefore, it is challenging for website users to keep track of a privacy policy as it evolves.

In this paper, we describe a system to facilitate access and comprehension of privacy policies of some of the most popular websites. Our hypothesis is that while the privacy policy content at different web sites may vary, it can be represented using a set of common **themes**, which can be extracted by automatically analyzing policies. To link privacy themes across multiple companies, we represent the content of a privacy policy document in terms of paragraphs. We then model the content of each paragraph as a random mixture over K latent topics using the latent Dirichlet allocation (LDA) method, where each topic is a probabilistic distribution over keywords. The LDA method is used as a dimensionality reduction technique which represents each paragraph as a probability distribution over topics. The topic vectors obtained from the LDA method are further clustered to find themes. We define a metric called the **coherence metric** to study the quality of the themes generated.

Our system analyzed polices at 46 popular websites including Amazon, Angry Birds, Facebook, Google, etc. The themes discovered highlighted non-obvious, latent policy information about user passwords, IP addresses, and children under 13 years of age. The theme structure discovered by our approach can be used by companies and regulatory agencies to comprehend the current privacy policy landscape. Using our system, privacy policies can be understood, not as a jumble of words, but based on a theme structure containing a set of high-level themes which immediately provide access to the parts of the policies that fit the theme. The proposed system has wide applicability and can analyze policies from several domains including finance and health care.

2. OUR APPROACH

Document Set: We downloaded 46 privacy policies from the following web sites (listed in alphabetical order) which appear in the top 100 most frequently visited websites: *Amazon, Angrybirds, Apple, Badoo, Bebo, Bing, Cartoon Network, Dropbox, Facebook, Firefox, Flickr, Foursquare, Friendster, Google, Groupon, Instagram, Keek, Kik, Line, Linkedin, Live-toolbar, Microsoft Messenger, Microsoft, Miniclip, Myspace, Netflix, Oovoo, Pandora, Pheed, Pinterest, Qooh, Quizlet, Reddit, Samsung, Shazam, Skype, Snapchat, Tumblr, Twitter, Vine, Wechat, Whatsapp, Xbox, Yahoo, Youtube,*

*Work completed while working as an intern at UNO.

KDD Workshop on Data Mining for Social Good 2014

and Zynga.

Constructing a Common Feature Space: Documents were tokenized, stop words were removed, and the rest of the words were stemmed with the Porter's Stemmer tool to get a bag of keywords. Two-word phrases were computed and combined with bag of keywords to construct the common feature space. To reduce the sparsity of the data set, only the top 200 most frequently occurring keywords and phrases were used as the features.

Model of a Document: The 46 policy documents were processed to generate a total of 1063 paragraphs using sentence detection tools from the LingPipe [2] software repository. A feature vector was built for each of the 1063 paragraphs. The feature vector for a paragraph is the bag of keywords that are in the intersection of the common feature space and the paragraph. This method resulted in empty feature vector for just 1 out of 1063 paragraphs.

Latent Dirichlet Allocation (LDA): We used the JGibbLDA [1] tool to find the topics underlying the document set. The LDA method discovers a small number (K) of latent topics from a set of documents by identifying the keywords that frequently occur together. The two important outputs of the LDA are the following:

- Topic-word map: For each of the K topics, this map specifies the m (m was set to 20) most important keywords that constitute that topic.
- Topic Vector: A vector of length K, $v = (v_1, , v_K)$, whose i^{th} component v_i specifies the probability that the document belongs to the i^{th} $(1 \leq i \leq K)$ topic.

Clustering of Topic Vectors The LDA method computes, for each document, a mixture of K topics underlying that document. Two documents are considered *similar* if their topic vectors have similar probability distributions. We employ the *Hellenger* distance [7] to compute the distance between two topic vectors. Let p and q be two topic vectors over K topics. Then, the Hellenger distance between p and q is given by the following equation. If $He(p, q)$ is small (large), then the two corresponding topic vectors have a similar (dissimilar) topic distributions, and therefore, contain a similar (dissimilar) content.

$He(p, q) = \sum_{i=1}^{n} (\sqrt{(p(x_i))} - \sqrt{(q(x_i))})$

Let D be the number of documents analyzed using the LDA method. Then we compute a $D \times D$ Hellenger distance matrix over D topic vectors where entry (i, j) in the matrix records $He(t_i, t_j)$ where t_i and t_j are two topic vectors. We then use the complete linkage clustering method [8] from LingPipe package [2] to cluster the topic vectors. Each cluster of topic vectors is then referred to as a *theme.*

3. RESULTS

We applied our method to the 1062 feature vectors, each corresponding to a paragraph in the document set. We set the value of α to $50/K$ where K is the number of topics and the value of β to $200/W$ where W is the number of features in the feature set. To decide the number of topics suitable for the data set, we ran the LDA method on our data set for different values of K (K = 5, 6, 7, 8, 9, 10, 50).

3.1 Determining the appropriate number of themes and topics

Determining the values for parameters K and G, a priori, was difficult. Since our system uses LDA as a dimensionality reduction technique, small number (around 10) was chosen for K. Based on the earlier P3P based models, we inferred that the privacy policies of websites discussed around a dozen or so issues. So, we set the value of G to around 10.

In order to measure the quality of the themes computed by our approach, we defined a **coherence** metric to measure the cohesiveness of the content in each theme. It uses the top p (p was set to 20) keywords of the K topics obtained from LDA and measures the ratio of the occurrences of topic keywords from the most frequently occurring topics in a theme to the total number of occurrences of feature keywords in the theme. If the ratio or the coherence metric is high, then the content in the theme is considered cohesive since it can be generated from a handful of topics. If the ratio is small, then the theme is not considered cohesive and hence, is considered noisy.

Let Γ be a theme and let $t_1, t_2, ..., t_m$ be the topics with highest probability values for the paragraphs in Γ. That is, each t_i appears with a high probability value in a large number of topic vectors in Γ. Let $t_1, ..., t_K$ topics were computed by LDA. Let W_m denote the topic keywords obtained by a union of topic keywords from $t_1, ..., t_m$. Let W be the feature set. Let $f(w, \Gamma)$ denote the occurrence frequency of word w in the content of theme Γ. We use $M(\Gamma)$ to denote the coherence metric value of theme Γ.

$$M(\Gamma) = \frac{\sum_{w \in W_m} f(w,\Gamma)}{\sum_{w \in W} f(w,\Gamma)}.$$

We considered the themes obtained for values of K = 5, 10, and 15, and values of G = 5, 10, and 15, and computed the coherence metric for each theme for each pair of K and G values. The results are shown in Figure 1(a). There are 9 box plots in the figure each labeled g_k where g stands for the number of themes and k stands for the number of LDA topics chosen. Each box plot plots the coherence measure of each of the g themes computed. For example, the box plot for 5_10, i.e.; 5 themes and 10 LDA topics, shows that the average coherence measure for the 5 themes is around 34% with the smallest value around 16% and highest around 45%. As is evident from the figure, the coherence measures are the highest for the 10_5 and 15_5 boxplots. Therefore, the most similar paragraphs in the themes appear for the cases where $G = 10, 15$ and $K = 5$.

We also examined the number of paragraphs in each of the themes computed for different values of K and G. Figure 1(b) shows 9 box plots, one for each pair of G and K values as in Figure 1(a). Each box plot g_k in the figure shows the variation in the number of paragraphs for g themes and k topics. For example box plot 5_10 contains themes with as few as 20 and as many as 600 paragraphs. The number of paragraphs per theme shows less variation and fewer outliers in cases where $G = 10, 15$ and $K = 5$.

Based on these results, we conclude that 5 is a suitable number of LDA topics for the chosen data set. The number of themes can be 10 or 15. We further examined the themes computed when $G = 10$ and $K = 5$. We labeled each theme with the most frequently occurring topic keywords in the theme. Figure 2 displays four most significant themes uncovered by our method and some of the content in each theme. These results show that the themes discovered

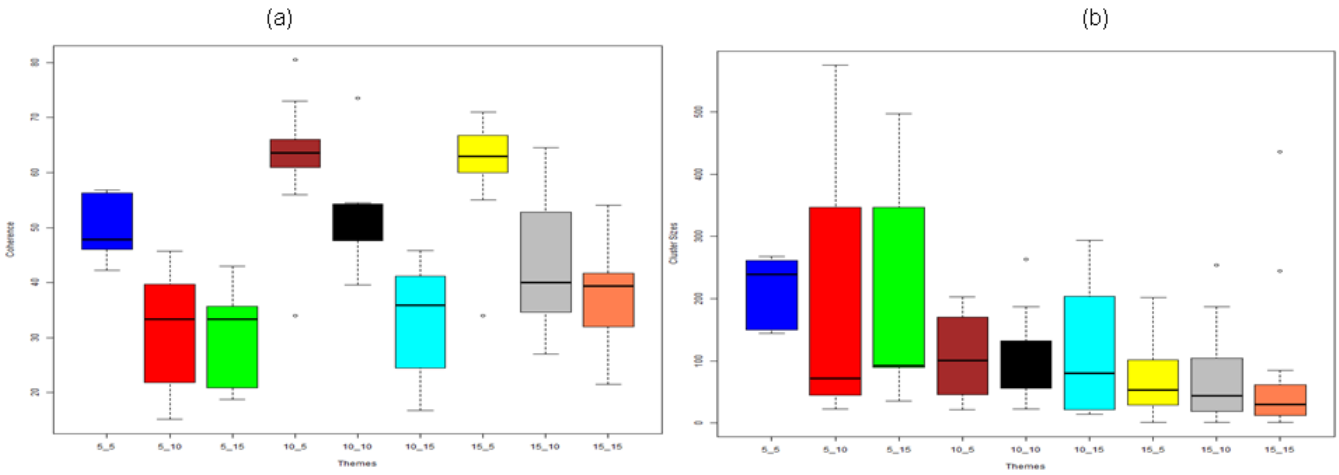

Figure 1: (a) Coherence Values of Themes Obtained, (b) Cluster Sizes of Themes, for Different Values of K and G.

by our system: *collection of personal information, sharing of information with third parties, tracking actions using cookies, etc.* are specified in the P3P standard as well. However, some themes such as *cookies and mobile devices* and *policies for young (children under 13 years of age) users* are not explicated in these systems. Hence, our method can complement the existing methods in that it can explicate the different themes that are latent in the current policy documents. (Please note that we are unable to include all 10 themes due to lack of space.)

4. RELATED WORK

User understanding of the privacy policies at websites has been a challenging problem [9, 12]. Prior research to address this problem has largely focused on creating non-textual formats for specifying privacy policies. The platform for privacy preferences (P3P), a non-textual machine-readable standard to specify privacy policies was created by the World Wide Web consortium [5]. However, the adoption of the P3P standard has been somewhat slow and software to help users read and analyze P3P policy specifications has been limited. Recently, grid based and "nutrition label" like single page summary of the privacy policy of a website based on the P3P specification have been developed to help users [6].

Unlike P3P-based approaches, our system works with policies in the same textual format in which they are typically posted on the websites using text mining methods. Recently, authors in [10] have also used LDA to study how good privacy policies are for identifying requirements. We have designed a novel combination of the generative topic modeling approach LDA [3, 4] and clustering [8] to automatically mine the themes from text policies.

The proposed bottom-up approach is more powerful than the top-down approaches such as the privacy nutritional label since it does not categorize policies into fixed topics. Themes are discovered dynamically and this enables updates to contents to be automatically reflected in the updates to the discovered themes. Content from different privacy policies is grouped into themes so that users can immediately see different aspects of the theme as well as compare content from different websites.

5. CONCLUSION AND FUTURE WORK

A system for automatically discovering themes from website privacy policies to facilitate the comprehension of these documents is described in this paper. A novel combination of LDA and complete linkage clustering was designed to discover themes from paragraphs of policy documents. A coherence measure to determine if the themes computed were meaningful in terms of their content. The system was successfully applied to policies of 46 websites and showed that the content can be described using 10 themes each having coherent content. Our future plans are to build a browser over the theme structure and investigate the use of LDA for studying different versions of a privacy policy.

Theme: children, security, personal information

Apple: We do not knowingly collect personal information from children under 13 except where a parent has set up an Apple ID for their child through the Apple ID for Students program and provided Apple with verified parental consent. Learn more about the Apple ID for Students program and device parental controls. If we learn that we have collected the personal information of a child under 13 without first receiving verifiable parental consent we will take steps to delete the information asap..

Quizlet: Parents and legal guardians of children under 13 who are members of Quizlet.com have certain rights under COPPA, and Quizlet recognizes those rights. Parents/guardians can consent to collection and use of a child's personally identifiable information (PII) without consenting to the disclosure of information to third parties. Currently the only identifiable information we collect from children under 13 is parent's email address.

Twitter: Our Services are not directed to persons under 13.If you become aware that your child has provided us with personal information without your consent, please contact us at privacy@twitter.com. We do not knowingly collect personal information from children under 13.If we become aware that a child under 13 has provided us with personal information, we take steps to remove such information and terminate the child's account. You can find additional resources for parents and teens here.

Wechat: Children under the age of 13 are not allowed to use our services. We do not knowingly collect Personal Information from any children under the age of 13.Please contact our Privacy Officer if you believe we have any Personal Information from any children under the age of 13 and we will promptly investigate (remove) Information...

Theme: password, security, protection.

Amazon: It is important for you to protect against unauthorized access to your password and to your computer. Be sure to sign off when finished using a shared computer. Click here for more information on how to sign off.

Friendster: Friendster account of every Member is password-protected. Friendster takes every precaution to protect the information of the Members, as well as information collected from other users of the Friendster website. We use industry standard measures to protect all information that is stored on our servers and within our database. We limit the access to this information to those employees who need access to perform their job function such as our customer service personnel. If you have any questions about the security at our website, please contact us .

Samsung: using encryption where appropriate; using password protection where appropriate; and limiting access to your information (that is, only employees who carry out the purposes referenced above have access to your information).

Tumblr: Email Communications with Us: As part of the Services, you may occasionally receive email and other communications from us. Administrative communications relating to your Account (e.g., for purposes of Account recovery or password reset) are considered part of the Services and your Account, which you may not be able to opt-out from receiving. We also may send you other kinds of emails, which you can opt-out of either from your Account Settings page or by using the Ã¢<80><9c>Opt-OutÃ¢<80><9d> link in the mails themselves. Note that we will never email you to ask for your password or other Account Information; if you receive such an email, please forward it to us.

Theme: third-party, comply, change, partner, sharing

Angrybirds: Please note that Rovio's partners may have their own policies related to tracking technologies for analytics and ad-serving purposes. While Rovio requires that companies comply with Rovio's privacy policy before Rovio allows those companies to access our sites and services, Rovio does not actually control third-party data collection and use. Rovio encourages parents to review the list of operators below and to become familiar with those parties policies and practices.

Apple: Apple websites, products, applications, and services may contain links to third-party websites, products, and services. Our products and services may also use or offer products or services from third parties; for example, a third-party iPhone app. Information collected by third parties, which may include such things as location data or contact details, is governed by their privacy practices. We encourage you to learn about the privacy practices of those third parties.

Cartoon Network: External Links. Our Sites may contain links to third-party websites, services, content, and applications whose information practices may be different from ours. Following these links may allow the user to go to or interact with third party websites, services, content, and applications that may offer content or games, or conduct contests or sweepstakes that are separate from our Sites. We are not responsible for the practices of these third parties, and we recommend that visitors review their privacy policies.

Dropbox: Others working for Dropbox. Dropbox uses certain trusted third parties to help us provide, improve, protect, and promote our Services. These third parties will access your information only to perform tasks on our behalf and in compliance with this Privacy Policy.

Theme: cookie, location, store, preferences, website, mobile, device

Amazon: The Help feature on most browsers will tell you how to prevent your browser from accepting new cookies, how to have the browser notify you when you receive a new cookie, or how to disable cookies altogether. Additionally, you can disable or delete similar data used by browser add-ons, such as Flash cookies, by changing the add-on's settings or visiting the Web site of its manufacturer. Because cookies allow you to take advantage of some of Amazon.com's essential features, we recommend that you leave them turned on. For instance, if you block or otherwise reject our cookies, you will not be able to add items to your Shopping Cart, proceed to Checkout, or use any Amazon.com products that require you to Sign in.

Bing: Cookies & Similar Technologies. When you use Bing services with a web browser, we will place one or more "cookies" on your machine. For example, Bing uses a cookie with a unique identifier known as the Search ID to operate the service and enable certain search features. If you sign into Bing or other Microsoft services using a Microsoft account, we will set or read one or more additional cookies. We use these cookies to operate Bing services and provide you a more relevant search experience. You can use your browser settings to remove or block cookies...

Firefox: By default, the activities of storing and sending cookies are invisible to you. However, you can change your Firefox settings to allow you to approve or deny cookie storage requests, delete stored cookies automatically when you close Firefox,.

Skype: We don't recommend that you restrict or block cookies as this may impact on the functionality of our websites and products. However, if you choose to do so, please see the options available to you below. You can also find comprehensive information about cookies at www.aboutcookies.org.

Figure 2: Themes Discovered by our Approach.

6. REFERENCES

[1] JgibbLDA (2008), *A Java Implementation of Latent Dirichlet Allocation (LDA) using Gibbs Sampling for Parameter Estimation and Inference.*

[2] Alias-i (2008), *LingPipe 4.1.0.*.

[3] Blei, D. M., Ng, A. Y., and Jordan, M. L., (2003), *Latent Dirichlet Allocation*, Journal of Machine Learning Research.

[4] Chundi, P., Go, S. (2013), *Latent Dirichlet Allocation Method for Text Mining*, To appear in the Third Edition of Encyclopaedia of Information Technology.

[5] Cranor, L., Langheinrich, M., Marchiori, M., Presler-Marshall, M., Reagle, J. (2002), *The Platform for Privacy Preferences 1.0 (P3P1.0) Specification*, 2002.

[6] Kelley, P.G., Bresee, J., Cranor, L., Reeder, R. (2009), *A Nutrition Label for Privacy*, International Symposium On Usable Privacy and Security.

[7] Krstovski, K., Smith, D.A., Wallach, H.M., McGregor, A. (2013) *Efficient Nearest Neighbor Search in the Probability Simplex*, Internatioal Conference On Theory of Information Retrieval.

[8] Jain, A. K, and Murty, M. N., and Flynn, P. J. (1999), *Data Clustering: A Review*, ACM Computing Surveys, 31(3), 264-323.

[9] Jensen, C. and Potts, C. (2004) *Privacy policies as decision-making tools: an evaluation of online privacy notices*, ACM SIGCHI.

[10] Massey, A.K., Eisenstein, J., Anton, A.I., Swire, P.P. (2013), *Automated Text Mining of Requirements Analyses of Policy Documents*, IEEE Conference on Requirements Engineering.

[11] McDonald, A. and Cranor, L. (2008), The Cost of Reading Privacy Policies, Telecommunications Policy Research Conference.

[12] Turow, J. Feldman, L., and Meltzer, K. (2005), *Open to Exploitation: American Shoppers Online and Offline*, The Annenberg Public Policy Center.