PrivaSeer: A Privacy Policy Search Engine

Mukund Srinath(✉), Soundarya Nurani Sundareswara, C. Lee Giles, and Shomir Wilson

Pennsylvania State University, University Park, State College, PA, USA
{mukund,sxn5310,clg20,shomir}@psu.edu

Abstract. Web privacy policies are used by organisations to disclose their privacy practices to users on the web. However, users often do not read privacy policies because they are too long, time consuming, or too complicated. Attempts to simplify privacy policies using natural language processing have achieved some success, but they face limitations of scalability and generalization. While this puts an onus on researchers and policy regulators to protect users against unfair privacy practices, they often lack a large-scale collection of policies to study the state of internet privacy. To remedy this bottleneck, we present PrivaSeer, the first privacy policy search engine. PrivaSeer has been indexed on 1,400,318 English language website privacy policies and can be used to search privacy policies based on text queries and several search facets. Results can be ranked by PageRank, query-based document relevance, and the probability that a document is a privacy policy. Results also can be filtered by readability, vagueness, industry, and mentions of tracking technology, self-regulatory bodies, or regulations and cross-border agreements in the policy text. PrivaSeer allows legal experts, researchers, and policy regulators to discover privacy trends and policy anomalies in privacy policies at scale. In this paper we present the search interface, ranking technique, and filtering techniques for PrivaSeer. We create two indexes of privacy policies: one including supplementary non-policy content present in privacy policy web pages and one without. We evaluate the functionality of PrivaSeer by comparing ranking techniques on these two indexes.

Keywords: Privacy · Search engine · Ranking

1 Introduction

A privacy policy is a legal document that an organisation uses to disclose how it collects, uses, shares and secures its customers' personal data. Laws around the world such as the General Data Protection Regulation (GDPR) and the California Consumer Privacy Act (CCPA) require organisations to make their privacy policies readily available to their users. These laws assume that users will read the privacy policy of an organisation and either accept the practices or abstain from using the offered services. However, a number of studies have shown that

M. Brambilla et al. (Eds.): ICWE 2021, LNCS 12706, pp. 286–301, 2021.
https://doi.org/10.1007/978-3-030-74296-6_22

although average internet users have a basic interest in online privacy [19], they rarely read privacy policies as they are either too long [15] or too complicated to understand [7]. Additionally, suggestions to improve the comprehensibility of privacy policies [10] have not been adopted by most organisations.

Natural Language Processing (NLP) techniques to simplify privacy policies tested on small corpora of privacy policies [24,26,27] have shown promising results. However, they face issues of accuracy, scalability and generalization due to their small datasets, often consisting of fewer than 10K policies. This paucity of large datasets leads to a lack of robust techniques which could be used to easily understand the wide range of privacy policies on the web. Without automated tools, regulators in some jurisdictions (such as the European Union) rely on user complaints to investigate privacy practices [23] while others (such as the United States) rely on organisations to self-certify their compliance[1] and only investigate when a privacy policy is at odds with real world privacy practices.

To remedy the lack of a publicly accessible large-scale privacy policy resource, we present PrivaSeer[2], a privacy policy search engine that currently indexes 1,400,318 privacy policies collected from the web. PrivaSeer can be used to find policies based on policy text using facets such as sector of commerce, policy vagueness, policy readability, tracking technology mentioned, regulatory bodies mentioned and regulations or cross-border agreements mentioned in the policy text. Search results can be ranked by popularity of the website of the policy, relevance based on the query and the probability that a document is a privacy policy. To the best of our knowledge, PrivaSeer is the first search engine specifically designed to support privacy research.

PrivaSeer gives researchers the ability to quantify and examine sets of policies based on key features, enabling them to discover trends in privacy practices online. Similarly, policy regulators and legal experts can use PrivaSeer to find anomalies in policies, thereby empowering them to protect users from privacy-eroding practices. The simple and intuitive search interface allows privacy concerned web users to find features of particular privacy policies and search for privacy-friendly alternatives for everyday services.

2 Related Work

The related work can be categorised into two areas: collections of privacy policies and methods to simplify privacy policy documents.

To the best of our knowledge, PrivaSeer is the first privacy policy search engine, but a few prior attempts have focused on making privacy policies more accessible to the public. The Usable Privacy Policy Project made available a collection of 115 privacy policies with fine grained human annotation of privacy practices in policies [24]. These policies can be accessed through the website and filtered based on the URL[3]. They display the sector of activity, readability

[1] https://www.privacyshield.gov/Program-Overview.
[2] https://privaseer.ist.psu.edu/.
[3] https://explore.usableprivacy.org/.

and popularity of the website from which the policy was originally obtained. In addition, they also created a collection of seven thousand privacy polices which contain machine annotated privacy practices which can be filtered by URL. Similarly, Polisis[4] is a collection of about 31,000 privacy policies which generates automatic summaries of privacy policies based on various data practices. While the privacy policies can be filtered based on their URL, they cannot be searched based on user queries [9]. More recently, Amos et al. [1] released a longitudinal corpus of privacy policies collected from around 130,000 websites.

Privacy policies have been simplified using various machine learning approaches. PrivacyCheck is a an application that automatically summarizes privacy policies online and answers ten basic questions on any privacy policy [26]. Similarly, Privee uses both rule-based and machine learning methods to classify privacy policies based on predefined categories of privacy practices [27]. Question answering techniques to simplify privacy policies have achieved some success. The PrivacyQA corpus was introduced to promote question answering in the privacy domain [16]. Opt-Out Easy is a web browser extension designed to present available opt-out choices to users as they browse the web [3]. Additionally, Apple has begun displaying privacy labels in its MacOS and iOS app stores having collected the information from App developers; however, they are available exclusively for apps in the Apple ecosystem.

While all the above techniques are geared towards simplifying privacy policies for everyday internet users, there is a lack of tools to aid privacy researchers and help regulators manage the vast number of privacy policies online. PrivaSeer has the capacity to help researchers and regulators analyse privacy policies based on required features and enforce regulations at scale.

3 Data Collection

The privacy policies for the PrivaSeer search engine come from the PrivaSeer Corpus[5] [21,22]. Srinath et al. built the PrivaSeer Corpus using two separate crawls of the web. The first crawl occurred in July 2019 with seed URLs from Common Crawl[6], a non-profit organisation which has been releasing large monthly archives of the internet since 2008. The URLs in the Common Crawl archive were first filtered based on a selection criteria that took advantage of the fact that most privacy policy URLs either have the word 'privacy' or the words 'data' and 'protection' in them. The candidate URLs were then re-crawled. The crawled documents were put though a filtering pipeline which included language detection, document classification, duplicate and near-duplicate removal, URL re-verification and non-policy content removal.

The second crawl, in February 2020, used seed URLs from the Free Company Dataset[7]. Candidate documents were filtered using the crawl pipeline after which

[4] https://pribot.org/polisis.

[5] We refer to the corpus as PrivaSeer Corpus and the search engine as simply PrivaSeer.

[6] https://commoncrawl.org/.

[7] https://docs.peopledatalabs.com/docs/free-company-dataset.

duplicates between the first and second crawls were resolved. The Free Company Dataset provided additional website metadata such as year founded, industry, size range, country, and employee estimate. The final set in the PrivaSeer Corpus consists of around 1.4 million English language website privacy policies.

4 Search Interface

The user interface of PrivaSeer resembles a standard search engine, in order to keep the system familiar and easy to use. A user enters a query in the search text box on the landing page and can opt to search either the privacy policy URLs or the policy text by selecting a radio button. Figure 1 shows a screenshot of the landing page. Clicking *Search* takes the user to the results page.

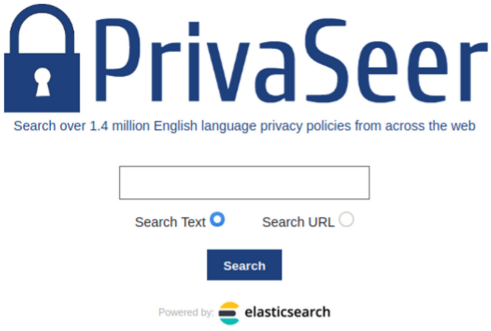

Fig. 1. Snapshot of landing page

By default, the privacy policies in the results page are ordered based on a custom ranking function discussed in Sect. 6. The result page displays the top ten results with options to go to the next page. Each result has the title of the webpage, the URL, the date it was crawled and snippets of text in the document with words matching the query words highlighted. The user can re-filter the results based on search facets on either side of the page. Figure 2 shows a screenshot of the results page for the query 'address', a common personal information type mentioned in privacy policies.

5 Indexing

We created two separate indexes: one for privacy policy web pages with non-policy content included and one without. Non-policy content refers to content in a privacy policy web page such as header, footer and navigation menus which are irrelevant to the privacy policy as a legal text. We used Elasticsearch [8] to create an inverted index and divided the documents into the title, URL, and body for indexing. We tokenized the body and title of the privacy policy using grammar based tokenization that works based on the Unicode text segmentation algorithm [5] and tokenized the URL based on a regex tokenizer.

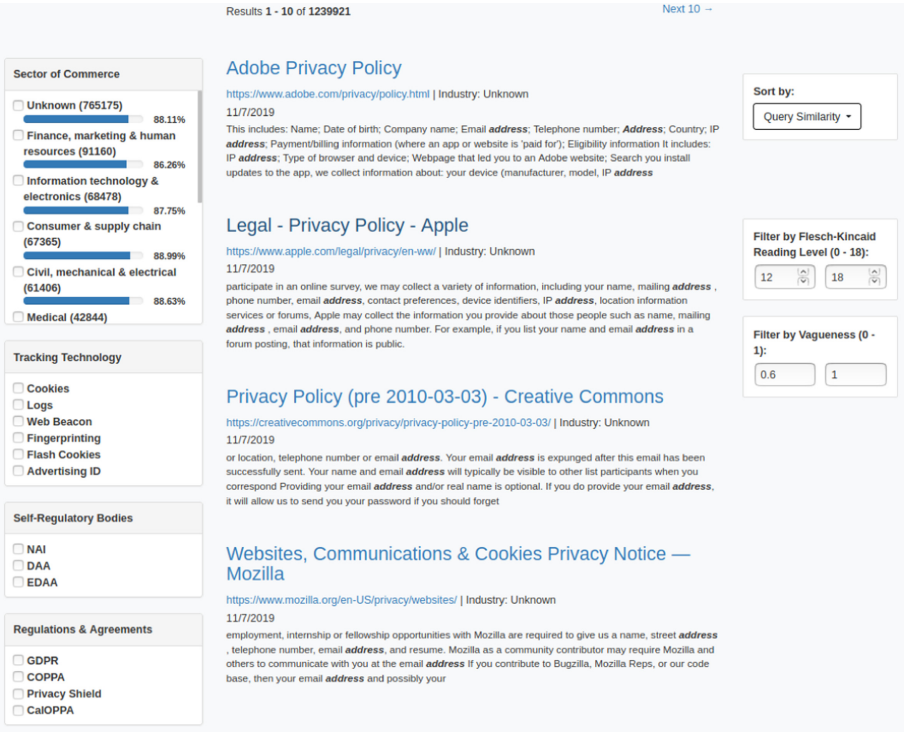

Fig. 2. Snapshot of results page

6 Ranking

The results in PrivaSeer are ranked based on PageRank, query based document relevance and the probability of the document being a privacy policy.

PrivaSeer uses the bag-of-words based Okapi BM25 [17] ranking function to estimate the relevance of a document given a search query. Given a search query Q with terms q_i where $i = i...n$, the score of a document D is given by the following function.

$$\sum_{1}^{n} idf(q_i) \times \frac{(k_1 + 1).tf(q_i, D)}{tf(q_i, D) + k_1(1 - b + b.|D|/dl_{avg})} \tag{1}$$

Where, $idf(q_i)$ is given by the following equation.

$$idf(q_i) = log\frac{N - n(q_i) + 0.5}{n(q_i) + 0.5} \tag{2}$$

In the equations, N is the total number of documents in the collection, $n(q_i)$ is the number of documents containing q_i, $tf(q_i, D)$ is the term frequency of q_i

in document D, k_1 and b are tuned constants, $|D|$ is the number of words in document D and dl_{avg} is the average document length in the collection.

We extracted the PageRanks of the domains in the corpus Common Crawl's web graph. Common Crawl use the Gauss-Seidel algorithm [2] to calculate the PageRanks in the web graph. Only a few domains have a high PageRank suggesting that ranking based only on PageRank might limit the discovery of privacy policies from the domains that are not very popular. The custom ranking function combines the scores derived from query based document relevance with PageRank and the probability of the document being a privacy policy. The final score of a document D given query Q is given by the following function.

$$P(D) \times Relevance(D, Q) \times log_{10}(D_{pr}) \tag{3}$$

In the equation, $P(D)$ is the probability that the document is a privacy policy, $Relevance(D, Q)$ is defined in Eq. 1, and D_{pr} is the PageRank of the website from which the document was crawled.

The PrivaSeer Corpus was created by training a random forest model to classify whether a document is a privacy policy. Srinath et al. [21] labeled 1000 crawled documents as either a privacy policy or not. We used the labeled data to train a machine learning model and obtained the probability of a document being a privacy policy. We used 100 documents as the validation set to tune hyperparameters and divided the rest of the documents into train and test sets in the ratio 4:1. We then tokenized and removed stop words before using term-frequency inverse-document-frequency features extracted from the URL and document. The average precision and recall score after 5-fold cross validation were 0.96 and 0.97 respectively.

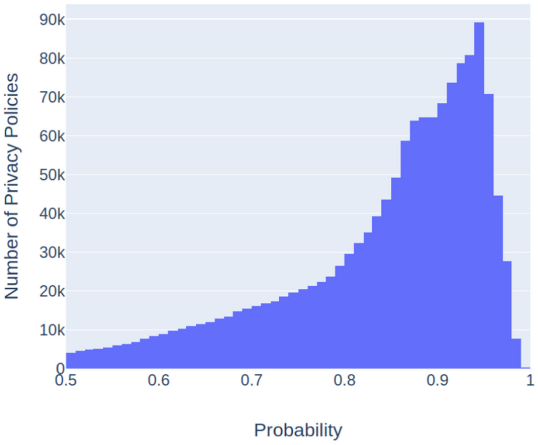

Fig. 3. Distribution of probabilities of documents (being a privacy policy) in the PrivaSeer Corpus

Figure 3 shows the distribution of the probabilities of documents classified as a privacy policy for all the documents in the PrivaSeer Corpus. The horizontal axis begins at 0.5 since the binary classification cutoff probability was 0.5, with 1 being the label for a privacy policy and 0 being the label for its negation. The figure shows that most of the documents are classified with high confidence with only a few documents having a probability less than 0.7.

7 Filtering and Observations on the Document Set

7.1 Sectors of Commerce

The Free Company Dataset, which was used to obtain seed URLs for the PrivaSeer Corpus, maps website URLs to a set of 148 unique industries. Two researchers worked independently and arrived at a consensus to consolidate the industries into 11 sectors of commerce. Table 1 shows the distribution of privacy policies across different sectors of commerce in the PrivaSeer Corpus. *Unknown* consists of extracted privacy policies whose sector of commerce information could not be found on the Free Company Dataset. Expected norms for privacy practices differ based upon sectors of commerce. For example, privacy policies in the medical sector are more likely to address users' health information, which has its own privacy laws (i.e., HIPAA in the US). Thus, we provide sector of commerce as a filter facet to enable policy regulators to compare and find anomalies between privacy policies of the same sector.

Table 1. Distribution of privacy policies across different sectors of commerce

Sector of commerce	Number
Unknown	858,395
Finance, Marketing and Human Resources	106,732
Information Technology and Electronics	82,192
Consumer and Supply Chain	77,477
Civil, Mechanical and Electrical	70,209
Medical	49,918
Sports, Media and Entertainment	43,912
Education	35,468
Government, Defense and Legal	29,037
Travel, Food and Hospitality	28,290
Non-Profit	18,688

Figure 2 shows the screenshot of the results page with sector of commerce as a filter facet. For a given query, the number of privacy policies in each sector is specified next to sector name. The progress bar and percentage value for each sector indicate the number of privacy policies retrieved for the query out of the total number of privacy policies for that sector in PrivaSeer Corpus.

7.2 Readability

Readability of a text can be defined as *the ease of understanding or comprehension due to the style of writing* [12]. One of the main critiques of privacy policies is that they are too complicated to read and understand. Studies have found that privacy policies are difficult to read and require a college-level reading ability [6,7]. The online privacy paradigm follows the *Notice and Choice* framework. *Notice* is a presentation of terms by an organisation, usually in the form of a privacy policy and *choice* is an action by a user signifying the acceptance of terms [20]. When privacy policies are difficult to understand, the notice and choice framework breaks down. To assess readability, we calculate the Flesch-Kincaid Grade Level (FKG) [11] for all the policies in the corpus and include it as a facet to filter privacy policies. FKG gives the United States school grade level an average user would need to be in order to understand the text. Srinath et al. show the distribution of readability scores in the PrivaSeer Corpus based on a number of readability techniques [21].

7.3 Tracking Technology

Tracking technologies are used by organisations to keep track of web users' browsing habits. We selected six different types of tracking technologies and extracted their mentions in all the policies in the PrivaSeer Corpus. Table 2 shows the distribution of these mentions, extracted using regex queries in an approach similar to Amos et al. [1]. To study the effectiveness of the regex technique, we manually sampled 10 privacy policies from each category of the facet and found no false positives for any category. Intuitively, privacy policies rarely mention tracking technologies which they do not use. While there is a possibility that mentions of some tracking technologies were not captured by the regex technique, thereby leading to false negatives, we believe that the regex expressions captured the common terms for all tracking technology thereby minimizing false negatives.

Studies have found a misalignment between the use and mentions of various tracking technology in privacy policies [1]. While tracking technologies are common in practice, they may not always be mentioned in the privacy policy. Thus, to enable further investigation of discrepancies and trends in the use of tracking technology we include tracking technologies as a facet.

Table 2. Distribution of tracking technology

Tracking technology	Number of policies	% of total
Cookies	1,179,351	84.2%
Logs	249,901	17.8%
Web Beacon	236,099	16.9%
Fingerprinting	73,969	5.3%
Flash Cookies	39,199	2.8%
Advertising ID	15,366	1.1%

7.4 Self-regulatory Bodies

Some jurisdictions rely on organisations to self-certify their privacy regulation compliance. Organisations therefore work with self-regulatory bodies to provide them with privacy seals and certificates verifying the organization's adherence to certain specified privacy standards [18]. Amos et al. presented a longitudinal analysis of self-regulatory compliance by examining the mentions of self-regulatory bodies in privacy policies [1]. We applied the same method of using regex queries to extract mentions of nine self-regulatory bodies in privacy policies of the PrivaSeer Corpus. Similar to the regex technique applied to extracting tracking technologies, we sampled ten random privacy policies for each item in the facet and found no false positives.

Table 3. Distribution of self-regulatory bodies

Self-regulatory bodies	Number	% of total
NAI	88,964	6.35%
DAA	72,754	5.19%
EDAA	22,874	1.63%
BBBOnLine	3,190	
TrustArc	2,300	
CNIL	1,767	<1%
ePrivacy	899	
VeraSafe	180	
Evidon	109	

Table 3 shows the percentage of privacy policies mentioning each of the self-regulatory organizations in the PrivaSeer Corpus. Only initiatives such as Network Advertising Initiative (NAI), Digital Advertising Alliance (DAA) and European Interactive Digital Advertising Alliance (EDAA) that develop self-regulatory standards for online or digital advertising, have significant number of mentions (present in over 1% of privacy policies in the corpus). Therefore, we only provide these as filterable items in PrivaSeer.

7.5 Regulations and Agreements

While some jurisdiction rely on organisations to self-certify their privacy compliance, others rely on concrete regulations and cross-border agreements. Similar to self-regulatory bodies, we extracted mentions of eight regulations and cross-border agreements in privacy policies of the PrivaSeer Corpus. We also included a sector-specific government regulation, the Health Insurance Portability and Accountability Act (HIPAA). Table 4 shows the percentage of privacy policies mentioning different regulations and agreements in the PrivaSeer Corpus. GDPR

and Privacy Shield have the highest number of mentions among regulations and cross-border agreements respectively. We include regulations and agreements as a filter facet in PrivaSeer to enable users to identify privacy policies that refer to these regulations.

Table 4. Distribution of regulations & agreements

Regulations & agreements	Number	% of total
GDPR	228,726	16.33%
COPPA	73,745	5.27%
Privacy shield	62,778	4.48%
CalOPPA	57,819	4.13%
CCPA	13,215	
SCC	6,834	
HIPAA	3,713	<1%
BCR	2,105	

7.6 Vague Language

A term is regarded as *vague* if it admits borderline cases, where speakers are reluctant to say either the term definitely applies or does not apply [4]. Vagueness in privacy policies is a pervasive problem [13], limiting the ability of readers to precisely interpret their contents. Uncertainly in future needs prompts organisations to resort to using vague language to describe their privacy practices [4]. This diminishes the effectiveness of policies making them unclear to users, thereby reducing trust and causing potential user privacy issues. Thus, we calculate the vagueness scores of all policies in the PrivaSeer Corpus and make them available as a search facet, for regulators and researchers to study at scale.

We use the corpus on vagueness in privacy policies made available by Lebanoff and Liu [13] to calculate the vagueness of policies in PrivaSeer. To create their corpus, Lebanoff and Liu first extracted sentences from 100 privacy policies that contain 40 cue words for vagueness [4]. Each sentence was considered separately and the vague words/phrases in it were identified and annotated. Since the sentence context was not considered, co-referential words were annotated as vague. For example, in the sentence *You can find out more about this on the anonymous edits page*, the word 'this' was annotated as vague. Since our aim was to find the vagueness of the privacy policy as a whole, we ignored annotations on words that were only annotated as vague due to co-reference issues. We ignored the annotations on the following words when they were annotated as vague: *it, this they, them, that, these, here, there, you, us, we,* and *following.* We also ignored annotations on the following phrases when they were annotated as vague: *personal information, personally identifiable information,* and *third party (parties)* as they might have been defined in the privacy policy prior to usage.

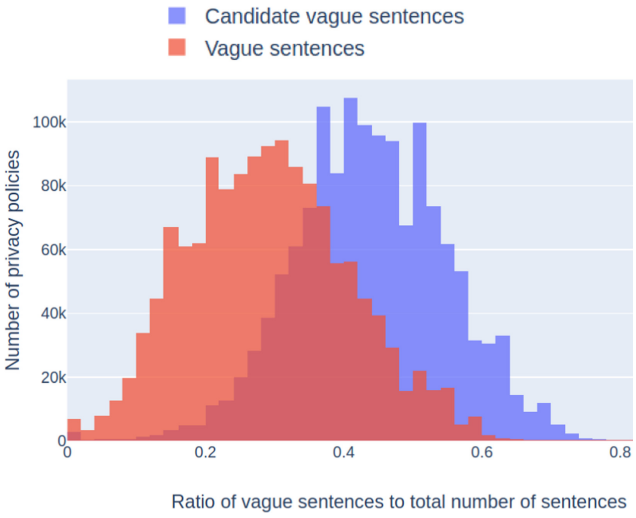

Fig. 4. Distribution of vague sentences

We treated the problem as a token classification problem where each word in a sentence would be predicted as either vague or not. We fine-tuned a pre-trained transformer based model, namely Roberta [14], using the Roberta token classification head from Huggingface [25]. We divided the corpus into train, development, and test sets in the ratio 3:1:1. We used the development set for hyper-parameter tuning. Table 5 shows the results for vague word prediction. While Lebanoff and Liu achieve better precision and recall scores, the corpus that we report on is a modified version due to the removed co-references.

For predicting vagueness of all the policies in the corpus, we extracted sentences form the corpus that had any one of the 40 cue words [4] similar to Lebanoff and Liu [13]. We call these *candidate vague sentences*. Following the candidate sentence extraction, if any word in a sentence was found to be vague by the Roberta model, we considered the sentence to be vague. We then normalised the number of vague sentences with the total number of sentences in the privacy policy to obtain a vagueness score for each policy.

Table 5. Vague word prediction results

Model	Precision	Recall	F1
Lebanoff and Liu	68.4	53.8	60.08
Roberta	65.1	52.6	58.3

The distribution of candidates and vague sentences is shown in Fig. 4. From the figure we can see that on average around 50% of the sentences in a privacy

policy are candidates for vague sentences, while about 30% are actually vague. There appears to be a long tail with some privacy policies having almost no vague sentences and some with almost all vague sentences. Manual evaluation of the policies in tail shows that most of them are very short with at most three or four sentences. Figure 2 shows the filter facet for vagueness based on a measure of the ratio of vague sentences to the total number of sentences in the policy. Users can filter results by entering a range between 0 and 1 to select the proportion of vague sentences they would like to see in policies.

8 Ranking Evaluation and Discussion

We perform an exploratory evaluation, since no prior work exists on evaluating a privacy policy search engine. Precision at k or $P@k$ measures the number of relevant results among the top k returned results. We report precision at 10 and precision at 5 scores for two indexes of privacy policies as discussed in Sect. 5, one with the context provided by non-policy content and one without.

Prior research identified ten categories of privacy practices that lawyers expect privacy policies to contain [24]. To evaluate PrivaSeer, we created three themes for queries based on the ten categories in prior work. These themes comprise of *personal information type (PI)*, *security information (S)*, and *privacy practice type (PP)*. The queries were designed and evaluated so that even if a returned result for a query was a privacy policy with the query words, it was deemed irrelevant if it did not fall in the expected category. For example, The query 'health information' is from the category *personal information type*. If a returned result for the query was a privacy policy from a hospital which did not mention how users' health information would be collected or managed, then the query was deemed irrelevant.

Table 6. Queries and their categories

Category	Queries
Personal information type	Payment information, health information, social security number, phone number, photos, private messages, microphone
Security	Firewall, encryption, SSL, data breach, deletion
Privacy practice type	Opt-out, retention period, change notification, do not track, European audience

The categories of queries and each query that was used for evaluation are shown in Table 6. Table 7 shows the comparison of P@5 and P@10 results between three ranking schemes over the different type of query categories and indexes. The results suggest that the custom ranking technique works best followed by PageRank and finally the simple query-document matching.

Table 7. PrivaSeer evaluation results

	Non-policy content Excluded						Non-policy content Included					
	PI		S		PP		PI		S		PP	
	@5	@10	@5	@10	@5	@10	@5	@10	@5	@10	@5	@10
Relevance	0.54	0.48	0.28	0.38	0.66	0.63	0.37	0.31	0.48	0.44	0.6	0.62
PageRank	0.6	0.56	0.6	0.62	0.48	0.5	0.51	0.41	0.56	0.6	0.72	0.7
Custom	**0.88**	**0.9**	**0.92**	**0.92**	**1**	**0.9**	**0.83**	**0.76**	**0.76**	**0.78**	**0.9**	**0.9**

We tested a version of the custom ranking technique without document probability scores and found that the results were slightly better than either PageRank or query based document relevance individually. Although this technique was able to leverage PageRank and query based document relevance scores together, we found that it performed poorly in cases where false positive privacy policies came from domains with a high PageRank. It was only able to perform well when both PageRank and query based document relevance scores presented reasonable results on their own. The use of document probabilities significantly improved ranking performance. The document probability scores suppress documents with a high PageRank or high query based document relevance scores but which might not be a privacy policy in reality.

Performance of all the techniques deteriorated on the index with non-policy content, across all the categories. This suggests that content in the header, footer or navigation menu do not provide much context while ranking queries related to privacy practices. It is likely that non-policy content would improve ranking in cases where users would like to filter results based on a specific industry. While the 'sector of commerce' facet allows users to filter results based on course grained industry categories, queries which include industry specific words on an index with non-policy content might serve as a stronger filter.

The custom ranking technique outperforms the PageRank and query based document relevance techniques and also has a higher variability in the returned results when compared to the PageRank technique. The PageRank technique usually returns the same set of documents for most queries. We hypothesize that this behaviour is because most popular websites have a comprehensive coverage of privacy practices.

9 Conclusion

We present PrivaSeer, the first privacy policy search engine. PrivaSeer is a necessary tool that is the first of its kind and is helpful to several distinct groups with goals in furthering user privacy. Documents can be ranked by query based document relevance scores, PageRank values, and document probabilities. They also can be filtered based on sector of commerce, policy vagueness, policy readability, tracking technology mentioned, regulatory bodies mentioned, and regulations/cross-border agreements mentioned in the policy text.

On average about 30% of the sentences in a privacy policy were found to have at least one vague word in them. This suggests that vagueness in privacy policy documents is a pervasive problem. We used regex text matching to extract details about tracking technology, regulatory bodies, and regulations/cross-border agreements and found non instances of false positives. We believe this is because privacy policies only record elements of privacy that they use/comply while rarely mentioning other elements/alternatives that exist.

An exploratory evaluation of PrivaSeer based on PageRank, query based relevance, and our custom ranking technique found that the custom ranking technique outperformed the others in all categories. We found that our custom technique had higher variability in returned results and was able to overcome limitations caused by the presence of false positive privacy policies in the results. Future work could concentrate on adding a temporal component to the collection of privacy policies and explore alternative ranking methods.

Acknowledgements. This work was partly supported by a seed grant from the College of Information Sciences and Technology at the Pennsylvania State University. We also acknowledge Adam McMillen for technical support and Ellen Poplavska for providing feedback.

References

1. Amos, R., Acar, G., Lucherini, E., Kshirsagar, M., Narayanan, A., Mayer, J.: Privacy policies over time: curation andanalysis of a million-document dataset. arXiv preprint arXiv:2008.09159 (2020)
2. Arasu, A., Novak, J., Tomkins, A., Tomlin, J.: Pagerank computation and the structure of the web: experiments and algorithms. In: Proceedings of the Eleventh International World Wide Web Conference, Poster Track, pp. 107–117 (2002)
3. Bannihatti Kumar, V., et al.: Finding a choice in a haystack: automatic extraction of opt-out statements from privacy policy text. In: Proceedings of The Web Conference, vol. 2020, pp. 1943–1954 (2020). https://doi.org/10.1145/3366423.3380262
4. Bhatia, J., Breaux, T.D., Reidenberg, J.R., Norton, T.B.: A theory of vagueness and privacy risk perception. In: 2016 IEEE 24th International Requirements Engineering Conference (RE), pp. 26–35. IEEE (2016). https://doi.org/10.1109/RE.2016.20
5. Davis, M., Iancu, L.: Unicode text segmentation. Unicode Stand. Annex **29**, 1–30 (2012)
6. Ermakova, T., Fabian, B., Babina, E.: Readability of privacy policies of healthcare websites. Wirtschaftsinformatik **15**, 1–15 (2015)
7. Fabian, B., Ermakova, T., Lentz, T.: Large-scale readability analysis of privacy policies. In: Proceedings of the International Conference on Web Intelligence, pp. 18–25 (2017). https://doi.org/10.1145/3106426.3106427
8. Gormley, C., Tong, Z.: Elasticsearch: The Definitive Guide: A Distributed Real-Time Search and Analytics Engine. O'Reilly Media, Inc., Newton (2015)
9. Harkous, H., Fawaz, K., Lebret, R., Schaub, F., Shin, K.G., Aberer, K.: Polisis: automated analysis and presentation of privacy policies using deep learning. In: 27th USENIX Security Symposium, pp. 531–548 (2018)

10. Kelley, P.G., Cesca, L., Bresee, J., Cranor, L.F.: Standardizing privacy notices: an online study of the nutrition label approach. In: Proceedings of the SIGCHI Conference on Human factors in Computing Systems, pp. 1573–1582 (2010). https://doi.org/10.1145/1753326.1753561
11. Kincaid, J.P., Fishburne Jr, R.P., Rogers, R.L., Chissom, B.S.: Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel (1975). https://doi.org/10.21236/ada006655
12. Klare, G.R., et al.: Measurement of readability (1963). https://doi.org/10.1177/002194366400100207
13. Lebanoff, L., Liu, F.: Automatic detection of vague words and sentences in privacy policies. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pp. 3508–3517 (2018). https://doi.org/10.18653/v1/D18-1387
14. Liu, Y., et al.: Roberta: a robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692 (2019)
15. McDonald, A.M., Cranor, L.F.: The cost of reading privacy policies. Isjlp **4**, 543 (2008)
16. Ravichander, A., Black, A.W., Wilson, S., Norton, T., Sadeh, N.: Question answering for privacy policies: combining computational and legal perspectives. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pp. 4949–4959 (2019). https://doi.org/10.18653/v1/D19-1500
17. Robertson, S.E., Walker, S., Beaulieu, M., Willett, P.: Okapi at TREC-7: automatic ad hoc, filtering, VLC and interactive track. Nist Spec. Publ. SP **500**, 253–264 (1999)
18. Rodrigues, R., Wright, D., Wadhwa, K.: Developing a privacy seal scheme (that works). Int. Data Priv. Law **3**(2), 100–116 (2013). https://doi.org/10.1093/idpl/ips037
19. Rudolph, M., Feth, D., Polst, S.: Why users ignore privacy policies – a survey and intention model for explaining user privacy behavior. In: Kurosu, M. (ed.) HCI 2018. LNCS, vol. 10901, pp. 587–598. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-91238-7_45
20. Sloan, R.H., Warner, R.: Beyond notice and choice: privacy, norms, and consent. J. High Tech. L. **14**, 370 (2014). https://doi.org/10.2139/SSRN.2239099
21. Srinath, M., Wilson, S., Giles, C.L.: Privacy at scale: introducing the privaseer corpus of web privacy policies. arXiv preprint arXiv:2004.11131 (2020)
22. Sundareswara, S.N., Wilson, S., Srinath, M., Giles, C.L.: Privacy not found: a study of the availability of privacy policies on the web. In: Sixteenth Symposium on Usable Privacy and Security (SOUPS 2020). USENIX Association (2020)
23. Supervisor, F.E.D.P.: What to expect when we inspect (2018)
24. Wilson, S., et al.: The creation and analysis of a website privacy policy corpus. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, pp. 1330–1340 (2016). https://doi.org/10.18653/v1/P16-1126
25. Wolf, T., et al.: Transformers: state-of-the-art natural language processing. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pp. 38–45 (2020). https://doi.org/10.18653/v1/2020.emnlp-demos.6

26. Zaeem, R.N., German, R.L., Barber, K.S.: Privacycheck: automatic summarization of privacy policies using data mining. ACM Trans. Internet Technol. (TOIT) **18**(4), 1–18 (2018). https://doi.org/10.1145/3127519
27. Zimmeck, S., Bellovin, S.M.: Privee: an architecture for automatically analyzing web privacy policies. In: 23rd USENIX Security Symposium, pp. 1–16 (2014)