A Privacy Settings Prediction Model for Textual Posts on Social Networks

Lijun Chen[1], Ming Xu[1(✉)], Xue Yang[1], Ning Zheng[1], Yiming Wu[2], Jian Xu[1], Tong Qiao[2], and Hongbin Liu[1]

[1] Internet and Network Security Laboratory, School of Computer Science and Technology, Hangzhou Dianzi University, Hangzhou, China
{151050053,mxu,153050004,nzheng,jian.xu,162050127}@hdu.edu.cn
[2] School of Cyberspace, Hangzhou Dianzi University, Hangzhou, China
{ymwu,tong.qiao}@hdu.edu.cn

Abstract. Privacy issues of social media are getting tricky due to the increasing volume of social media users sharing through online social networks (OSNs). Existing privacy policy mechanisms of OSNs may not protect personal privacy effectively since users are struggle to set up the privacy settings. In this paper, we propose a privacy policy prediction model to help users to specify privacy policies for their textual posts. We investigate the semantic of posts, social context, and keywords associated with users' privacy preferences as possible indicators of decision making, and build a multi-class classifier based on their historical posts and decisions. During the cold-start periods, the proposed model integrates crowdsourcing and machine learning to recommend privacy policies for new users. Experimental results shows that the overall match rate for all the data with random forest classifier is over 70%, with more than 50% correct prediction rate for new users.

Keywords: Social networks · Privacy · Policy recommendation

1 Introduction

Wide spread of online social networks (OSNs) make the volume of personal resources publicly available on OSNs has drastically increased. The public shared content often contains sensitive information (e.g., likes, friendships, education and work experience) about people, which may be threats to one's privacy. Most OSNs allows users to manage the audiences of their posts by specifying privacy settings [12]. However, existing access control of OSNs mostly require users to manually setting their privacy policies for each post. Recent studies show that users frequently mis-configure the privacy settings [8] of their posts.

Many researchers have acknowledged the need for policy recommendation systems which can assist users to easily and properly configure privacy policies [6,7,16,18]. However, these works focus on privacy preferences of photo, location and profile etc, few have focused on predicting the privacy policies of text-based post. Text-based information contains users' behavior or personal opinions is

I. Romdhani et al. (Eds.): CollaborateCom 2017, LNICST 252, pp. 578–588, 2018.
https://doi.org/10.1007/978-3-030-00916-8_53

considered to be particularly risky from the privacy perspective. Previous works about predicting the policies of text-based posts focused on predicting whether a post is high or low privacy [9] rather than predicting a fine-grained privacy policy. Furthermore, [1,9] did not deal with the cold start problem.

In this paper, we present a random forest based privacy policy prediction model, which aims to recommend fine-grained privacy policies to users. We define privacy policy inference as a multi-class classification problem. The model takes user's past decisions and text-based content as input and output appropriate privacy policies to them. In order to build the classifier, we utilize factors in the following criteria that influence one's privacy settings of posts:

- Social context of posts. Such as users' emotion, timestamps of publish, location semantics.
- The semantics of posts' content. Generally, users with similar tend to have similar privacy preferences.
- Keywords associated with corresponding privacy policy. For instance, Alice publishes a post: {How vexing! Boss gets me overtime every day! }, she may specify that her colleague members are not allowed to see this post. Hence keywords may be "Boss", "overtime".

As for new users, our model blends crowdsourcing and machine learning techniques, and predicts a new user's privacy policies by training the data of other users. Our contributions in this paper are twofold: (1) We propose a privacy policy recommendation model aims at helping OSNs users to configure privacy settings for text-based posts. (2) Our model can predict privacy policies for new users to protecting their privacy.

The rest of the paper is organized as follows. Section 2 reviews related works. Section 3 presents the methodology of privacy recommendation model. Evaluation and user study are described in Sect. 4. Section 5 outlines some future works and summaries the paper.

2 Related Work

Several works have studied how to assist users with privacy policy configuration. [4] proposed a privacy wizard based on supervised learning to help users grant privileges to their friends whether profile attributes allow someone to see or not. Similarly, an semantics-based privacy configuration system is proposed to configure the users' privacy settings on Facebook profile [6]. To protect the privacy of location on social networks, methods of machine learning and recommendation systems are used to refine user's location privacy settings [2,14,18]. There also exist large body of work on photo sharing preferences [7,15,17], including build a binary-class classifier [17], consider social context, image content, and meta data as possible indicators of user's privacy preferences [7].

Aforementioned approaches focus on deriving policy settings for profiles, images or locations, and they mainly consider social context such as one's friend list, image content and then using recommendation system or machine learning

technology to help users to specify policies for corresponding social media. While interesting, these methods may not be sufficient to address challenges brought by text-based posts. For the reason of the privacy preference may vary substantially not only influenced by social context but also due to the semantic of texts. Textual posts contain users' social interaction and background knowledge which are considered to be particularly risky from the privacy perspective. Previous works about predicting privacy policies of textual posts were focused on judging the post is high privacy or low privacy [9] rather than a fine-grained privacy policy. [1] propose a policy recommendation system to assist users with privacy policy configuration. The authors built a classifier based on users' historical policies and posts. Nevertheless, their system may not perform well for new users during the cold-start periods. In our work, we attempt to investigate the feasibility of deploying both social context and semantics of posts, and provides users a personalized privacy policies (even during the cold-start period).

3 Methodology

Users can manage their sharing content's privacy via privacy policies. For ease of description, some basic concepts are given:

- Audience: The one who can access the user's sharing content.
- Social Groups: Subset of a user's socially connected users.

The privacy settings in most OSNs regarding the audience of a post can be one of the following four main alternatives: $\{\text{everyone}, \text{allfriends}, \text{custom}, \text{self}\}$. With setting *custom*, user deliberately specifies a customized privacy settings that includes or excludes specific social groups. For example, Alice publishes a textual post and selects two social groups(e.g., friends, colleagues) visible. Anybody other than friends and colleagues can not access this post.

3.1 Model Overview

Privacy policy prediction based on users' historical decision is suitable for our work. We focus on users who assign multiple privacy settings rather than just one to their past posts. As shown in Fig. 1, the proposed model process two types of users (New users and others). When a user publishes a post, the model builds keyword repository based on user's historical posts. After that, five features are constructed and fed into multi-class classifier to predict the privacy policy. Note that, the keyword feature is extracted based on this user's keyword repository.

In order to alleviate the cold start problem (new users on OSNs), our model integrates cowdsourcing [11] and machine learning to recommend privacy policies for new users relies on the other users' past decisions. New users include users that don't aware or understand the privacy setting mechanism for existing OSNs (their historical decisions are all *everyone*) and users that just register on the site. In this case, machine learning model may not perform well because the training data are insufficient or the historical decisions are default privacy policy.

Predictions based on crowdsourcing, which uses opinions from a group of users, are also used to provide privacy preference recommendations.

For a new user, the process of generating keyword repository is not the same as before since new user does not have enough posts and policies stored. The model transforming crowd users' policies into four labels (*everyone*, *all friends*, *custom*, *self*) and building global keyword repository based on their past posts and policies. Afterwards, extracting features and building a global classifier. On a separate note, the train set is crows users' data for predicting new user's policies.

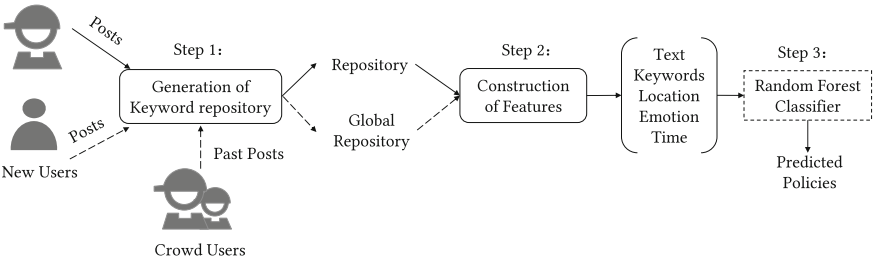


Fig. 1. Overview of the proposed model.

3.2 Feature Extraction

Word Features. Similar pieces of user-generated content have similar privacy policies. we apply short text classification technology to extract word features. All users' historical posts are considered as corpus $\Omega = \{p_1, p_2, \ldots, p_n\}$. For each post p_i in corpus Ω, tokenizing (transform the corpus Ω into small units) it to individual words. After that, we removing stop words and useless characters (e.g., special symbols) from Ω and building a term vocabulary V. Finally, we vectorizing Ω, each post p_i is represented as a m-dimensional (m represented V's dimension) vector $\mathbf{c_i} = \langle f_{i1}, f_{i2}, f_{i3}, \ldots, f_{im} \rangle$ in a vector space. TF-IDF [13] is used for calculating each word's term-weighting (f_{ik}). After above steps, each post is transformed to a row feature vector $\mathbf{c_i} = \langle f_{i1}, f_{i2}, f_{i3}, \ldots, f_{im} \rangle$.

Context Features. The sharing content always contain personal information in corresponding contexts (e.g., time, location, emotion). We consider three values (weekday, weekend, night [14]) for time, since privacy preferences are found to be time sensitive in many scenarios. Finally the time vector $\mathbf{t_i}$ has three attributes.

For location, 20 location semantics [14] are considered in our work. These location semantics are supported by Google Places: Airport, Art Gallery, Bank, Bar, Bus Station, Casino, Cemetery, Church, Company Building, Convention Center, Hospital, Hotel, Law Firm, Library, Movie Theater, Police Station, Restaurant, Shopping Mall, Spa, Workplace. $\mathbf{loc_i}$ represent the location vector.

For emotion, most of previous works obtain users' emotions by guiding users to choose one from several emotion dimensions they assume. For a deeper insight, we investigate how does emotion affect users make their privacy decisions? We randomly sample 385 posts from our data set which cover 86 participants, then utilize the lexicon-based sentiment analysis [10] technology to compute the sentiment score of each post.

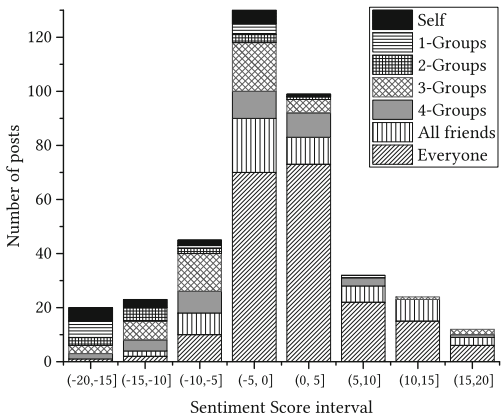


Fig. 2. The relationship between sentiment and sharing policies.

Figure 2 summarizes the relationship between sentiment score and sharing policies. The x-axis represents the sentiment score interval, ranging from -20 to 20 in our samples. The y-axis is the number of posts with selected privacy policy. To aid interpretation, privacy policies are categorized into several groups: 1-Group, 2-Group, 3-Group, 4-Group, Self, Everyone, All friends. The $x - Group$ indicates that x (x = 1, 2, 3, 4) social groups (defined in Sect. 3) are selected in the privacy settings. For instance, policy $[families, friends]$ and $[friends, colleagues]$ belong to 2-Group. From our observation, the general trend is the lower sentiment score of a post, the less visible the audiences. For example, in the interval $[-5, 5]$, more than half of the users selected *everyone* policy. But there also exists posts with low sentiment score but with *everyone* policy. This is due to different users may have different privacy preferences, yet the general trend is mostly identical.

We extract emotion features (positive, negative) $\mathbf{e_i}$ based on sentiment scores. It should be note that, normally, the score ≥ 0 represents positive emotion, vice vise. However, the dividing value of positive and negative is not zero. This is consistent with ground truth that users usually publish negative posts to vent their emotions. Hence, the threshold δ that used for deciding whether the emotion is positive or negative is -5 in our work. And this has been validated in our experiments.

Generation of Keyword Repository. The author in [9] considered the keyword feature to captures whether a post includes a keyword that might be related to certain concept like family, friends, work, etc. Their manually complied 20 representative words for all the users. Nevertheless, different users may have different privacy preferences and the keywords related to each user is different. For instance, Alice may not grant permission on her parents when she publishes a post about she is sick. Therefore, "hospital", "doctor" may be the keywords that related to the Alice's privacy preference. The process of extracting keyword features consists three steps. First, we build a keyword repository for each user. Different users have different keyword (associated with privacy preferences) repositories. We set the rule as equation Eq. (1)

$$\begin{aligned} &Dict = \{\langle P_1, Kw_1\rangle, \ldots, \langle P_n, Kw_n\rangle\}, \\ &where\ Kw_i = [k_1, \ldots, k_m]\ and\ 0 \leq i \leq n \end{aligned} \tag{1}$$

Kw_i is a list of keywords belongs to policy P_i. The process of building keywords dictionary is as follows: scan all posts that have been assigned policy *custom* or $self$, because of posts with fewer audiences may contain sensitive terms (keyword). Then adapting keyword extraction technology to extract keywords of each post and adding to Kw_i. Each P_i belongs to policy *custom* or $self$ in historical policies will maintain a keyword list Kw_i (there are no duplicate elements in this list). Finally, update the list if the next private post contains new terms.

The keyword feature vector $\mathbf{K_i} = \langle f_{k1}, f_{k2}, \ldots, f_{kn}\rangle$ of a post is built based on keyword repositories. $\mathbf{K_i}$'s dimensions is the total number of different policies (except for *everyone* and *all friends*) the i^{th} user has used in his/her historical posts. Different people may have various policies usage, thus the number of dimensions of $\mathbf{K_i}$'s for i^{th} user is different from other each users'. For instance, Alice has used the following policies: $\langle$friends, colleagues$\rangle$, $\langle$everyone$\rangle$ and $\langle$families$\rangle$ in her past decisions. $\mathbf{K_i}$'s dimensions: $n = 2$ (except policy *everyone*).

$$w_j = \begin{Bmatrix} n & \text{if } Pst \text{ has } n \text{ words that in } Kw_i \\ n \times 0.5 & \text{else } Pst \text{ has } n \text{ synonyms that in } Kw_i \end{Bmatrix} \tag{2}$$

According to Eq.(2), if a post Pst contains n words and n synonyms that exist in the Kw_i, then the value of f_{ki} is $\sum w_j$, where j=1, 2.

For example, assume Bob has a post Pst: {Today, I go to the **college**!} with policy P_1. There also exists a dictionary built from Bob's historical posts: $Dict = \{\langle P_1, [Sad,\ university]\rangle, \langle P_2, [happy]\rangle\}$, thus the Pst has one synonym: "university" in $Kw_1 = [Sad,\ university]$ and no word completely matched in Kw_1, then f_{ki} is 0.5. Note that, Chinese Open Wordnet[1] is utilized here to extract words'synonyms.

The main process of building global repository is similar to the process described above. However, the keywords in global *Dict* is the most representative words and reflect the privacy preferences of most users. The keyword extraction

[1] Chinese Open Wordnet http://compling.hss.ntu.edu.sg/cow/.

technology is adapt to extract keywords from the crowd users' historical posts. In order to prevent keyword list Kw_i from being to long, we keep top -1000 most words with the highest TF-IDF scores among all keywords.

4 Evaluation

4.1 Data Collection

In this section, we collected actual user-specified policies to be used as ground truth for our evaluation. The process of data collection is similar to the author in [7]. We recruited 160 volunteers to accomplish our online survey during December 2016 to March 2017. Volunteers are required to be Chinese-speaking and have created at least 60 text-based posts over the past eight months. The survey contains two parts. The first part contains users demographic questions, usage frequency, privacy concerns, etc. The second part is to acquire user's privacy policies. Each user will receive distinct set of posts (crawled from OSNs: QQ Zone and Facebook) according to volunteers' choices (publish frequency) in the first part. The larger the publish frequency of a volunteer, the more posts he/she received. For each post, we asked the user to choose the privacy settings by assuming these posts as their own posts. In addition, assuming volunteers have four Social Groups (SGs): friends, colleagues, families, classmates. For each question, the volunteer may choose one among the following options: *everyone*, *all friends*, *custom*, *self*. Note that *custom* policy is a combination of one or more social groups. For instance, [friends, colleagues] and [classmates] belong to policy *custom*. If the participants choose policy *custom*, the options of *everyone*, *all firends*, *self* will be invalided, vice versa. This constraint is implemented in our online survey [3] to prevent participants from inputting noisy data. Out of 160 volunteers, we just keep data for 94 of them, as the remaining ones (56) generated poor quality data (e.g., data was incomplete, policies for all posts are the same).

4.2 Experimental Settings

We conduct three sets of experiments based on the data collected from Sect. 4.1. The privacy policies in our data set were regarded as labels and we transform policies into numbers (e.g., *everyone*: 0, *all friends*: 1, *self*: 2, *custom*: 3). Note that our method can be easily adapted to different social media platforms, because the privacy policy mechanism is compatible with the existing OSNs. Three representative classification methods are considered: Random Forest (RF), Support Vector Machine (SVM), K-Nearest Neighbor (KNN).

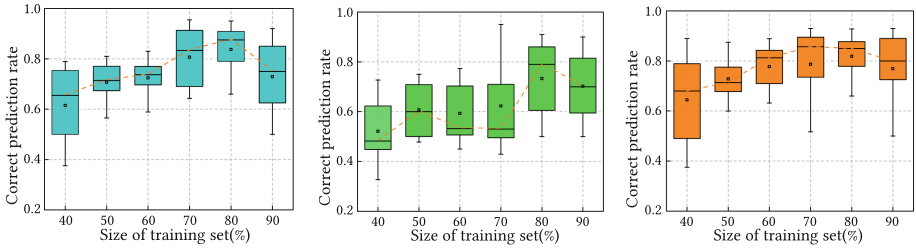
To evaluate the accuracy of our recommended policies, the following metrics are used:

- Correct prediction rate (CPR). The proportion of correctly predicted policies.
- Overall match rate. The proportion of predicted policies among all policies in our data set.

4.3 Evaluation and Analysis

The first experiment focus on evaluate the performance of the proposed model with respect to each user. We select different proportions (from 40% to 90%) of each user's data to train a classifier, and evaluate the classifier on the rest of the data (test set). The value of neighbors K for KNN we set $K = 2$, $K = 3$ and $K = 4$. Among the three values, best results can be gained when $k = 3$. Therefore, we keep using $k = 3$ for the rest of our experiments.

As shown in Fig. 3(a), (b) and (c), KNN and RF outperform SVM. KNN keeps stable performance even in the case of 40% training set, while RF outperform KNN when obtain sufficient training set. We observes that the highest median CPR reachs 0.85 at a training set of 80% with RF classifier. Above the training set of 50%, most users obtain the CPR higher than 80% with RF. Moreover, with the change of the training set size, the median of RF and KNN's CPR is similar, but KNN has more anomalous points than RF. e.g., the minimum value of CPR is lower than RF. Therefore, we take random forest classifier as our core predictive engine. Even at a training set of 40%, the median CPR is over 65%. This means we could already build an acceptable model using a very small number of posts and their past policies.



(a) RF (b) SVM (c) KNN

Fig. 3. Correct prediction rate at different sizes of training sets with RF, SVM and KNN.

Table 1 presents the overall prediction performance of our model. We calculate the proportion of correct matched policies among all policies in our data set (all users') at different training ratio. As can be seen, even we only use 40% as training data and 60% as test data, our model can still correctly predict 60.98% policies among all users with random forest and 63.41% policies with K-Nearest Neighbor.

The second experiment is to evaluate the global classifier described in Sect. 3.2. For each user u, we train a classifier with random forest on the data of the remaining users, which is then evaluated on the data of user u. New users will get one of the following four privacy policies: *everyone*, *all_friends*, *custom*, *self*. The policies of selected users (random select 50 from 94 users) in our data set are transformed into these four privacy policy. Table 2 illustrates

the results of correct prediction rate for recommending privacy polices to the 50 new users. The median correct prediction rate of all selected new users is above 68.42%. The prediction performance of our model for new users is not as good as the personalized classifier built on each user's own data. Because users may have different behaviors and privacy attitudes towards sharing content. However, such a crowdsourcing-based classifier could already provide a more acceptable performance better than a random guess.

Table 1. Correct match rate

Ratio of Training Data (from 40% to 90%)	RF Correct match rate	SVM Correct match rate	KNN (K = 3) Correct match rate
0.4	0.6098	0.4927	0.6341
0.5	0.6576	0.4971	0.7193
0.6	0.7080	0.5912	0.7591
0.7	0.7467	0.6214	0.7573
0.8	0.7941	0.6912	0.7647
0.9	0.7889	0.6865	0.7941

Table 2. Correct prediction rate for new users

Boxplot Kennwert	Maximum	Oberes Quartil	Median	Unteres Quartil	Minimum
CPR	0.7536	0.6953	0.6642	0.5410	0.4281

At the end, we evaluate the performance of decision making on different combinations of different features. As mentioned before, we use Random Forest classifier as the core prediction method and set training set ratio at 70%. The correct prediction rates of 94 users are shown in Fig. 4, the prediction rate increases along with the features. The combination of text, sentiment, keyword features achieves the highest median correct prediction rate, which is larger than 73%. It proves that combining these features together can achieve a better performance. We also found that, the median correct prediction rate will not significantly change or be improved by adding time feature and location feature. This implies that the time or location has very weak or even negative influences on decision making. The reason is that some users are unaware of location privacy and time privacy when publishing a post. However, this is not always the case for every user.

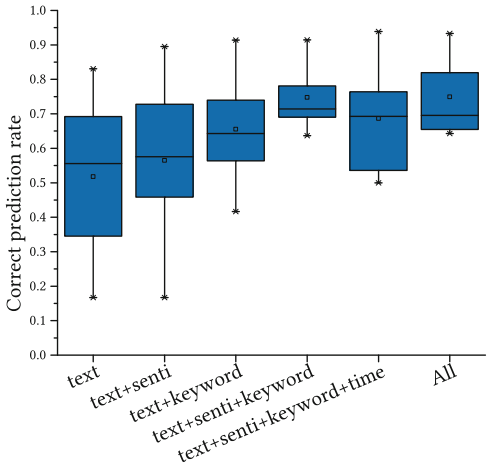


Fig. 4. Feature combination.

5 Conclusion and Future Work

In this paper, we discussed a prediction model for privacy management within text-based social media. The proposed model utilizes content and contextual information of posts to recommend fine-grained privacy policy to users. The experimental results show that our model can achieve 79.41% overall match rate with random forest (ratio of training data is 0.8). As for cold start problem, the proposed model integrates crowdsourcing and machine learning to recommend privacy policies for new users. Results show that our model can reducing the risk of being overexposed for new users.

Regarding the future work, we will try to obtain more social context information(e.g., user's social groups, photos) because users with a similar background tend to have similar privacy concerns (as been in previous research studies). The proposed model adopt TF-IDF (based on bag-of-words model) to compute the weight of each word, however, it does not capture position in text semantics, co-occurrences in different posts. Therefore, extracting word features by utilizing neural networks (e.g.,Word2Vec [5]) is a part of our future work.

Acknowledgments. This work is supported by the National Key R&D Plan of China under grant no. 2016YFB0800201, the Natural Science Foundation of China under grant no. 61070212 and 61572165, the State Key Program of Zhejiang Province Natural Science Foundation of China under grant no. LZ15F020003, the Key research and development plan project of Zhejiang Province under grant no. 2017C01065, the Key Lab of Information Network Security, Ministry of Public Security, under grant no C16603.

References

1. Abuelgasim, A., Kayem, A.: An approach to personalized privacy policy recommendations on online social networks. In: International Conference on Information Systems Security and Privacy, pp. 126–137 (2016)
2. Bilogrevic, I., Huguenin, K., Agir, B., Jadliwala, M., Hubaux, J.P.: Adaptive information-sharing for privacy-aware mobile social networks. In: ACM International Joint Conference on Pervasive and Ubiquitous Computing, pp. 657–666 (2013)
3. Chen, L.: Questionnaire about privacy preference on social networks (2016). http://sec.hdu.edu.cn/questionnaire/
4. Fang, L., Lefevre, K.: Privacy wizards for social networking sites. In: International Conference on World Wide Web, pp. 351–360 (2010)
5. Goldberg, Y., Levy, O.: word2vec explained: deriving Mikolov et al'.s negative-sampling word-embedding method. Eprint Arxiv (2014)
6. Li, Q., Li, J., Wang, H., Ginjala, A.: Semantics-enhanced privacy recommendation for social networking sites. In: IEEE International Conference on Trust, Security and Privacy in Computing and Communications, pp. 226–233 (2011)
7. Lin, D., Wede, J., Sundareswaran, S.: Privacy policy inference of user-uploaded images on content sharing sites. IEEE Trans. Knowl. Data Eng. **27**(1), 193–206 (2015)
8. Madejski, M., Johnson, M., Bellovin, S.M.: A study of privacy settings errors in an online social network. In: IEEE International Conference on Pervasive Computing and Communications Workshops, pp. 340–345 (2012)
9. Naini, K.D., Altingovde, I.S., Kawase, R., Herder, E., Niederée, C.: Analyzing and predicting privacy settings in the social web. In: Ricci, F., Bontcheva, K., Conlan, O., Lawless, S. (eds.) UMAP 2015. LNCS, vol. 9146, pp. 104–117. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-20267-9_9
10. Taboada, M., Brooke, J., Tofiloski, M., Voll, K., Stede, M.: Lexicon-based methods for sentiment analysis. Comput. Linguist. **37**(2), 267–307 (2011)
11. Toch, E.: Crowdsourcing privacy preferences in context-aware applications. Pers. Ubiquitous Comput. **18**(1), 129–141 (2014)
12. Watson, J., Besmer, A., Lipford, H.R.: +your circles: sharing behavior on Google+. In: Eighth Symposium on Usable Privacy and Security, p. 12 (2012)
13. Wu, H.C., Luk, R.W.P., Wong, K.F., Kwok, K.L.: Interpreting TF-IDF term weights as making relevance decisions. ACM Trans. Inf. Syst. **26**(3), 55–59 (2008)
14. Xie, J., Knijnenburg, B.P., Jin, H.: Location sharing privacy preference: analysis and personalized recommendation. In: International Conference on Intelligent User Interfaces, pp. 189–198 (2014)
15. Yeung, C.M.A., Kagal, L., Gibbins, N., Shadbolt, N.: Providing access control to online photo albums based on tags and linked data. In: AAAI-SSS: Social Semantic Web (2011)
16. Yuan, L., Theytaz, J., Ebrahimi, T.: Context-dependent privacy-aware photo sharing based on machine learning. In: IFIP International Conference on ICT Systems Security and Privacy Protection, pp. 93–107 (2017)
17. Zerr, S., Siersdorfer, S., Hare, J., Demidova, E.: Privacy-aware image classification and search. In: International ACM SIGIR Conference on Research & Development on Information Retrieval, Portland, Oregon, pp. 35–44 (2012)
18. Zhao, Y., Ye, J., Henderson, T.: Privacy-aware location privacy preference recommendations. In: International Conference on Mobile and Ubiquitous Systems: Computing, NETWORKING and Services, pp. 120–129 (2014)