Empirical Data Analysis on User Privacy and Sentiment in Personal Blogs

Simon S. Woo
University of Southern California
Information Sciences Institute
Los Angeles, CA
simonwoo@usc.edu

Harsha Manjunatha
University of Southern California
Information Sciences Institute
Los Angeles, CA
hmanjuna@usc.edu

ABSTRACT

Web blogging serves as a popular platform for users to express and share ideas, opinions, and information about personal interests and life. However, users might post sensitive personal information unintentionally and inadvertently based on sentiment. That could potentially lead to a compromise of users' privacy and security. In this work, we focus on extracting named entities, part of speech tags, and users' like/dislike expressed in a personal blog as a form of private information. Furthermore, we explore the relationship of private information with users' sentiments in a blog. Our empirical result with Spinn3r blog data shows that there exists positive correlation between user sentiment and potentially private information in a personal blog.

1. INTRODUCTION

Currently, we share lots of personal, social, and professional information in online social network sites (OSNs) such as Facebook and LinkedIn, as well as personal web blog sites. These OSNs and web blog platforms have contributed significantly in connecting people, sharing information, and expressing users' ideas and opinions. However, users have to face security and privacy issues as they post and share information about themselves or their close kin and friends. We believe that the seriousness of this issue become paramount as we share and post more information on the Internet via various OSN and web blogging platforms. In this work, we focus on mining user's personally related information in web blogs. Also, we hypothesize that when people are happy (high positive sentiment) or sad (high negative sentiment), people might express more about their personal attributes, interests, and personal digital footprints. This can be a significant problem which online attackers can exploit for inference attack to infer private information from individual.

2. PREVIOUS WORK

Research by [6] revealed that predicting personal traits and attributes can be a serious privacy issue. Also, [4] provided comprehensive research in finding personal information from Twitter data. In addition, [9] highlighted the potential information that can be leaked from texts, provided the examples in LinkedIn, and proposed different approaches to mining personal information. How technology and public datasets enabled the inference of private information about users easier was discussed in [1]. Further, [2] collected publicly available personal information about users in Facebook and developed an automated profiling tool to gather more information about users. However, none of the research addresses the relationship between user sentiment and personal information in web blogs.

Privacy-Preserving Information Retrieval 2015, SIGIR 2015 Workshop, August 13th, 2015, Santiago, Chile

3. OUR APPROACH

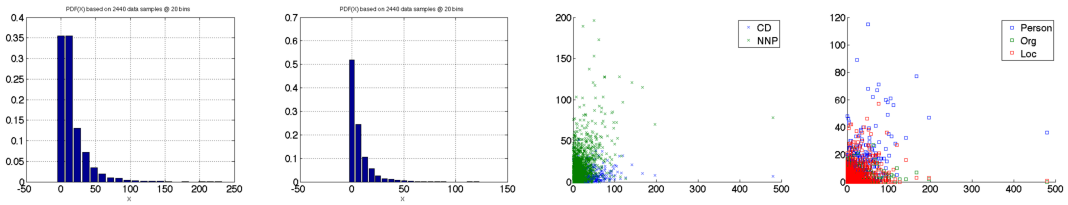
We use natural language processing tools to extract users' personal information using a natural language processing parser and the named entity recognizer (NER). Especially, we extract the Part-of-Speech (POS) tags for proper nouns (NNPs) and numbers (CD), respectively, where NNPs are often unique names for people, location, products, etc., and CD captures date, numbers, address, etc. In addition, we extract the occurrence of users' likes and dislikes in personal blogs, where users' preferences reveal personal traits as well. We used linguistic features and synonyms to identify users' like/dislike using verb patterns such as *like, enjoy, love, prefer* etc., to detect users' likes and *hate, dislike, loathe* etc., for users' dislike. Also, we use NERs to detect person, organization, and location information. We assume those linguistic POS tags, likes/dislikes, and named entities (NEs) are potential candidates for private information in users' personal blogs. For sentiment analysis, we used the SentiStrength [8] to capture positive and negative sentiment in each blog. We chose to aggregate the sentiment score over words in a sentence and also over multiple sentences for a particular blog. We believe that a wider distribution of scores allows for a better judgment of the sentiment, rather than basing the sentiment of an entire blog post on the score of a single word or sentence.

4. DATA COLLECTION

We used the ICWSM 2009 Spinn3r [3] datasets for our evaluation, where the Spinn3r datasets are a crawled collection of millions of blog posts, news articles, classifieds, and forum posts. We employed the Google Protocol Buffers API [5] to parse and cleaned up the data to obtain the pure textual content of weblog posts. Also, we used the Spinn3r API [7] to decode the protostream files into individual payload objects, each of which corresponds to a crawled webpage entry.

5. EVALUATION

We used 2,440 unique web blogs from the Spinn3r dataset. For each blog, we analyzed the empirical probability distribution function (PDF) of the sentiment score, POS tags, likes/dislikes, and



(a) Empirical PDF of POS tags (b) Empirical PDF of NEs (c) abs(SentiSum) vs. POS (d) abs(SentiSum) vs. NEs

Figure 1: Empirical PDF and Correlation Scatter Plots of POS and NEs

Case	CD	NNPS	POS Tags	Like	Dislike	Like/dislike	Person	Org.	Location	NEs
abs(SentiSum)	0.099	0.202	0.193	0.337	-0.113	0.298	0.128	0.11	0.182	0.188
abs(SentiSum)	0.335	0.456	0.462	0.391	0.146	0.407	0.429	0.177	0.22	0.42

Table 1: Pearson correlation coefficient (PCC) between sentiment score and POS tags, like/dislike, and NEs

NEs. Due to a page limit, we only present the empirical PDFs for POS and NEs, respectively, in Fig. 1. The average occurrences of POS and NEs are 16.1 and 7.1 per blog, respectively. Also, we computed the Pearson correlation coefficient (PCC) between the sentiment score and the other three features, and we summarized our correlation results in Table 1. The sum of positive and negative sentiment score denoted as *SentiSum* is used to capture the net sentiment of the story, while the absolute value of *SentiSum*, denoted as, *abs(SentiSum)*, is used to capture the strength of sentiment score. For the other PDFs, generally, we observe a heavy tail distribution across sentiment core, POS, like/dislike, and NEs. Therefore, certain blogs would contain more specific details about personal information than others. However, we detected POS and NEs much more frequently than high sentiments or like/dislike in a blog. Hence, this suggests that we can use POS and NEs as basic features to detect personal information and combine like/dislike and sentiment score as an additional features to mine personal user information. Following summarizes our findings:

1. **Sentiment vs. POS tags:** As we can see from Table 1, the Spinn3r dataset shows weak correlation (about 0.193) between the net sentiment score and POS tags. We believe that there are many different positive and negative sentiments in a single blog such that the net sentiment cancels out. Thus, we think *abs(SentiSum)* is a better metric to use. We observe higher correlation between the absolute sentiment strength and POS tags. This validates our hypothesis that when people are happy, “high positive sentiment", or sad, “high negative sentiment", people might talk about more personal information.
2. **Sentiment vs. Like/dislike:** The overall PCC between like /dislike and *abs(SentiSum)* is about 0.4. People tend to talk more about their (positive) likes than (negative) dislikes in a blog, since *like* has a higher positive correlation with sentiment than *dislike*.
3. **Sentiment vs. NEs:** The overall correlation between NEs and *abs(SentiSum)* is above 0.42. We observe a much higher sentiment correlation with Person than either Location or Organization. Therefore, people tend to express their sentiment about people rather than places or organizations.

6. CONCLUSION AND FUTURE WORK

Our preliminary results show that the interesting and positive correlations exist between user sentiment and the linguistic features we considered. The future work is to develop a classifier to accurately predict the personal information based on the features we analyzed. Also, we plan to explore if this personal blog information can be combined with information from other OSNs and public information to asses users’ security risks.

7. REFERENCES

[1] A. Acquisti and R. Gross. Predicting social security numbers from public data. *Proceedings of the National academy of sciences*, 106(27):10975–10980, 2009.

[2] M. Balduzzi, C. Platzer, T. Holz, E. Kirda, D. Balzarotti, and C. Kruegel. Abusing social networks for automated user profiling. In *Recent Advances in Intrusion Detection*, pages 422–441. Springer, 2010.

[3] K. Burton, A. Java, and I. Soboroff. The icwsm 2009 spinn3r dataset. In *Proceedings of the Third Annual Conference on Weblogs and Social Media (ICWSM 2009), San Jose, CA*, 2009.

[4] A. Caliskan Islam, J. Walsh, and R. Greenstadt. Privacy detective: Detecting private information and collective privacy behavior in a large social network. In *Proceedings of the 13th Workshop on Privacy in the Electronic Society*, pages 35–46. ACM, 2014.

[5] Google Protocol Buffers. https://developers.google.com/protocol-buffers/.

[6] M. Kosinski, D. Stillwell, and T. Graepel. Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the National Academy of Sciences*, 110(15):5802–5805, 2013.

[7] Spinn3r. https://code.google.com/p/spinn3r-client/.

[8] M. Thelwall. Heart and soul: Sentiment strength detection in the social web with sentistrength. *Cyberemotions*, pages 1–14, 2013.

[9] S. Zhang, H. Yang, and L. Singh. Increased information leakage from text. In *Proceeding of the 1 st International Workshop on Privacy-Preserving IR: When Information Retrieval Meets Privacy and Security (PIR 2014)*, page 41, 2014.