Automatically Identifying Relations in Privacy Policies

John W. Stamey
Department of Computer Science,
Coastal Carolina University
Conway, SC 29585
jwstamey@coastal.edu

Ryan A. Rossi
Jet Propulsion Laboratory,
California Institute of Technology
Pasadena, CA 91109
ryan.a.rossi@jpl.nasa.gov

ABSTRACT
E-commerce privacy policies tend to consist of many ambiguities in language that protects companies more than the customers. Types of ambiguities found are currently divided into four patterns: mitigation (downplaying frequency), enhancement (emphasizing nonessential qualities), obfuscation (hedging claims and obscuring causality), and omission (removing agents). A number of phrases have been identified as creating ambiguities within these four categories. When a customer accepts the terms and conditions of a privacy policy, words and phrases (from the category of mitigation) such as "occasionally" or "from time to time" actually give the e-commerce vendor permission to send as many spamming email offers as they deem necessary . Our study uses techniques based on Latent Semantic Analysis to discover the underlying semantic relations between words in privacy policies. Additional potential ambiguities and other word relations are found automatically. Words are clustered according to their topic in privacy policies using principal directions. This provides us with a ranking of the most significant words from each clustered topic as well as a ranking of the privacy policy topics. We also extract a signature that forms the basis of a typical privacy policy. These results lead to the design of a system used to analyze privacy policies called Hermes. Given an arbitrary privacy policy our system provides a list of the potential ambiguities along with a score that represents the similarity to a typical privacy policy.

Categories and Subject Descriptors
K.4.1 [**Public Policy Issues**]: Privacy, Ethics, Regulation. I.2.7 [**Natural Language Processing**]: Text Analysis, Language models, Language parsing and understanding. H.3.1 [**Content Analysis and Indexing**]: Dictionaries, Linguistic Processing

General Terms: Design, Usability, Languages, Legal Aspects, Information Systems, Algorithms, Experimentation, Documentation.

Keywords: Privacy Policies, Latent Relations, Ambiguities.

SIGDOC'09, October 5–7, 2009, Bloomington, Indiana, USA.

1. INTRODUCTION
Online e-commerce privacy policies should address both the protection of user information, as well as outlining safeguards about the potential vulnerabilities of Personally Identifiable Information (PII). This is initially based on the Fair Information Practice Principles Act of 1973, and later on the Electronic Communications Privacy Act of 1968. Earp, et. al. [12] identifies four characteristics of privacy protection that should be addressed: Notice/Awareness (users can find out an organization's information practices before any information is collected from them); Choice/Consent - users have the option to decide on the information that may be collected on them; Access to the site based on providing PII; and, Integrity/Security (ensuring data is both accurate and secure). Vulnerabilities in privacy policies are based on information monitoring and aggregation, information storage and transfer, collection and personalization of data, along with the ability to contact users.

Jonathan Ezor, author of Clicking Through [11] (a book about digital commerce) states that privacy policies generally contain sections regarding the following topics:

- **A.** Types of information collected;
- **B.** Information collection procedures;
- **C.** Use of information collected;
- **D.** Measures used to protect information collected;
- **E.** User access to information collected; and,
- **F.** Interaction with subsequent offers based on information collected (opting-out, etc.).

It was shown [1] that privacy policies are ambiguous in nature and protect the organization more than the users. In this work we propose a system to analyze privacy policies submitted by a user. The system identifies latent relationships between words and privacy policies automatically. The latent relationships can help users design or examine privacy policies. Our study furthers the understanding of privacy policy design and usability.

In the next section we describe the data collection and processing as well as the underlying concepts used in our analysis. In the third section we discover a ranking of privacy policy topics as well as a ranking of words from the topics. We also find various word relations between the identified privacy policy ambiguities [1] such as synonyms, antonyms and contextual similarities. These results ultimately lead us to the design of our system that identifies potential ambiguities and similarity between privacy policies. In the last section we describe the design of our system 'Hermes' in detail and show how it is used.

2. METHODOLOGY

We selected the top fifty e-commerce web sites (Apple, Ebay, Dell, Bank of America, Target, ….) and extracted their privacy policies. From this collection, we parse the text into words or tokens. We take into account any digits, hyphens, and punctuations. In addition articles and other non-distinguishable words such as {the, that, to, a, as, and, be, ….} are eliminated using a standard list of stopwords. The list of sites and stopwords are located at http://www.softwareengineeringonline.com/policy/.

We consider the standard vector-space model where a privacy policy is represented as a vector in $\Re^n$ whose coefficients are the frequency of occurrence of the words in a dictionary of size n. We further assume that we have m privacy policies, and that we use the dot product of vectors, that is the cosine of the angle formed by two vectors, to define a notion of similarity between vectors, 1 if the two vectors are identical, 0 if they are orthogonal. It is common to assume that two vectors close to each other represent two semantically related privacy policies.

From the fifty privacy policies where we have n words and m privacy policies we construct a sparse word by privacy policy matrix M. An example of a word by privacy policy matrix is shown below where the entry $[w_2, p_1]$ is a measure of the amount of times w_2 occurs in p_1.

	p_1	p_2	…	p_m
w_1	–	–	–	–
w_2	–	–	–	–
⋮	–	–	–	–
w_n	–	–	–	–

Figure 1. Word by privacy policy matrix M

In this work we apply the log entropy (log(tf + 1)entropy) weighting scheme to M [9]. The weighting is based on the distribution of words over privacy policies.

2.1 Singular Value Decomposition

Let $M \in \Re^{nxm}$, we decompose M into three matrices using Singular Value Decomposition:

$$M = U S V^T$$

where $U \in \Re^{nxm}$, $S \in \Re^{mxm}$ and $V^T \in \Re^{mxm}$. The matrix S contains the singular values located in the $[i, i]_{1,..,n}$ cells in decreasing order of magnitude and all other cells contain zero. The eigenvectors of MM^T make up the columns of U and the eigenvectors of M^TM make up the columns of V. The matrices U and V are orthogonal, unitary and span vector spaces of dimension n and m, respectively. The inverses of U and V are their transposes.

$$\begin{bmatrix} | & | & & | \\ d_1^p & d_2^p & \cdots & d_k^p \\ | & | & & | \end{bmatrix} \begin{bmatrix} s_1 & 0 & 0 & 0 \\ 0 & s_2 & 0 & 0 \\ 0 & 0 & \ddots & 0 \\ 0 & 0 & 0 & s_k \end{bmatrix} \begin{bmatrix} — & d_1^w & — \\ — & d_2^w & — \\ & \vdots & \\ — & d_k^w & — \end{bmatrix}$$

$$U \qquad\qquad S \qquad\qquad V^T$$

The columns of U are the *principal directions of the privacy policies* and the rows of V^T are the *principal directions of the words*. Similarly the rows of U are the coordinates of the words and the columns of V^T are the coordinates of the privacy policies. The principal directions are ordered according to the singular values and therefore according to the importance of their contribution to M. The singular value decomposition is used by setting some singular values to zero, which implies that we approximate the matrix M by a matrix:

$$M_k = U_k S_k V_k^T$$

A fundamental theorem by Eckart and Young states that M_k is the closest rank-k least squares approximation of M [8]. This theorem can be used in two ways. To reduce noise by setting insignificant singular values to zero or by setting the majority of the singular values to zero and keeping only the few influential singular values in a manner similar to principal component analysis.

2.2 Latent Semantic Analysis

In LSA we extract information about the relationships between words and privacy policies as they change when we set all, but the most significant, singular values to zero. The singular values in S provide contribution scores for the principal directions in U and V^T. We use the term principal direction because (assuming unit vectors) the principal eigenvector is an iterative centroid where outliers are given a decreasing weight. There have been numerous techniques based on LSA to discover signatures, synonymy, features, motifs, and many others [2-7]. In this work we take advantage of a few of these techniques.

First we give a rough example of how word (and privacy policy) relationships are automatically discovered. Let P = $\{p_1, p_2, p_3\}$ be a collection of privacy policies where w is a word such that $w_1 \in p_1$, $\{w_1, w_2\} \in p_2$, and $w_2 \in p_3$. It is clear that p_2 is 'linking' or creating a measure of similarity between both privacy policies p_1 and p_3 indirectly since p_2 shares a word with each. In this light LSA can be thought of as attempting to find the likelihood that $w_1 \in p_3$ or $w_2 \in p_1$ given the latter relationships. Let M be the word by document matrix (as seen in Figure 2) where the dimensionality is reduced by setting all but the first singular value (k = 1) to zero. Therefore M_1 is the least squares approximation of *M*. One can think of this as projecting the points onto a line where similarities that were obscured in the original space are now apparent in the reduced space. In the lower dimensionality related words are closer together.

$$M = \begin{bmatrix} 1 & 1 & 0 \\ 0 & 1 & 1 \end{bmatrix} \quad and \quad M_1 = \begin{bmatrix} 0.5 & 1 & 0.5 \\ 0.5 & 1 & 0.5 \end{bmatrix}$$

Figure 2. Construction of M and M_1

The latent relationships are automatically created. It is important to note that finding meaningful latent relationships depends on the chosen dimension to represent the data. This method has also been shown to work best on large datasets. In this example we see that the words $\{w_1, w_2\}$ are given identical definitions (in space they are represented by the same point) indicating they are strong synonyms if not the exact same word. Two or more words are recognized as synonyms if their cosine is close to 1 where the

closeness is defined by some threshold. Similarly, we see that the privacy policies $\{p_1, p_3\}$ are also represented by the same point in space. A more detailed explanation of LSA can be found in [4].

3. RESULTS

In this section we describe results and specific examples that lead to the design of Hermes a system that allows a user to analyze an arbitrary privacy policy automatically. The system identifies potential ambiguous relations that may be of concern to the user as shown in this section. The user also receives a score of how similar the user's privacy policy is to a typical privacy policy as defined by our corpus.

3.1 Topical Separation

Privacy policies can be partitioned into topics or sections (as shown in the introduction) where every topic represents a user or organizational concern.

We find that the first k *principal directions of privacy policies* (columns of U) represent the most significant topics of privacy policies. The directions are ranked according to importance and can be considered super-positions where the greatest variance between words (or documents) is captured. This can be thought of as a type of topic clustering where the corresponding word coordinates indicate how strongly a word belongs to the cluster. From every clustered topic we select the ten words with the largest word coordinate. These words are viewed as the most important words in that specific topic of a privacy policy. The topic separation is not as clear as it might be if our corpus did not strictly contain privacy policies. Nevertheless the results are interesting as this provides us with a ranking of topics in privacy policies and consequently a ranking of the corresponding words that belong to the topics.

Table 1. Ranking of Topics and Words in Privacy Policies

Rank	Topic A	Topic B	Topic C
1	privacy	browser	subscribe
2	information	secure	organization
3	products	partners	reserve
4	use	cookie	collect
5	email	phone	inaccuracies
6	services	providers	optional
7	collect	serve	demonstrate
8	provide	website	potential
9	personal	stored	participation
10	contact	track	agreed

The first topic pertains to what personal information is collected and how the personal information might be used. The second topic is about the technology used in the collection of information. The third topic is about how or whom the personal information may be disclosed. This is seen by looking at the most significant words of the topics. These topics are considered the most important topics in privacy policies. Furthermore, they are weighted according to their corresponding singular value {88.07, 27.40, 20.69} where the value represents the contribution of the topic in privacy policies. Therefore the first topic can be seen as approximately three and four times more important than the second and third topics, respectively.

It will be interesting to redo the same experiment in the future to see how the companies have evolved their privacy policies to cope with changing needs of the organization as well the user. In the future it is likely that security procedures will be more rigorously outlined in privacy policies as security continues to be a growing problem.

3.2 Discovering Word Relations

Phrases that are considered ambiguous in privacy policies have been manually identified [1]. From these phrases we selected the words that contain the most meaning with regards to privacy policies. We denote our set **A** of previously identified ambiguities where

A = {*occasionally, time, times, sometimes, trustworthy, reputable, carefully, screened, selected, interest, value, might, perhaps, discretion, except, limited, basis, reserve, right, including, authorize, without, consent, permission, knowledge, unless, sharing, shared, receive, send*}

The ambiguities 'perhaps', 'trustworthy' and 'screened' do not appear in our corpus.

It is important to see the difference between our latent relationships and the ambiguities defined in [1]. Context is used when defining ambiguities manually while we are throwing away the ordering of the words and finding relationships automatically. The latter is more formally known as the bag of words model in natural language processing [3]. Singular Value Decomposition is used to model relationships between words that appear in similar contexts[5].

To discover relationships between the words we measure the similarity of a word from the set **A** to all other words in our lexicon. Similarities that do not appear in M but appear in M_k are called Latent Relations. They appear in M_k either because noise has been removed or major components have been hidden by less important ones.

Table 2. Ambiguity Synonyms

Ambiguity		Synonyms
sometimes	≈	primarily
carefully	≈	effectively
interest	≈	appropriate
value	≈	importance
authorize	≈	permit
consent	≈	allow confirmation
knowledge	≈	acceptable
permission	≈	accordance

Synonymy is where different words describe the same idea. These words are often represented by very close points in the reduced space M_k as seen previously in our example in section 2.2. In Table 2 we show a few synonyms of the set **A** that were found automatically.

Interestingly we also find words that are antonyms clustered together as seen with the words preserve and disclose. If we look at a typical sentence from a privacy policy that contains either word: "…effort to *preserve* user privacy, we may be required to *disclose* personal information.." This relationship implies the

'true' meaning of the word preserve from a privacy policy point of view.

We also find words used in describing a specific part of a privacy policy clustered together. As an example, the word discretion is clustered with terms relating to personal information such as the users email, credit card number, address and phone number.

Table 3. Ambiguity Contextual Similarities

Ambiguity		Similarity
time	≈	change
reserve	≈	email
circumstances	≈	sending
receive	≈	purposes
shared	≈	third parties

In Table 3, we show another type of latent relationship found using our method. These words, called *contextual similarities* are not direct synonyms of the ambiguities but are used in a similar context. This can be seen by looking at the original ambiguous phrases described in [1]. A few examples are 'from time to time we may change…' or 'we reserve the right to … email …'

The process of identifying latent relations can be applied recursively to the discovered words. Using M_k one could perform a type of ontology alignment (possibly weighted) between the significant concepts of a privacy policy. It would be interesting to see a visualization of the latent relationships between the words. This will be looked at in future work.

3.3 Privacy Policy Signature

Assuming that our corpus is representative of privacy policies we can find a signature of a typical policy. The signature is extracted using the *principal direction of the privacy policies*. Let P_S be our privacy policy signature where

$$P_S = U_1 S_1$$

The rows of P_S are the coordinates of words in the reduced space where the i^{th} value can be thought of as a measure of likelihood that this word will appear in a typical privacy policy. The larger the value the more likely it is to appear. This signature can be used in many ways. In this work we use it to partly build our system.

3.4 Hermes: Protector of Privacy

We define a system based on the previous formalisms called Hermes. Given an arbitrary privacy policy or query q as input we represent the policy as a vector in $\Re^n$ whose coefficients are a measure of the frequency of occurrence of the words in our dictionary. We project the query into the reduced space by

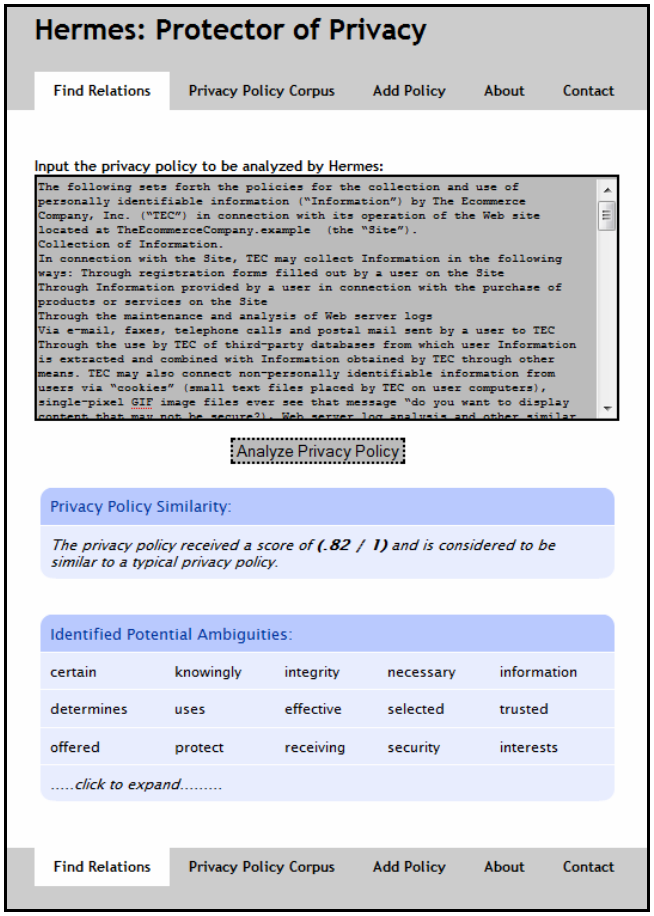
$$q_1 = q^T U_1 S_1^{-1}$$

Where U is the principal direction of the privacy policies and S^{-1} is the inverse of the contribution score of U_1. Therefore q_1 is considered a pseudo-policy with a coordinate value that can be easily compared with the 'typical' coordinate from V^T. The value is normalized between 0 and 1. This gives us a score of how 'similar' the user's query is to a typical privacy policy. If the score is greater than a threshold then the policy is considered to be similar to our typical privacy policy. We also output the ten most relevant or similar privacy policies. This allows the user to see the specific privacy policies that are most similar to the user's policy. The comparison is made with M_k where the latent relationships in the privacy policies are apparent.

The user receives as output a set of words that might be of concern or ambiguous to the user in regards to privacy. The similarity between the words in our lexicon are found by

$$M_k M_k^T = (US)(US)^T$$

Therefore given a user's privacy policy a list of potential ambiguities are returned based on the words used in the user's privacy policy.



Hermes is designed and developed using AJAX and PHP. The user interface (UI) design is very simple. The user inputs a privacy policy using the textbox and clicks the 'Analyze Privacy Policy' button. The similarity score and the potential ambiguities will appear within a few seconds. We list only the fifteen most significant potential ambiguities, but the user also has the option to view more by clicking 'expand' at the bottom of the table. The user can look at the current privacy policy corpus that was used to find the latent relations as well as the lexicon that was extracted. The user also has the option of adding a privacy policy to the current collection under certain constraints. On a regular basis we

plan to add the appropriate policies into our collection and redo the computations described previously. Hermes might perform significantly better with a larger corpus of privacy policies.

We believe Hermes will help protect users by providing a quick automatic way for them to clearly see the potential ambiguities in a given privacy policy. If a privacy policy is not 'similar' to a typical privacy policy (as defined by our collection) then this also warrants a closer look at the particular policy. The output can be used to design a privacy policy. The potential ambiguities can be used as a reference when writing or examining a privacy policy. The user can investigate whether the discovered words are used in a way that surrenders their rights or used in other potentially harmful ways.

4. CONCLUSION

We have automatically identified the most significant topics of a privacy policy and consequently the most significant words of the identified topics. We have also found potential ambiguities and extracted a signature. Using these as a basis we build a prototype system to analyze privacy policies called Hermes. The system automatically identifies potential ambiguities and outputs a score that represents the similarity of a user's policy to a typical privacy policy. In future work we plan to investigate the use of data visualization techniques to further the understanding and analysis of the latent relationships between the ambiguities and words in privacy policies.

5. ACKNOWLEDGMENTS

This work was supported in part by the National Science Foundation under Grant ATM-0521002. We thank Jean-Louis Lassez for his insightful discussions.

6. REFERENCES

[1] Pollach, I. 2007 What's Wrong with online Privacy Policies?, Communications of the ACM, Volume 50-9, 103-108.

[2] Lassez, J-L., Rossi, R., Sheel, S., Mukkamala, S. 2008 Signature Based Intrusion Detection System using Latent Semantic Analysis, IJCNN, 1068-1074.

[3] Landauer, T. K., Foltz, P. W., Laham, D. 1998 Introduction to Latent Semantic Analysis. *Discourse Processes,* **25**, 259-284.

[4] Landaur, T. K. and Dumais, S. T. 1997 A Solution to Plato's Problem: The Latent Semantic Analysis Theory of the Acquisition, Induction, and Representation of Knowledge, Psychological Review, vol. 104, pp. 211-240.

[5] Landauer, T. K. and Littman, M. L. 1990 Fully automatic cross language document retrieval using latent semantic indexing, Proceedings of the Sixth Annual Conference of the UW Centre for the New Oxford English Dictionary and Text Research., 31-38.

[6] Deerwester, S., Dumais, S. T., Landauer, T. K., Furnas, G. W. and Harshman, R. A. 1990 Indexing by latent semantic analysis. JSIS, 41(6), 391-407.

[7] J-L. Lassez, J-L., Rossi, R., Jeev, K. 2008 Ranking Links on the Web: Search and Surf Engines, Lecture Notes of Artificial Intelligence, IEA/AIE, 199-208.

[8] Eckart, C. and Young, G. 1936 The approximation of one matrix by another of lower rank, Psychometrika, 1, 211-218.

[9] Berry, M. & Browne M. 1999 Understanding Search Engines: Mathematical Modeling and Text Retrieval, SIAM.

[10] Golub, G., Reinsch, C. 1970 Singular value decomposition and least squares solutions. Numer. Math. 14, 403-420.

[11] Ezor, Jonathan, Clicking Through. Bloomberg Press, 1999, and personal communication with the authors (August 2006).

[12] Earp, J.D., Anton, A.I., Aiman-Smith, L & Stufflebeam, W.H. 2005 Examining Internet Privacy Policies Within the Context of User Privacy Values. IEEE Transactions on Engineering Management, 52(2), 227-237.