USABLE WEB 2.0 PRIVACY MANAGEMENT AND MEDICAL IMAGING SEARCH : AN ONTOLOGY-BASED APPROACH

by

NITYA VYAS

A thesis submitted to the

Graduate School—New Brunswick

Rutgers, The State University of New Jersey

in partial fulfillment of the requirements

for the degree of

Master of Science

Graduate Program in Computer Science

Written under the direction of

Danfeng Yao

and approved by

_____

_____

_____

New Brunswick, New Jersey

May, 2010

ABSTRACT OF THE THESIS

Usable Web 2.0 Privacy Management and Medical Imaging Search : An Ontology-based Approach

by Nitya Vyas

Thesis Director: Danfeng Yao

Ontology is the study of categorization of concepts and their relations. In this thesis, I provide insight towards an idea of using ontology-based approach for systems of which I present it in two different contexts. First application is in User Privacy Management in Web 2.0 and the other is in Medical Imaging Search. Both applications use user-defined annotations within a standard framework to achieve the desired results. The central idea of our technique is that users are not required to have prior knowledge about the structure of the ontology to use the system. In the first application we use annotated data in Web 2.0 social networking applications to predict privacy preferences of users and automatically derive policies for shared content. We carry out a series of user studies to evaluate the accuracy of our prediction techniques. Our analysis gives encouraging results on the feasibility of using annotations for privacy management in Web 2.0. The second application is a system for annotation and retrieval of medical images, and is built on semantic web standards. By annotating data according to standard medical ontologies, it allows the user to construct complex queries that utilize background knowledge from the underlying ontologies. Our ontology-based approach allows for several features not available in existing keyword-based search engines.

Acknowledgements

In last two years at Rutgers, I have received exposure to cutting edge research, direct application of it in the industry and gained deep knowledge in various subjects. I am thankful to several people who have contributed in making my journey worthwhile.

First and foremost, I wish to thank Prof. Danfeng Yao, without whom this work would not have been possible. I want to thank her for her encourangement, guidance, support and belief in me. She has given me invaluable advice during the times of need. I would also like to thank the members of my defense committee, Prof. Vinod Ganapathy and Prof. Alex Borgida for taking time out of their busy schedules to oversee my defense.

Department of Computer Science at Rutgers has been a great place for work and studies and I owe it to a lot of people here including Prof. Danfeng Yao, Prof. Vinod Ganapathy, Prof. Amélie Marian, Prof. Endre Szemeredi, Prof. Liviu Iftode, Prof. Vladimir Pavlovic, Prof. Rich Martin, Prof. Eric Allender for teaching me Computer Science. I thank Dr. Saikat Mukherjee for giving me the opportunity to work in Siemens Corporate Research as an intern. I wish to thank Carol DiFrancesco for helping out in all the paper work and Prof. Kate Goelz for giving me opportunity to teach. I thank my seniors Brian Thompson, Vikas Menon, Mohan Dhawan, Manuel Möeller, Gayatree Ganu, Chih-Cheng Chang for their guidance. I specially thank Brian for his k-means clustering code that we use in privacy generation tool and Dr. Khamir Mehta for giving invaluable advice related to my work and thesis.

Last but not the least, I thank my friends Vaidehi, Rajvi, Arpan, Dhruv, Vishwajit, Anand, Savan and Digant for their love and encouragement. I also thank Vatsal, Krutik, Ruchi, Ronak, Ateet for their help here in USA. I thank extended family members in India and in USA for their love and support.

Dedication

I dedicate this thesis to my Ba, Dadaji, Mummy, Pappa and Uditi.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

Today software applications are used in almost all fields. People have a greater familiarity with web applications, services and stand-alone applications. They are using web to store and share personal data more and more. Annotations on this data provide easy organization and search capabilities. Annotating is better known as tagging. Tagging is most used in Web 2.0 applications on pictures, blogs, photographs, articles etc. Annotations concisely describe the content and so they are useful in search and retrieval as well. Today medical imaging applications have also started using the annotations where each annotation describe a part of an image. `flickr.com` uses Geo-tagging which is location based tags of the photographs. Recently in `youtube.com` one can tag a part of video which is known as Video-annotations. Computer Science research community uses the data from social annotations for classification of data [55] and search optimization [8]. Ontology is a description of formal concepts for a particular domain and contains relationships between those concepts. I present here two ontology based applications one for user privacy management and other for medical imaging search both of which leverage annotations on the content. We use ontology for English dictionary words - *WordNet* [60] for privacy policy generation in Web 2.0 social network while standard ontologies developed by medical community for the medical imaging search interface.

The first application tries to address the challenge of automated user privacy management in Web 2.0. User privacy has been the most important problem with the advent of social networking and blogging applications. A personalized, quantified and easy-to-use method for managing all the content online is desirable for the users as not all users are technolgy-savvy and can not easily understand the access control mechanisms

currently available. There have been numerous cases of user privacy being breached and the consequences of it have been quite frustrating and agonizing for the users [53]. We use personal and social group annotations on the user content to develop automatic tool for managing content sharing. The framework is referred to as *APPGen* (standing for Automatic Privacy Policy Generator). It utilizes WordNet based semantic similarity analysis to generate privacy policy for the user uploaded content. It finds semantic similarity between the user annotated content and pre-defined privacy profile of her. It also does this by similarity of users in a social group of given user.

Today many medical imaging applications for X-ray, Computed Tomography, Magnetic Resonance Imaging etc. are in use for different reasons and many of them also use annotations on the medical image. These annotations basically describe part of body in image or a defect or a disease. However, all these applications store and retrieve annotations differently. The application specific annotations mean that they cannot be shared with other applications and central storage is also not possible. There are standard ontologies for medical concepts or terms available developed by medical community which describe anatomy, disease, visual characteristics for human body. These include Foundational Model of Anatomy (FMA) [54], Lexicon of Radiology (RadLex) [26] and International Classification of Diseases version 10 (ICD-10) [3]. Since these ontologies provide a hierarchy of concepts, they can be stored in semantic web format to standardize the whole storage and retrieval part. DFKI Institute in Germany developed a medical image ontology under THESEUS MEDICO [39] project again using semantic web standards to express complex structure of medical image annotations. Moreover, in this project all radiological findings of disease, patient metadata etc. can be stored in single standard form which is of much importance. I describe a search plug-in *FastMedSearch* that we implemented at Siemens Corporate Research in collaboration with DFKI Institute, Germany in second part of thesis which uses ontologies like RadLex, FMA and ICD-10 to create complex search queries on top of the THESEUS MEDICO ontology based storage. This plug-in will retrieve previously annotated images so analysis of a disease can be done easily by clinicians. The search-as-you-type plug-in *FastMedSearch* implements a visual query construction of RadLex and ICD-10

ontologies and it does not require any background knowledge for them.

1.1 Our Contributions in User Privacy Management in Web 2.0

1. We describe a new framework for automatically inferring the privacy policies for personal Web 2.0 contents, which is to improve the privacy, usability, and manageability of personal contents. The framework produces privacy policies for the content owner based on a small amount of annotation information.

2. We design privacy inference mechanisms based on the relatedness of new contents to existing knowledge by utilizing a k-means clustering method for discrete objects. Specifically, we implement three independent privacy inference techniques:

   - *social group analysis*
   - *personalization with static tag classification*
   - *personalization with dynamic tag clustering*

3. We carry out a Web-application based user study to evaluate the accuracy and usability of the privacy inference system. Our experiments show that the majority of the participants think that the framework is accurate in inferring the privacy policies. 94% of the participants voted the policy generated using our tag clustering technique as the best policy in terms of both accuracy and closeness with their "ideal policy".

1.2 Our Contributions in Medical Imaging Search

1. We provide a search interface implementation $FastMedSearch$ to build complex queries using formal ontologies for anatomy, disease, visual characteristics of an image and patient metadata in a single query to retrieve images annotated previously to analyze the defect or disease.

2. Our visualization approach of standard ontologies in $FastMedSearch$, aids inexperienced users to navigate through hierarchy and gain knowledge on the ontology.

3. RadLex ontology is light-weight and so we use it in $FastMedSearch$ for loading of visualization. However, we map RadLex terms with Foundational Model of Anatomy (FMA) terms for access to comprehensive anatomy terms from FMA which contain around 80,000 concepts.

4. We provide a sub-tree search for anatomy and disease concepts instead of equality match so that users are not required to remember specific concepts.

Part - I

Towards Automatic Privacy Management in Web 2.0 with Semantic Analysis on Annotations

Chapter 2

Motivation for User Privacy in Web 2.0

Web 2.0 revolutionizes how people store and share personal data and content today. Desktop applications are being more and more replaced by Web services. Digital documents such as photos used to be kept on the owners' hard disks, whereas today sharing of personal information and documents on the Web is pervasive. From `flickr.com` for photo sharing to `myspace.com` for profile sharing and `facebook.com`, which has the highest image uploading rate among all social network sites. The change in sharing of information has multi-faceted implications, among which privacy is the most important aspect. *Access control* is the art of defining and determining the privileges of users to certain resources. The focus of conventional access control literatures are more on the security and robustness of the authorization systems and less on the usability [57]. As conventional authorization policies are designed for use by trained professionals (e.g., system administrators), they are complex to manage and use [10, 57]. As a result, users are exposed to a number of privacy threats [63]. A significant privacy threat is raised by an increasing amount of media content posted by users on Web 2.0 platforms. User provided digital images are an integral and exceedingly popular part of profiles on social network sites. For example, Facebook hosts 10 billion user photos (as of 14 October 2008), serving over 15 million photo images per day [9]. Pictures are tied to individual profiles and often either explicitly (through tagged labeled boxes on images) or implicitly (through recurrence) identify the profile holder [5].

Web 2.0 users have to take the responsibility to manage the access of their shared contents. Although social networking and photo sharing websites provide mechanisms and default configurations for data sharing control, they are usually not intuitive, and many users do not take the appropriate time to configure their privacy preferences [6].

This type of sharing control mechanisms do not effectively protect user's content, and have resulted in privacy breaches of data in Web 2.0. As documented in public news media [53], user-provided content can be stolen, sold, used for blackmailing and have serious consequences, such as stolen identities and financial losses. Directly borrowing conventional access control approaches to Web 2.0 is not a suitable solution, as both paradigms have drastically different requirements for the authorization model. In Web 2.0, the emphasis for such models is on the *usability and manageability*. In traditional information systems, resources are owned by an organization and controlled by a team of trained professionals, whereas in Web 2.0 environments, content owners are individuals who may not be technology-savvy. A personalized, quantified, and easy-to-use method for users to manage their shared contents in Web 2.0 environments is highly desirable in order to protect the personal information of participants.

In this work, we take the first step to address the challenge of automated privacy management by presenting an *automatic policy generator based on the semantic analysis of annotations and social communities*. Our approach takes advantages of user-specified annotations, i.e., tags. The purpose of tagging is to help users organize and maintain their own contents – profiles, photos, blogs, or videos – with free-form keywords, i.e., tags. We leverage personal and social group annotations to develop automatic tools for managing content sharing. Our technique utilizes folksonomy [32] and semantic similarity analysis for automatically inferring policies in content based access control. Folksonomy is different from traditional taxonomy in that tags used to label and classify Web 2.0 contents are generated by users, not by certain authorities. Specifically, the *APPGen* system draws knowledge from two main sources: i) the similarity of users in a group of related users; ii) a pre-defined privacy profile of the user. We demonstrate the potential of our new approach by experimental evaluation and user study, which show promising initial results. In next chapter I present related work in this field. I provide formal definitions for concepts used in our framework in Chapter 4. Our privacy management framework is presented in Chapter 5. In Chapter 6, we describe our approaches of computing similarity among tags and clustering similar tags for dynamic classification, respectively. Our experiments are described in Chapter 7.

Chapter 3

Related Work in Web 2.0 Privacy Management

Several solutions related to the access control management in Web 2.0 environments [18, 12, 20] have been proposed. Gollu et. al. [18] proposes an access control scheme with social attestations. Social attestation is a piece of data that certifies a social relationship. It contains four fields: an issuer, a recipient, a social relationship between two parties, and a relationship key. Social Access Control List for an object contains owner's public key and public key of relationships of all who can access that object. Similarly Carminati, Ferrari and Perego proposed a rule-based access control model for online social networks [12]. Their solution requires data owners to issue digital certificates to participants in their social relationships. The certificates are then used for enforcing the access control rules that the data owners define. The certificates are based on the social network graph between the users and the edges of this graph represent relationships or trust. This technique uses semantic web standards to store and distribute the access rules. They are generated as triples of relationships. Relationship is verified by the chain of certificates. Digital certificate is an important security primitive that has been demonstrated useful in numerous e-commerce settings, e.g., online banking and online shopping. However, the process of generating and verifying digital certificates requires a relatively high degree of sophistication from the users, which may not be appropriate in Web 2.0 settings. In comparison, our framework is easier for average Web 2.0 users to learn and use, as access control policies are automatically generated based on social annotations rather than specified by the data owners. Compared to the work [12], a more practical but coarse-grained solution for enforcing social relationship was proposed by Mannan and van Oorschot [31]. Their idea was to leverage the existing circle of trust in Instant Messaging (IM) networks.

Apart from the work mentioned above there have been lot of work in access control using semantic web standards. [13] proposed a role based access control mechanism using OWL (Web Ontology Language) language. They considered providing synergy between access control models and semantic web-based policy language for emerging dynamic environments. Moreover, Gates [17] has described relationship based access control as one of the new security paradigms that addresses the requirements of the Web 2.0, whilst [20] proposed a content-based access control model, which makes use of relationship information available in SNs for denoting authorized subjects. Also, [19] proposed an access control mechanism for SNs using the annotations on the content. Users were required to provide annotations and the users allowed to access the post as per those annotations. However, these frameworks rely on the users input indicating their access control policies for each protected object, in order to effectively protect users' privacy. There has been much work on the customization and personalization of tag-based information retrieval as well [27, 48, 28, 61]. [8] observed that annotations are usually good summaries of corresponding web pages, and the count of annotations by different users indicare popularity of that page. Several techniques involved in exploring social annotations include association rule mining [28] and EM-based probabilistic learning approach [27, 48, 61].

The ability to evaluate the semantic similarity of words has important applications in many research fields such as psychology, linguistics, cognitive science, and biomedicine. Semantic similarity measures and tools are mostly developed by the natural language community. Most of the word similarity measures make use of WordNet [60] ontology, and these include Jiang-Conrath [23], Resnik [52], Lin [30], Banerjee-Pedersen [7] and Pirro'-Seco method [47]. The above metrics cannot be applied to phrases, as WordNet does not contain general phrases. To address this limitation, a solution for assessing phrase similarity is proposed by measuring the edit distance of parse trees and single term similarity [59]. Sentence similarity has also been studied using corpus statistics and lexical databases [29]. Motivated by the need of Web 2.0 privacy management, our work studies the categorization and clustering properties of a large number of words based on

their semantics, which differs from the existing word-word semantic analysis. Clustering methods have previously been used to cluster documents for information retrieval purpose [24], or group contexts in a large corpus of text, for example, Kulkarni and Pedersen developed SenseCluster by analyzing the lexical features and co-occurrence of phrases [25, 46]. Our clustering method differs from the existing bisecting spherical clustering approach in that we leverage the quantified distance (i.e., similarity) values provided by WordNet, and are able to significantly simplify the k-means algorithm to meet our needs.

Chapter 4

Model and Definitions

In this chapter, I provide the fundamental notions underlying our solution. We cast our techniques for managing content sharing in the context of a social application, called *APPGen*. *APPGen* helps social-network users or bloggers predict privacy policies of their shared content. The framework allows users to annotate their content (hypertext, pictures, or videos) using *tags*. A *tag* (τ), or social annotation, is a single English word, freely chosen. The *APPGen* framework predicts a privacy policy for the content just added, based on semantics of the tags, leaving the user the option to accept or decline the predicted policy. In an *initialization phase*, *APPGen* requires the user to explicitly indicate some general topics of her interest, along with privacy preferences, as she creates a Web space in the considered domain. This initial set of topics is then dynamically updated by *APPGen* as new content is added to the user's Web space. A simple use scenario of *APPGen* is as follows.

Example 1 *Suppose Alice is a new blogger, and she wishes to create her blog within the* TheSpotToBlog *social network, to reach out to old friends, and share her pictures taken while working on her favorite hobbies and activities. In the initialization phase, Alice generates a simple* privacy profile *where she indicates her topics of interest and* sensitivity values *possibly associated with the topics. This setup is a one-time process. As Alice adds new content, APPGen predicts a privacy policy to be applied to the uploaded material. Alice can choose to accept it or modify it as she wishes. In this work, we assume one tag per content for our application but this can be easily generalized to more than one tag.*

4.1 Definitions for Social Network and User Profile

We notice that social networks represent only one of the possible social computing platforms where *APPGen* could be successfully used. The requirements for *APPGen* to guarantee accurate predictions, are the use of annotations and, as discussed later in the paper, the existence of users who are *similar* to the user in the same domain. Hence, policy predictions can be applied in other Web 2.0 platforms, such as blogs, wikis, etc. We begin our formal presentation by defining social networks, tags, and users profiles.

- A *social network* is denoted by the tuple $\langle U, \mathcal{R} \rangle$, where U denotes a collection of users U, connected by social relationships $\mathcal{R}$ of different types $\{R_1, \ldots, R_k\}$. (e.g., family, friends, colleagues, school network). We assume relationships to be explicit and mutually accepted by the involved users. For simplicity we focus on binary user relationships, and denote a relationship as $u : R : u'$, u and u' being users' unique identifiers, and R the relationship that connects them. By assumption, each user is connected by at least one relationship to another user in U.

- Each user $u \in U$ has one associated Web space or *profile*, $prof$. Each $prof$ is related to one or more topics $\gamma_1, \gamma_2, \ldots, \gamma_k$ indicated by the user at the time of registration. A *topic* or a *subject* is a word that represents an area of interest or a concept. For example, a topic may be: *alcohol, adult, religion, schoolwork, sport, technology, travel, food, animal*, or *gathering*. We assume the existence of a pre-defined set of general topics Γ, which can be dynamically expanded. We assume set of topics to be universal i.e., they are known to everyone.

Users populate their profiles (or Web spaces) by adding content of different types, and content can be annotated with tags. User groups, referred to as *Social Group* represent cluster of users, sharing certain properties, such as their relationships, their interests, etc.

Definition 1 (Social Group) *Let SocG be a subset of users in U. SocG is a* social group *if and only if at least one of the following condition is satisfied: group of friends*

of a user who

1. $\forall u' \in SocG$ *there exists* $\gamma \in \Gamma$ *s.t.* γ *is associated to* $prof'$,

2. $\forall u, u' \in SocG$ *there exists a relationship u:R:u'.*

The definition identifies social groups as groups of users who share a topic of interest (condition 1), are connected through a social relationship (condition 2) or both. This notion is useful to identify correlated users, in case not enough user information is available to accurately predict a policy. Social groups are also important to infer whether users with similar features are predictive of certain privacy preferences.

4.2 User's privacy policies

Expressing privacy preferences with *APPGen* is a simple task. The user simply has to assign a sensitivity score to the topics of interest, and indicate her privacy preferences. A *sensitivity value* for a topic γ is a non-negative numerical value w that a user u assigns to γ to indicate the degree of reluctance to share the contents related to it.

We model the indication of users' preferences by means of a *user expression.*

Definition 2 *(User Expression) A user expression is an expression of the form (* $\{R_1, \ldots, R_k\}$,*Cond); where:*

- $\{R_1, \ldots, R_k\}$ *is a list of relationship kinds,* $R_i, i \in [1, k]$ *is a relationship in* $\mathcal{R}$.
- *Cond is a boolean formula, against user profile attributes.*

Example 2 *Suppose that Alice indicates her preferred topic as 'photography', at time of registration. As part of the registration process, she indicates that photography is an interest she is willing to share with friends and relatives. This preference is summarized by the expression (*{*Friends, Colleagues*}*,* $\emptyset$*), since no further conditions are enforced. If sensitive content is added regarding the photography, she wants only friends which High School is 'Art School of London' to access her profile portion. In the latter case the expression used will be of the form* $(\{Friends\}, HighSchool = ArtSchoolLondon)$.

User expressions represent the building blocks for both privacy profiles, and privacy policies. The collection of sensitivity values along with related user expressions for the topics in $prof$ define the *privacy profile* of a user.

Definition 3 *(Privacy Profile) Let u be a user in U, and $prof$ be her profile. The privacy profile p of u is the list $[tup_1, \ldots, tup_n]$, where each tup_i, $i \in [1, n]$ is a tuple of the form $\langle \gamma_i, w_i, UExpr \rangle$, where γ_i is a topic, w_i the associated sensitivity value and $UExpr$ a user expression, specified according Definition 2.*

A compact representation of the privacy profile of a user u is synthesized as a vector $\overrightarrow{p_i} = [w_1, \ldots, w_n]$, where w_j is the sensitivity score for topic γ_j. As well as topics, tags are also coupled to a sensitivity score w, which value is subjective to the individual's privacy inclination. As we return later in the paper, this score is not manually input by the user, unless she wishes to do so, but inferred by *APPGen*.

In our context, a *privacy policy* (or policy for short) controls the access of a user's content. Given a Web space composed of multiple objects, the privacy policy applies to only one of these contents. The privacy policy specifies the scope of sharing, i.e., who is allowed to access the object/s posted in the profile.

Definition 4 *(Privacy Policy) Let $prof$ be the profile of a user u, and let c be some content in $prof$. A privacy policy pol is modeled as a predicate AccessTo(UExpr, Mode), where $UExpr$ is a user expression specified according to definition 2,* Mode *is a subset of admitted access modes that consists of view, modify, execute and delete.*

According to the definition, a policy constrains the set of users who can access certain content, based on the content sensitivity (namely, the w component) and on the viewers' properties (i.e., the user expression). The mode component indicates the granted access privilege.

Example 3 *Examples of policies are:* AccessTo(({U2Fans}, ϕ), read)*, and* AccessTo(({}, ϕ), read;write, { *pet* $\in prof$}). *The first policy is an example of policy with no access condition, while the second policy allows read and write operations to users who indicated* pet *in their preferred topics.*

Chapter 5

APPGen Privacy Policy Inferencing

The main goal of *APPGen* is to provide a semi-automated approach to privacy protection. A central technical question is, given the annotation of a content, how to infer the intended privacy policy for the user, while minimizing her intervention as possible. As introduced, a privacy policy essentially specifies which users are allowed to view the tagged content (say, s) of a user's profile. We can identify several approaches according to which a policy for some content s can be selected. The trivial approach would be to simply apply default policies according to the broad topic the tag falls into, and use the user's specified policy for the topic. Clearly, this approach would not allow fine-grained specification of policies, nor it would capture the user's inclinations with regards to content sharing. The opposite approach would require the user to continuously add policies each time new content is added, failing to provide any automation. *APPGen* overcomes the limitations of these approaches by using inferencing techniques to identify the *best* policies for some newly added content. Specifically, the system draws knowledge from two main sources: i) the similarity of users in a group of related users; ii) the sensitivity values of the content specified by users. We describe three main approaches, **i)** *personalization with static classification of tags*, **ii)** *personalization with dynamic clustering of tags*, and **iii)** *social-group based analysis*, for the privacy policy inference. The inference mechanisms are complementary to each other and can be integrated to yield a hybrid approach.

5.1 APPGen Policy Personalization

Given the inputs of a tag and a set of pre-defined topics or the user's previous tag history, the personalization process outputs an appropriate privacy policy for some

annotated content. The personalization component will first utilize semantic analysis techniques to discover the most similar tag in the topics or the user's profile, and then apply the appropriate policy accordingly. We present two different approaches for policy personalization, namely *static classification* and *dynamic clustering*. The two approaches are independent of each other.

- **Static Classification of Tags** utilizes a set of pre-defined topics (typically around 20), and aims to assign the tag τ to a topic γ that is semantically most similar to τ. Semantic similarity analysis is presented in more details in the next chapter. Once the topic γ is chosen, the user expression $UExpr$ associated to γ is used as the policy for the content tagged with τ.

- **Dynamic Clustering of Tags**. The analysis is between tag τ and all the previously annotated contents in the user's profile $Prof$, in order to identify the most similar content. In particular, we aim to discover a tag τ' in the user's history that is semantically most similar to tag τ. When such a tag τ' is found, the policy associated with τ' is applied to τ and the content associated with τ. In the dynamic clustering approach, the analysis is between tag τ and all the previously annotated contents in the user's profile $prof$, in order to identify the tag most similar to τ. In particular, we aim to cluster the tags in user's personal profile $prof$ into several groups based on tag semantic similarities, and then discover a cluster c whose cluster center is semantically most similar to tag τ. The cluster center is a tag in $prof$. When such a cluster and its center tag are found, the policy associated with the center tag is applied to τ and the content associated with τ.

The above approaches are called *personalization* because the analysis is based on the user's unique personal profile, as opposed to a set of uniform and generic rules defined by the system for every user.

5.2 APPGen Social Group Analysis

Social group analysis is an alternative, yet equally powerful approach, to automatically generate privacy policies for annotated content. The main idea is to leverage those users who have similar privacy preferences as the focal user (i.e., the user whose policy needs to be predicted), and to derive privacy policies based on their policy records and profiles. The users who have similar privacy preferences as the user are called *reference points* by us. We require the users who serve as reference points in this analysis to belong to the social group of the user. The purpose of this requirement is two-fold: to restrict the scope of reference points and to speed up the computation.

Once users have performed the one-time registration and we have obtained their privacy profiles that contain their specified sensitivity values for a set of pre-defined topics, a social group for a focal user can be identified. Precisely, given a certain user u and some content s tagged with τ, we identify u's social group $SocG$ (see Definition 1), as indicated by the users' specification. Users may specify how to select a social group that they belong to, by joining existing groups (aka. networks), or by indicating their own. Subsequently, the similarity of the user u with the users in $SocG$ can be computed.

In order to infer policies based upon the user's social group information, we first compute the *similarity of profiles*, that is, the similarities between a user's profile and group members' profiles. Formally, we denote $sim(p_u, p_v) \in [0, 1]$ as the similarity between user u and v where p_u and p_v are the privacy profiles of user u and v, respectively. Cosine similarity in Equation 5.1 (or more complex Pearson correlation coefficient) can be used as the similarity function. We use cosine similarity mainly because of its simplicity. The similarity is commutative, i.e., $sim(p_u, p_v) = sim(p_v, p_u)$.

$$sim(\overrightarrow{p_i}, \overrightarrow{p_j}) = cos(\overrightarrow{p_i}, \overrightarrow{p_j}) = \frac{\overrightarrow{p_i} \cdot \overrightarrow{p_j}}{\mid \overrightarrow{p_i} \mid\mid \overrightarrow{p_j} \mid} \tag{5.1}$$

We then sort the similarity scores and identify the most similar user (i.e., the most similar reference point). To obtain the privacy policy for tag τ, we directly apply the policy existed in this reference point's profile. For example, if the reference point,

say Bob, has given the policy *pol* to tag τ, then policy *pol* is returned at the end of the social group analysis. This method can be generalized to consider top-k similar reference points. This generalized top-k method will increase the chance of locating tag τ in the reference points' profiles. We do not handle the situation when there are more than one reference points for the user in the application. In case the tag τ cannot be found in the top-k profiles, the aforementioned personalization and semantic analysis techniques can then be incorporated, which are not limited to the syntax of words.

Note that in order for the inference to be feasible, the users' profiles in the social groups must be already populated with content, and users must have posted tagged content. As such, there is a necessary training phase during which the users cannot enjoy the advantages of the social groups. This problem is well known in recommendation systems as the *cold start problem*. Essentially, the problem arises in case of lack of historical data to use for inferencing. To solve the cold start problem, our personalization approach with dynamic tag clustering can be used in combination with the social group analysis.

Chapter 6

Semantic Similarity Analysis

The WordNet based semantic similarity analysis among tags plays an important role in our framework *APPGen* personalization. The user-user similarity described in Chapter 4 is for comparing users' privacy profiles and utilizes well-known metrics presented in equation 5.1. In comparison, the semantic similarity of tags is more challenging and requires developing and evaluating new methods beyond the existing semantic analysis tools. Our semantic similarity analysis problem is as follows. *Given a tag τ associated with a new content, how to find the tags that are semantically most similar to τ among the tags associated with existing contents.* Once the most similar tags are located, our privacy policy inference method described in Chapter 5 can be used to derive the sensitivity score of the tag τ and thus privacy policy for the new content. This inference process does not require user's participation and is automated.

The building block of all our semantic analysis is the pair-wise word similarity metric. Given two words w_1 and w_2, a similarity metric computes the words' semantic similarity or relatedness $sim(w_1, w_2)$ based on certain measurement. There exist several proposals on how to measure the semantic similarity of two words, including Jiang-Conrath [23], Resnik [52], Lin [30], Banerjee-Pedersen [7], and Pirro'-Seco method [47]. All of these above-mentioned metrics use WordNet [60] ontology as the dictionary, which is a large lexical database of English. In [47], WordNet is described as a light weight lexical ontology where concepts are connected to each other by well defined types of relations. It employs IS-A inheritance relation between words in its structure. It contains similar word sets known as *synsets*. Further details of the index and WordNet similarity can be found in following literature [52, 47]. An online WordNet similarity tool implementing several measures is available [45]. Similarity metrics

between concepts (words) can be divided into four general and not disjoint categories [47]: Ontology based approaches, Corpus based approaches, Information theoretic and Dictionary based approaches.

Pirro'-Seco metric [47] uses information theoretic approach to get similarity using WordNet ontology. This approach employs a notion of Information Content (IC), which can be considered a measure that quantifies the amount of information a concept expresses. The IC value is calculated by the Equation 6.1 by considering negative log likelihood.

$$IC(c) = -\log p(c) \tag{6.1}$$

In Equation 6.1 c is a concept in WordNet and $p(c)$ is the probability of encountering c in a given corpus. Intuition behind using negative likelihood is that, infrequent words are more informative than frequent ones. According to Resnik, similarity depends on the amount of information two concepts have in common, which is given by Most Specific Common Abstraction *(msca)* that subsumes both concepts. Pirro'-Seco metric is calcualted using the IC and the msca values and it generates a value between 0 and 1 for given two words/concepts. Equation 6.2 shows the formula to calcualte the similarity. In our implementation, we evaluate different similarity metrics with the focus on the most recent approach by Pirro' and Seco [47]. Their metric has been demonstrated to have good prediction accuracy by human users. We perform analysis for this metric on both our approaches of finding semantic similarity namely, static classification of tags and dynamic tag clustering using 914 tags retrieved from Flickr by human judge.

$$sim_{P\&S} = \begin{cases} 3IC(msca(c1, c2)) - IC(c1) - IC(c2) & \text{if } c1 \neq c2 \\ 1 & \text{if } c1 = c2 \end{cases} \tag{6.2}$$

6.1 Static Classification of Tags

As described earlier, the static classification of a tag involves assigning the tag to one (or more) pre-defined topics based on the computed semantic similarity of the tag-topic

pairs – the topic that is semantically most similar to the tag is chosen. Then, based on the chosen topic, we can derive an appropriate policy for the tag. To evaluate whether semantic similarity measures can be used to map a tag to one of the pre-defined topics, we manually choose a set of topics (20), each representing a general category. The topics are *alcohol, adult, religion, schoolwork, sport, politics, news, business, culture, technology, gathering, food, animal, pet, people, travel, relationship, entertainment, nature*, and *family*. We obtain 1544 tags from Flickr.com using the Flickr API. The tags were from the most popular photos on August 29, 2008. Some of the tags are non-English words. We use WordNet to filter out these non-English words, by keeping the ones that can be found in WordNet. We further remove identical words, which leaves 914 distinct tags. Our evaluation procedure for static classification is given below. Counter values in our semantic similarity analysis performed by a human judge for both tag classification and tag clustering methods. The analysis is done on a total of 914 tags retrieved from Flickr.

6.1.1 Evaluation on Static Classification of Tags

We evaluate the Pirro'-Seco similarity metric on the aforementioned 914 Flickr tags [47]. The pair-wise semantic similarity is a numerical value between 0 and 1, with more similar words giving higher score. Our analysis is as follows.

1. For each tag τ_i and each topic γ_j, compute their semantic similarity $sim(\tau_i, \gamma_j)$ using Pirro'-Seco metric.

2. For each tag τ_i, sort the values $sim(\tau_i, \gamma_j)$ for all j from high to low; select the top three highest ranking topics and denote them as the set $\Omega = \{\gamma^1, \gamma^2, \gamma^3\}$.

3. For each tag τ_i, a human judge evaluates the following:

   (a) Semantic similarity of τ_i and the topics in Ω: If Ω contains at least one topic semantically similar to τ_i, then the human judge sets variable $counter1_i$ to 1, otherwise 0.

(b) How well topics are selected: If $counter1_i = 0$ and Γ (which is set of twenty predefined topics in privacy profile) contains at least one topic semantically similar to τ_i, then the human judge sets variable $counter2_i$ to 1, otherwise 0.

Then, we compute the sums $C_1 = \sum_i counter1_i$, and $C_2 = \sum_i counter2_i$, respectively in Table 6.1. If $counter1_i = 1$ or $counter2_i = 1$ for all i's, then each of the tags can find at least one topic that is semantically similar. For tag τ_i with nonzero $counter2_i$, value $counter1_i$ represents how well the semantic similarity measure is in finding the most similar topic(s).

Table 6.1 shows that classification correctly identifies the most suitable topic among the top-3 hits for 53% of tags. However, for 31% of the tags studied, none of three most similar topics returned by the Pirro'-Seco algorithm are considered related by the human judge. We also evaluate the tags using Jiang-Conrath [23], Resnik [52], and Lin [30] metrics, which do not provide significantly better results. We do not report the analysis results here. In tag-topic classification, the assignment is computed based on a *single similarity value* between the tag and the topic, which may not be accurate for certain words. In addition, static and arbitrary choice of topics limits the accuracy of finding the suitable topic for a given tag.

The static classification relies on a set of pre-defined topics and thus is limited in its ability of locating the most suitable topic for a given tag. For example, if the tag represents a new concept that is not yet incorporated by the topics, the static classification may give inaccurate result. To improve the classification of tags and to group similar tags with high accuracy, we utilize a new clustering method for words, which is presented and analyzed next.

6.2 Dynamic Clustering of Tags

To accommodate the dynamic aspect of folksonomies, we apply a machine learning technique, namely k-means clustering, to cluster tags based on their pair-wise semantic similarity. Dynamic classification of tags does not require pre-defined topics, instead,

the method needs a large number of tags as inputs. Given a new tag τ, the method outputs a cluster of tags that is semantically most similar to τ. Then, based on the cluster information, we can derive an appropriate policy for τ. We carry out a set of experiments to investigate whether we can *automatically* group tags into clusters, each of which may represent a topic. Reclustering the tags periodically may be necessary as the cluster size gets bigger to improve the fine granularity of categorization. Next, we briefly explain k-means clustering algorithm.

6.2.1 Discrete k-means algorithm

Integer k in k-means clustering specifies the number of clusters being sought. We do not attempt to find a generalized value of k in this algorithm. Once k is determined, k data points are chosen at random as cluster centers, and all instances are assigned to their nearest cluster center according to a certain distance metric, e.g., typical Euclidean distance. At the next iteration, the centroids, or the means of the points in each cluster are computed that are taken as the new cluster centers for their respective clusters. The iteration terminates until an equilibrium is reached, i.e., the cluster assignments stop changing. k-means algorithm is simple and finds a local minimal, i.e., with respect to the cluster centers, the total distance of the instances to their cluster centers is minimized. Algortihm is defined as follows.

1. Arbitrarily choose k tags to be cluster centers and denote them as $\tau_{c_1}, \ldots, \tau_{c_k}$. Denote the k clusters by $c_1, \ldots, c_k$.

2. **Cluster assignment** For each tag τ_i: Add tag τ_i to the nearest cluster c_j, $j \in [1, k]$ according to a distance metric defined as the inverse of $sim(\tau_i, \tau_{c_j})$.

3. **Cluster update** Choose the new cluster center as the tag that is closest to the centroid of the cluster. If the new cluster center is the same as the previous one, then an equilibrium is reached and the algorithm terminates. Otherwise, repeat from Step 2.

Static Tag Classification		Dynamic Tag Clustering	
C_1	C_2	C_X	C_Y
496	278	564	252

Table 6.1: Analysis of Semantic Similarity Index

Conventional k-means algorithm does not work for discrete objects, and only works for numerical data. In order to use k-means to cluster words, the cluster recenter step of the algorithm needs to be modified. Instead of choosing the cluster center (i.e., means) as the new cluster center, we choose the object (i.e., tag) that is closest to the centroid. We refer readers to machine learning literature for details about k-means clustering algorithm [41]. The evaluation on this classification is as described.

6.2.2 Evaluation on Dynamic Classification of Tags

To analyze the clustering quality, we let the same human judge (as in the previous static classification of tags) to manually look into each tag and count the number of tags that are semantically related to their cluster centers. The human judge reports the following two counters, $counterX$ and $counterY$, which are defined as follows. For each tag τ_i in a cluster c_j with center τ_{c_j}, if τ_i is semantically related to the cluster center τ_{c_j}, then $counterX_i = 1$, otherwise, 0. If $counterX_i = 0$ and there exists at least one cluster center (among the rest of 49 centers) that is semantically related to the tag τ_i, then $counterY_i = 1$. Then, we compute $C_X = \sum_{i=1}^{n} counterX_i$ and $C_Y = \sum_{i=1}^{n} counterY_i$. We compare the performance of clustering method with the static analysis in Table 6.1.

Table 6.1 shows that dynamic tag clustering gives better results than the static tag-topic classification. Both C_1 and C_X represent the number of correctly assigned tags (either into a topic or into a cluster of tags). Out of 914 tags, the human judge finds 68 more tags (8% more) that are properly assigned by clustering than by classification. Counters C_2 and C_Y represent the number of tags that are mis-assigned while there exists a different topic or a cluster to which the tag should belong. Clustering gives us 26 fewer such misclassification cases. We plan to extend this analysis to a larger scale in future.

Cluster	fruit	indian	motion
Tags	flower	american	play
	cinnamon	persian	bw
	nature	iranian	crossing
	hair	barrage	reentry
	whiskers	aussie	jump
	seed	czech	art
	beard	irish	morning
	shoot	cuban	tilt
	wool	chinese	flying
	saskatoon	creek	flight
	delicious	italian	surprise
	cane	european	reflection
	europa	chin	drop
	chameleon	russian	travel
	watermelon	japanese	flare
		inca	kill
		inka	laugh
		tongue	buzz .....

Table 6.2: Examples of cluster outputs

Privacy inference using clustered tags For our privacy inference purpose, clustering is done on existing tags of the user or his social group. Each of the existing tags is already associated with a sensitivity score as we defined in our framework in Section 4. Given a new tag associated with a new content, we need to decide (1) which cluster c^* this new tag τ^* belongs to, and (2) what is the inferred sensitivity score w^*. To locate c^*, we compute the average distance from τ^* to all members of a cluster τ_j and choose the cluster that gives the minimal distance value as in Equation 6.3, where $|c_i|$ is the size of cluster c_i. Then, the sensitivity value w^* is computed as the average sensitivity score of the cluster as in Equation 6.4 where w_{τ_j} is the sensitivity score of a member τ_j of cluster c^*. This clustering and new tag assignment operations can be updated and carried out dynamically.

$$c^* = argmin_{c_i} \sum_{\tau_j \in c_i} sim(\tau_j, \tau^*)/|c_i| \tag{6.3}$$

$$w^* = \sum_{\tau_j \in c^*} w_{\tau_j}/|c^*| \tag{6.4}$$

Our clustering analysis is run on the same set of 914 Flickr tags with k being 50 and the k-means running for 10 iterations. We have also experimented clustering runs with 30 and 50 iterations that produce different clusters with similar quality. Table 6.2 gives examples of cluster outputs. Compared to classification, clustering provides a holistic picture of pair-wise similar tags, rather than based on a single point of computation. The words grouped into one cluster must be similar to one another, thus creating a web of inter-connected words. As the inter-connectivity among tags are based on multiple similarity values, misclassifying a tag into a wrong cluster is less likely. For pre-defined topics, classification *solely* depends on single tag-topic similarity values, which is less robust. Therefore, *clustering is a more robust method for finding semantically related tags than assignment to pre-defined topics.*

Chapter 7

Implementation and Evaluation

This chapter descibes implementation and evaluation carried out on *APPGen*. Our goal was to examine the accuracy of the *APPGen* techniques, in inferring users' most appropriate privacy policies based on the input provided both at the time of registration and during the users' lifetime within the social network.

7.1 Experiment Setup and Methodology

The implementation of our prototype consists of a Web server and a backend database that run on a Fedora 8 Linux machine. We used Apache Tomcat 5.5.27 as the Web server to run JSP and servlets. We also used MySql 11.18 Distrib 3.23.58 for redhat-linux-gnu (i386) as the database. All the JSP and servlets are implemented in Java and HTML/CSS. For the WordNet similarity, we use the Pirro' and Seco implementation Java Library [47]. Clustering based inference uses 914 Flickr tags. Finally, we use `Surveymonkey.com` to host the survey. For simplicity, we assign all participants into the same arbitrary social group; the social group analysis is based on *the most similar user* among all the participants. In the setup of the clustering method, we assign synthetic sensitivity scores to 914 Flickr tags (See also Chapter 13).

As shown in Figure 7.1, the whole procedure is divided into following steps. First in user study, we asked each participant to register to a fictitious social network and to provide privacy preferences of 20 pre-defined topics (listed in Chapter 6.1) on a scale ranging from 0 (least sensitive) to 9 (most sensitive). This becomes a privacy profile for a user. Next, whenever user uploads any content on the site and annotates it, three privacy policies will be generated for that content using three different techniques (i.e., (1) *social group analysis*, (2) *personalization with static tag classification*,

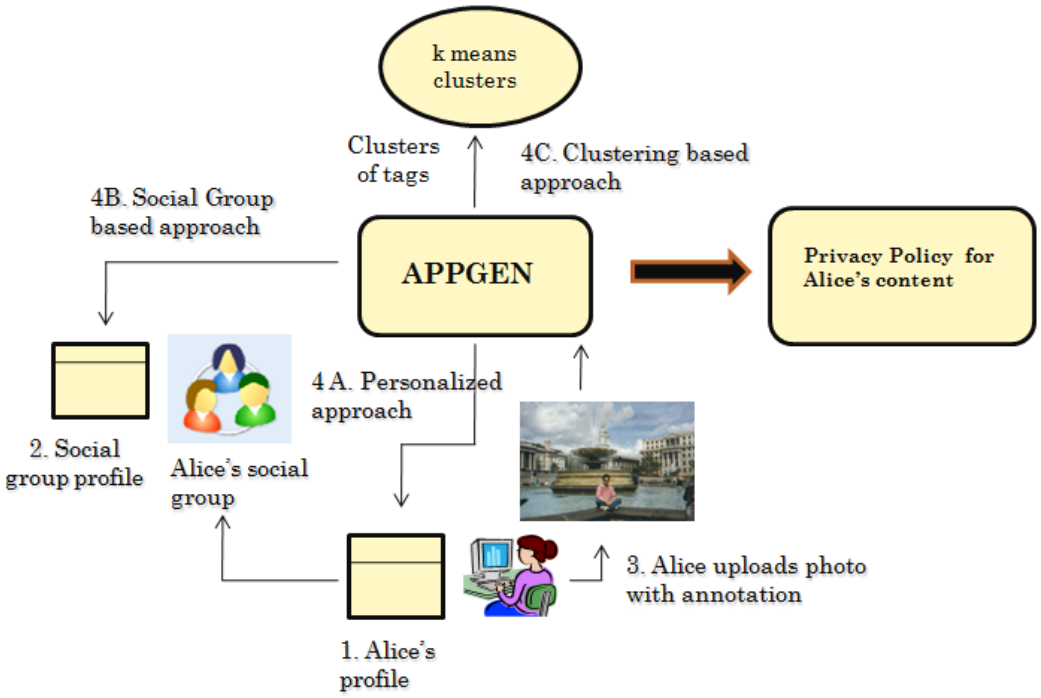

Figure 7.1: APPGEN System Architecture - Privacy Policy Generation

and (3) *personalization with dynamic tag clustering*). These techniques work with the use of privacy profile created by user and other users during the first step of the process. In user study, we ask users to tag three pictures (about cocktail party, traveling in London and drinking, respectively). The selected pictures had content that could be interpreted as sensitive. The tool, each time a picture is tagged, produces three types of policies based on our three privacy inference techniques shown in Figure 7.1. Specifically, the policies can include one or more of 10 pre-defined relationships, such as *Public, School/University Network, Friends of Friends, Local Community, Colleagues, Friends, Good Friends, Relatives, Best Friends*, and *Family*. We selected these groups as they reflect the most common relationships, and are general enough to summarize all possible relationships among social network users.

Participants were then asked to complete a post-session questionnaire. In order to evaluate the most effective technique we formulated questions using two different methodologies, namely *vertical comparison* and *horizontal comparison*. For the horizontal methodology we required each participant to evaluate *individually* each policy

generated by a specific technique. For each prompted policy we asked the participants three separate questions: to rate the overall perceived sensitivity of the picture, to state whether they thought it was a policy similar to their privacy inclinations, and to indicate whether the policy was appropriate for the content. The *vertical comparison* approach, instead, required the participants to compare the policies generated for the same picture. For each picture, we asked the participants to evaluate the three prompted policies and select the one that they perceived as the most adequate in terms of closeness with their thoughts, the most conservative in terms of privacy, and the most adequate with respect to the content. At the end of this procedure, the participants had to compile an exit questionnaire, where we asked some biographic information. Notice that an alternative design to the one described above would be to ask the participants' to manually input policies and compare them with the system's suggested ones. However, this approach is error-prone, as it depends on analysis of policies' similarity. Also, participants' would have the burden of commenting on their choices in order to make such approach effective.

Technique	Adequacy	Closeness
Social group analysis	19%	26%
Static tag classification	38%	26%
Dynamic tag clustering	43%	48%

Table 7.1: Adequacy and Closeness of Policy Generation

7.2 Experimental Results

Our initial sample consisted of 50 participants recruited using fliers. 15 participants had an age of under 20, 20 were aged between 20 and 25, and 15 were older than 25. Out of the 50 participants, 8 of them were not social network users. While 8 participants did not have their own blog the number of readers were higher, roughly 44 out of 50 participants declared they were blog readers, with varying degree of frequency. Data were discarded for all respondents who completed less than 80% of the tasks. For participants with modest amounts of missing data, we used a simple data imputation

method that has been found to be quite effective for factor analysis [14]. Specifically, we substituted item means (rounded to their integer value) for missing responses if a respondent omitted 1 item on a short scale (10 items or less) and up to 2 items on longer scales (more than 10 items). No imputation was used when 2 or more items were missing on short scales or 3 or more items were missing on long scales; rather, those participants were dropped from analyses involving these scales. Our final sample included answers of 42 participants.

7.2.1 Analysis Techniques and Participants' Preferences

According to the responses collected under the vertical comparison methodology, the policies were rated in terms of closeness with user's inclinations and adequacy of the policy with respect to the content. On a Likert scale from 1 (strongly agree) to 5 (strongly disagree) on the questions on both similarity, both personalization techniques (static tag classification and dynamic tag clustering) are rated equally well with a negligible difference, average 2.22 (agree) with standard deviation of 0.65 for static classification and 2.29 and sd=0.84 for dynamic clustering[1]. The policy returned by the social group method is rated at 2.5 (sd=0.721). Similar results were reported for the answers on adequacy of the policy with respect to the content. The lack of popularity for the social group technique can be motivated by the following considerations. First, in about 44 % of the cases, static tag classification produced a very similar policy to the social group technique. Users may select the static classification based policy for convenience, as it is listed at the top of the Web page (we did not scramble the ordering of policies when prompted to users). Second, we had to generate some synthetic data to bootstrap the social group technique. The synthetic data may have skewed the actual results, in that the randomly generated records may not be realistically significant for the similarity analysis.

The results from the horizontal comparison are reported in Table 7.1. Interestingly,

[1]The mean and standard deviation can only be calculated for interval and rational data. Many researchers argue that it is unclear whether Likert scales have interval properties. Nevertheless, Likert scales are often assumed to have interval properties (some researchers even refer to them as quasi-interval) and the mean and standard deviation are often reported[44].

the results do not exactly reflect the responses obtained using the vertical comparison. Thanks to this latter set of questions, we can clearly disambiguate the attitude of respondents' with respect to the prompted policies. When it comes to comparing the policies and select one policy over another, respondents preferred the clustering technique. As reported, in fact *the personalization with dynamic tag clustering technique outperforms the others, both in terms of perceived adequacy and closeness.* 94% of the participants voted the policy generated using the dynamic tag clustering technique as the best policy in terms of both accuracy and closeness with their privacy preferences. Votes were differentiated for policies generated by the other techniques, where the participants paired the answers about 80% of times. When they differentiated the answers, it was in most cases (90% of the cases) to indicate a more stringent policy as the most adequate one. On top of the analysis of above, we analyzed further our data, by running regression analysis for all three techniques. We used as independent variables age, social networks, and pictures' sensitivity as rated by the participants. These regression coefficients for our independent variable summarize the effects of the independent variable on the dependent variable when the effects of the other independent variables included in the regression analysis are controlled for or held constant.

The bivariate relationships obtained were inverse: as the sensitivity variable increased (that is, as participants perceived the pictures to be less sensitive) perceptions of policy generated by static tag classification decreased. So the policy generated with the static method was evaluated more positively when the pictures were more sensitive. The older people were, the less positively they evaluated the policy by static tag classification. Likely, this result can be justified by the fact that we noticed a tendency of *younger participants of rating the same pictures less sensitive than the elder observers.* Therefore, young users do not perceive stringent policies as useful. We report the results for this technique in Table 7.2. The table Coefficient gives results of the regression analysis while the model summary table reports the summary of results. In the Unstandardized Coefficients part of the Coefficient table, two statistics are reported: B, which is the regression coefficient, and the standard error. Notice that there are few statistics reported under B: one labeled as (Constant), age, soc_net, blog, sens_mod. These

Model	Unstandardized Coefficients		Standardized Coefficients	t	Sig
	B	StD Error	Beta		
(Constant)	3.645	.479		7.602	.000
age	-.428	.174	-.304	-.2455	.016
blog	-0.45	.106	.046	-.420	.675
soc_net	.102	.110	.118	.934	.353
sens_mod	-.197	.082	-.265	-2.418	.018

Table 7.2: Regression Analysis - personalization with static tag classification

statistics are the regression coefficients. The t-test (labeled as t) tests the significance of each b coefficient. The *sig* value indicates the confidence level. A *sig* below 0.05 indicates that the predictor is significant. No other predictors for the other techniques were found, although there are some clear tendencies for the clustering technique. The regression analysis showed that sensitivity of the picture is close to be a predictor variable (the significance variable is slightly below the threshold). The more sensitive is the picture, the more participants appreciated the policy. The lack of other significant predictors can be due to the relatively small sample size we had available. In light of the overall positive feedback obtained by the study, we interpret this as an encouraging sign. To certain extent, it implies that *no technical understanding of tags and blogs is required to appreciate our approach.* However, no stronger claims can be done at this time, and we reserve this investigation for future studies.

7.2.2 Increased privacy awareness

As part of our study, we asked users at the end of the experiment an overall opinion on this type of predictor tool and whether they thought this would be beneficial to them. The results obtained by these answers are extremely satisfactory, and clearly justify our efforts. *92 % of the respondents embraced the idea of a tool being able to adaptively provide privacy protection with little effort from the user end.* As this ideology is the goal of our *APPGen* framework, we feel that this outcome confirms our hypotheses of the need of *APPGen* type of tools. 86% of the respondents felt that tools like ours will increase their privacy awareness, and better protect their privacy. Finally, 83% of

the respondents expressed a positive opinion over the intention of using *APPGen* as a predictor tool for their current blog.

7.3 Summary

To summarize, we generate privacy policy by user annotations on the uploaded content. The *APPGen* system is based on WordNet ontology that is completely transparent from the users. We perform static and dynamic tag similarity analysis with limited training data available to us. The relationship-based policies are highly intuitive and the overall system is unobtrusive in policy generation. User study shows the feasibility of such systems. Using tags and social networks in Web 2.0, we demonstrate a new security application of social annotations based on ontology beyond conventional knowledge discovery and personalized information retrieval. Next, I present a Medical Imaging Search application which uses standard ontologies to search historical data and will conclude the thesis in Chapter 13.

Part - II

Visual Query Construction for Cross-Modal Medical Imaging Search

Chapter 8

Motivation for Ontology-based Medical Imaging Search

Today huge amounts of medical data is produced in hospitals and other clinical facilities every day. Despite the fact that this data is stored and made accessible electronically, searching for the content is not well supported. Radiological findings are kept separately from images which, in turn, are kept separately from patient accounting and billing information. Currently, these systems are more or less isolated from each other and do not allow queries to span across these systems. Today, such images can only be retrieved by attributes such as patient name, age or gender. However, these attributes do not contain any information about the anatomy or disease associated with the image. The research project THESEUS MEDICO [39] addresses these shortcomings by leveraging techniques from the Semantic Web to combine medical domain knowledge with image annotations in the same formalism. Within this project MEDICO ontology [38] hierarchy is developed which models various aspects of clinical data management and medical background knowledge. Storing the medical annotations as instances of well defined OWL (Web Ontology Language) classes [33]—rather than in a proprietary relational database—fosters an open interchange of this data with other applications and makes them easily available for other research goals, such as clinical data mining.

Many different medical imaging applications are in use today to view 3D volumes etc. to better visualize the humany body parts. Many of these applications use annotations sometimes from standard ontologies like RadLex [26], FMA [54] and ICD-10 [3] but are stored and retrieved in application specific way. There is no standard practice to this. The MEDICO ontology [38] provides a standardized way to store annotations in a semantic web format. This is another very important advantage of using it. However a fast and light-weight search interface is necessary to leverage the use of ontologies

and to retrieve images quickly even if the user is not aware of standard ontologies RadLex, FMA or ICD-10. We at Siemens Corporate Research built a search plug-in *FastMedSearch* that leverages MEDICO ontology based storage of annotations and searches with standard ontology concepts of RadLex, FMA, ICD-10 and patient metadata (DICOM standard). This plug-in can also be used with any other imaging application as long as the annotation storage is done in MEDICO ontology structure.

Existing approaches for semantic search often try to hide the complexity of the ontologies from the user. However, by using concepts and relations of the underlying ontologies only implicitly in the UI imposes a serious disadvantage. As a prior user study with clinical experts [40] showed, even experienced radiologists have difficulties with semantic annotation and search since they lack knowledge of what is modeled in the ontology. In contrast to that, our *FastMedSearch* addresses both inexperienced users as well as experts by providing an intuitive *visual query composition interface* combined with a powerful freetext query parser (Chapter 12). This parser transforms Lucene-like [1] queries into complex SPARQL [49] (SQL-like query language for RDF) queries which are used to retrieve search results from our central semantic data repository. Moreover, the plug-in provides an easy-to-use search-as-you-type textbox to write queries which is highly intuitive. Next Chapter 9 describes related work in this field, Chapter 10 briefly explains Semantic Web standards and the MEDICO ontology used. Chapter 11 introduces THESEUS MEDICO application in detail. Chapter 12 describes the implementation of search interface plug-in *FastMedSearch* and its functionalities. Chapter 13 concludes the thesis.

Chapter 9

Related Work in Medical Imaging Search

The need for representing high-level annotations of medical images on an abstract level has been emphasized in various publications in recent years, e.g., in [36, 62, 42]. [62] shows a semantic annotation and retrieval framework based on Error-Correcting Output Codes. It finds similarity between the images by matching the overall abstract labels generated with query image generated labels. Similarly, [42] also describes a framework that uses a hierarchy of concepts built for the system. Both of these rely on the generated concept set and do not use a standard ontology available for human body and diseases. Biomedical ontologies and terminologies received high attention in the last decade and they provide promising technologies. [11] evaluated popular large scale ontologies such as SNOMED (Systematized Nomenclature of Medicine - Clinical Terms), FMA (Foundational Model of Anatomy), and Gene Ontology and stated that "ontologies play an important role in biomedical research through a variety of applications". Besides efforts combining ontologies and radiological reports (e.g., [43]), other approaches using ontologies in medical image processing have been proposed [51, 58].

[50] describes a Support Vector Machine (SVM) classifier to extract low level features from the images to use it for image retrieval purpose. However, the system uses ImageCLEFMed dataset[1] in which the anatomy and disease are not associated in the same image metadata. Only recently, there has been work on creating an application ontology from RadLex and FMA [34]. This is done by incorporating subsets of the FMA into the organizational structure of RadLex. In contrast to this approach, we map RadLex to obtain an additional view to the FMA. This allows us to preserve the

[1] `http://ir.ohsu.edu/image/`

entire information from the FMA for automatic image/text annotation whenever necessary. And also this makes the search interface light weight and relatively fast in loading the ontologies.

In the area of visual semantic query composition some early prototypes are available online. Datao[2] provides a drag and drop interface for SPARQL queries but has slightly complicated interface for complex queries. Recent publications present different approaches for visually building SPARQL CONSTRUCT queries in client-side [37] as well as web-based applications [56]. These approaches are more generic than the one presented in our work as they address the use case of querying arbitrary data on the semantic web. In contrast, our approach is specifically directed towards the generation of queries with a more or less stable set of ontologies and querying for particular data structures.

[2] `http://datao.sourceforge.net`

Chapter 10

Ontological Modeling

Modern hospital information systems have become quite complex and in them Radiological findings are generally kept seperate from the images. Thus it has become challenging for clinicians to query relevant historical data for common cases [40]. Historical patient images can be very useful to them in understanding progression of abnormalities or may be recent trends of the disease for particular type of patients. These types of information can be very helpful in determining the cause and concerns regarding a particular disease. Currently in most applications search is carried out using DICOM metadata. DICOM [35] metadata headers of the image contain patient information such as patient id, age, gender etc. It does not contain any information about the anatomy and disease associated with the image. This makes searching the previous images very difficult and time consuming. Radiologists are often overwhelmed with irrelevant images not connected with the current examination or other extreme are unable to retrieve any similar cases. On top of this some applications do use annotations on the images but these annotations are stored and retrieved by application specific design and code. This creates a problem in sharing the annotation data and retrieval as well. So to store the annotation data in semantic web standards would be beneficial in many aspects. Semantic web standards describe a format of storage in OWL [33] and RDF [21].

10.1 Semantic Web

W3C has described Semantic Web as web of data. It is about assigning every data with a Uniform Resource Identifier (URI) that uniquely idenfies the object and all these objects are connected with each other. The relationships between objects help fetch chain

of data from the source. Notations such as Resource Description Framework (RDF), Resource Description Framework Schema (RDFS) and the Web Ontology Language (OWL) are intended to provide formal description of concepts, terms and relationships within a given knowledge domain. This representation of terms and their interrelationships is called an ontology [33]. *Classes* in OWL are subclasses of root class *owl:Thing*. A class may contain individual objects, which are a single instance of a class and it may also have subclasses. *Property* is a binary relation that specifies class characteristics. Object properties are the relation between instances of two classes while datatype properties are relations between instances of classes and XML schema datatypes. RDF is the underlying framework for OWL, but OWL is a stronger language with greater machine interpretability than RDF. It has a larger vocabulary than RDF. It has three sublanguages *OWL Lite, OWL DL (includes OWL Lite), and OWL Full (includes OWL DL)*. It is written in XML. Different notations in OWL using the RDF Schema are owl:class, rdfs:subClassOf, rdf:Property, rdfs:label etc. The medical ontologies for anatomy and diseases like RadLex, FMA and ICD-10 are all developed and maintained in OWL format since the hierarchy of classes and tree structure can be described easily with the available schema of OWL. We use these ontologies in OWL format in our search interface plug-in *FastMedSearch*. However, OWL is not expressive enough to describe the ontology for annotations on an image. Annotation on medical image should contain the location of annotation, the anatomy, the disease, the visual characteristic, the region, the patient metadata and other metadata of annotation creation. All these classes and relationships cannot be defined by the standard RDF Schema and so MEDICO ontology [38] is built for this purpose which I describe in next section. It was developed for the sole purpose of building a Semantic Web standard for medical imaging annotation data.

10.2 MEDICO Ontology

The MEDICO Ontology hierarchy consists of several components each modeling different aspects of the domain of our use case. It is structured across four different layers, based on the assumption that those elements at higher levels are more stable, shared

among more people, and thus change less often than those at lower levels. An extensive description of this ontology hierarchy can be found in [38]. Fig. 10.1 gives an illustration of the structure of a typical image annotation.

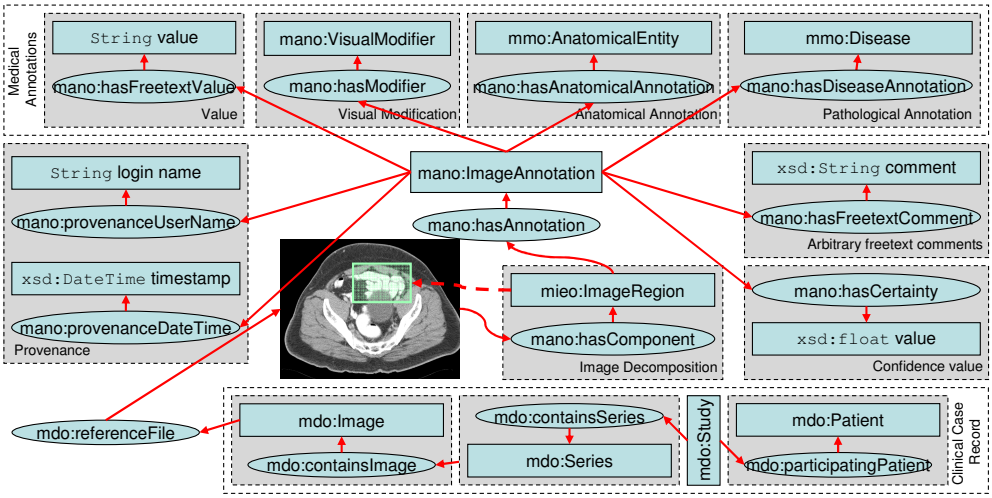

Figure 10.1: MEDICO ontology for annotation

The medical image in the center is decomposed into `ImageRegions`. These can then be annotated with `ImageAnnotations`. We differentiate between three medical aspects or *dimensions* of `ImageAnnotations`. For *anatomy* we use the RadLex and FMA. The concepts for the *visual manifestation* of an anatomical entity on an image is derived from the *modifier* and *imaging observation characteristic* sub-trees of RadLex. We consider the *disease* aspect as the interpretation of the combination of the previous two. Here we use the ICD-10. Additionally, a freetext value field can be used to save measurements, e.g., sizes of volumes. Provenance data is stored for the user (currently we use the user's login name) and timestamps. Additional comments can be saved using the property `hasFreetextComment`. This makes sure that annotations which cannot yet be expressed using concepts from the ontology can at least be stored in an informal way and do not get lost.

Additionally, the user can specify a continuous confidence value from the range from 0 to 1 to express his certainty about the actual correctness of each annotation. This can also be used to store the confidence values of automatic object recognition algorithms which we also plan to integrate. Unlike normal photos e.g., in JPEG format, medical

images usually contain a broad range of patient and image acquisition metadata in their file headers. The DICOM standard [35] is the most commonly accepted standard here for the interchange of digitized medical images. It provides a container format for data from different modalities such as X-ray, ultrasound, Computed Tomography (CT) etc. The MEDICO ontology also contains our own DICOM ontology which models the hierarchical data structure of the DICOM standard. This includes special annotations which control the automatic transformation of DICOM metadata into instances of classes of the MEDICO ontology [38].

Chapter 11

THESEUS MEDICO Application

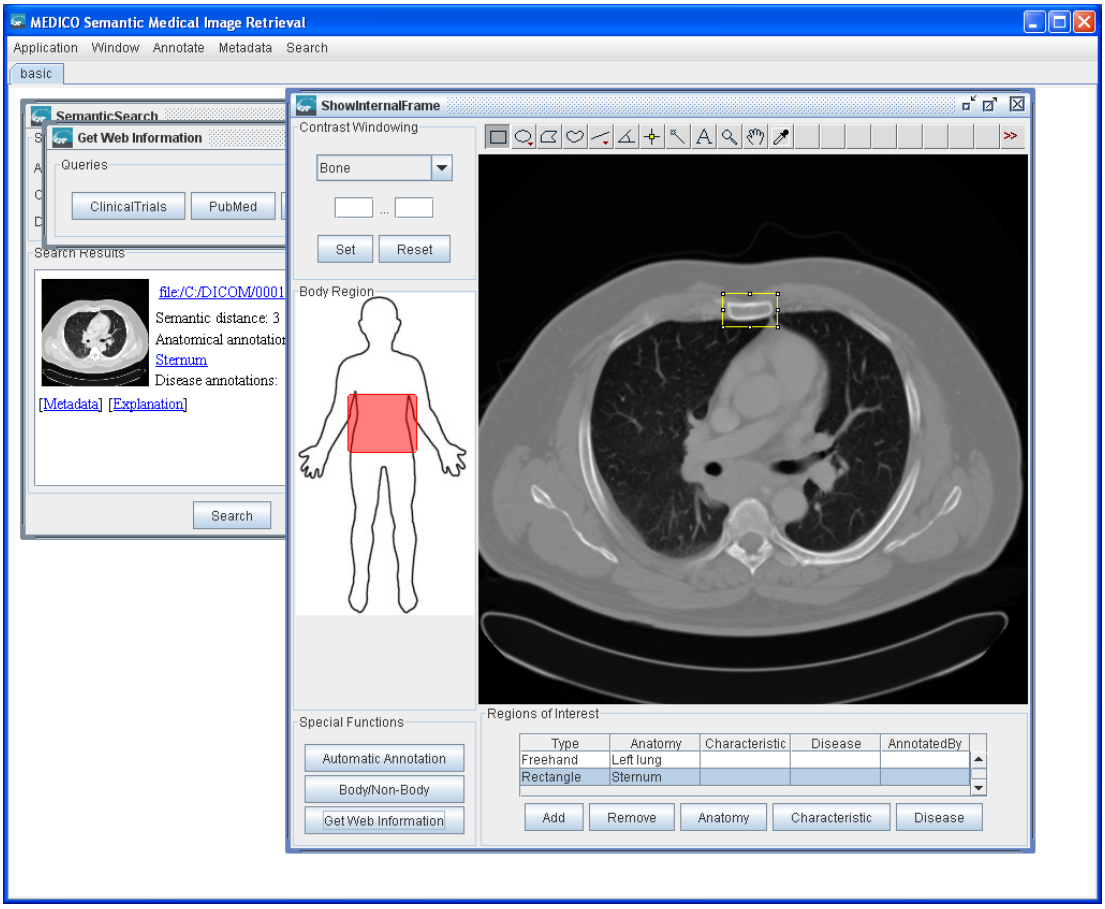

Figure 11.1: MEDICO Project Annotation Generation

THESEUS MEDICO is a German Government funded project [38] which is worked in collaboration with DFKI Institute, Germany and Siemens Corporate Research. It is basically built using the MEDICO ontology described in section 10.2. Figure 11.1 shows a user interface of the project. The main part of this interface is the two dimensional Image Viewer where the medical image can be opened. Next to the image viewer is the body region visualization that shows which body part image is shown in viewer. The

clinician can select a part of the image displayed using the drawing tools available to point to a certain part of the image where anomaly is seen. Every image region has annotation generated for it in the table seen under the image viewer. Every annotation has many parts in it as shown in the table. One annotation contains a field for anatomy, a field for disease, and a field for visual characteristic of the anatomy which are the main parts of an annotation. These are searched on standard medical ontologies created and maintained by medical community such as RadLex, FMA and ICD-10. Some other fields associated with an annotation are the user name, creation date and time and a confidence value for the annotation selected.

The annotations thus created are converted to OWL format (semantic web standard) and are stored in a central semantic data repository. The annotation data is converted into OWL format and stored using SPARQL [49] query language. SPARQL is an SQL type language made to work with OWL, RDF formats. It is very important to have a search interface on the annotations created on images for clinicians to easily get access to the historical data. This is as mentioned previously very important but it can be a time consuming process. For storing of the annotations MEDICO project uses a simple Lucene [1] based search-as-you-type interface. The initial user study done at the DFKI institute show that this is a time consuming and difficult process for someone who is not aware of the ontologies like RadLex, FMA, and ICD-10. The initial user study is described in following section. A search plug-in is developed which provides all the relevant features which ease the retrieval of images for inexperienced as well as expert users. Please note that this search plug-in is a plug-and-play application and so it is certainly not tied to the THESEUS MEDICO project. In fact, there are numerous applications available currently for medical imaging and annotations. The idea behind creation of this plug-in $FastMedSearch$ is to leverage the semantic web standards in annotations. Any application that will use the annotations data in this standard format can just plug this search interface into it. This way eventually all the annotations can be stored at a central location and a search on this will provide relevant historical data required by clinicians. The search plug-in details are explained in next chapter.

11.1 Initial User Study

In this section I present the results and feedback from a user study conducted at DFKI Institute, Germany *without* any visualization of the ontology contents. These results led to the RadLex FMA mapping and the implementation of the visual query composer presented in the next chapter. As the application so far had been developed by computer scientists it needed an external evaluation by the target audience and its medical assumptions were to be validated by medical experts. Therefore an evaluation was done in collaboration with a radiologist of the University Hospital of Erlangen.

In general, the application proved to be suitable for the semantic annotation of medical images with controlled vocabulary from formal ontologies. The majority of the clinical findings could be annotated. However, several shortcomings were noted. Choosing concepts only by relying on auto-completing combo boxes proved to be unsuitable for annotation and search. The reason for this is that it requires the radiologist to know: 1. which concepts are modeled in the ontologies like RadLex, FMA and ICD-10, and 2. the word sequence of the concept's label, which is standardized in the ontology but not in the everyday vocabulary of radiologists. Radiologists tend to use a context-dependent specificity for terms. If they are noting down findings for a particular disease or body region they tend to use terms which can have a different meaning in the context of another disease or body region. Due to this fact it is questionable whether the general-purpose FMA is able to reflect the radiologist's workflow at all.

Chapter 12

Search Interface Implementation

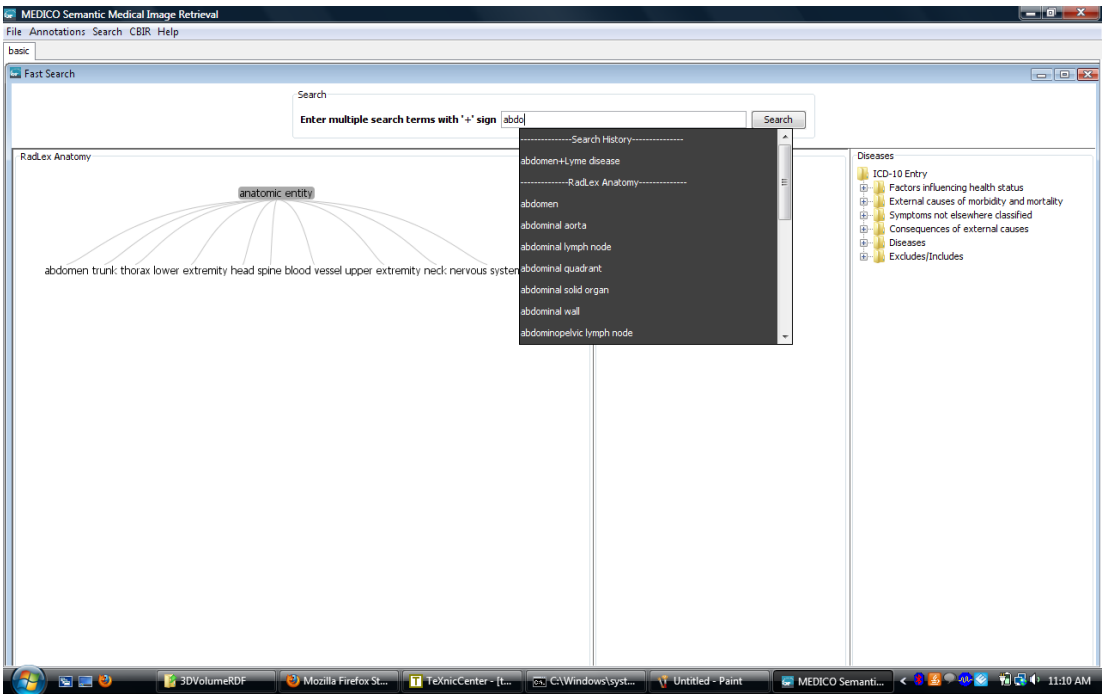

Figure 12.1: Search Interface of *FastMedSearch*

The search interface is a plug and play, single frame search interface desktop application. The header part contains a search-as-you-type text field created using the Apache Lucene [1] library with the annotations from RadLex and ICD-10 ontologies. In the text field provided, ICD-10 concept and Visual Characteristic can also be included in a single query with anatomy from RadLex using a simple '+' sign. The suggestions are provided as the user types in some annotation. The suggestion box also contains valid separators between Anatomy, Disease and Visual Characteristic values so that a correct concept can be selected without any confusion. The text field can also be

used to add DICOM Metadata fields into the same query again using the '+' sign between different concepts. DICOM metadata as explained previously contains patient data like age, gender, image type, name etc. These fields can also be given a specific value in single query e.g., a search query can be built using anatomy *abdomen*, disease *non-hodgkin's lymphoma*, visual characteristic *enlarged* and for all patients aged greater than 25. Mathematical notations '>', '<', '=' can be applied in such scenarios. The similar query can be built for specific date (creation date of annotation). To build a range query, 'TO' keyword can be used with two specific values entered in single query. In the backend this query is mapped to SPARQL queries which operate on the RDF [21] representation of the stored annotations. A *Search Log* is maintained at the client-side to help users see previous searches. This helps them select the whole query directly without typing it again.

12.1 Visualization

The search interface $FastMedSearch$ provides three visualizations on the frame. These visualizations are generated using prefuse API [2] from the OWL representation of the ontologies RadLex and ICD-10. RadLex and ICD-10 are created in OWL formats using protege toolkit [4] and made available to the desktop application by synchronizing it with the server everytime application starts. The updates to these ontologies are also possible from the client side which is explained in Section 12.3. These visualizations show Anatomy tree (part of RadLex ontology), Visual Characteristics tree (part of RadLex ontology) and Disease tree (part of ICD-10 ontology). Whenever user types and selects a particular concept in the search combobox, that particular concept is found in a relevant tree and the tree is expanded to that level automatically. This is very helpful in visualizing the annotation selected in standard hierarchy. Inexperienced users can also just navigate the tree without typing the exact term and use it in search. This way they will learn eventually the standard ontology used for medical imaging. Expert users can easily type in the concept without any delay.

12.2 RadLex FMA Mapping

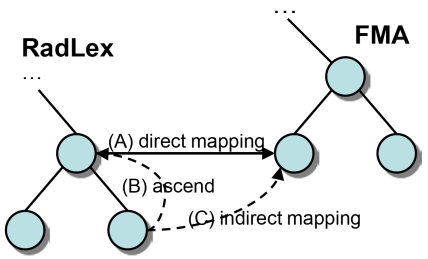

Figure 12.2: RadLex to FMA Mapping Approach

For the annotation of anatomical concepts on medical images there exist different standardized terminologies and ontologies. Two of the most prominent are RadLex, the Radiology Lexicon, and the FMA (Foundational Model of Anatomy). We use *RadLex*, which is maintained by the Radiological Society of North America[1], as the central terminology for annotating anatomical concepts. RadLex is designed to be a "Lexicon for Uniform Indexing and Retrieval of Radiology Resources" and its goal is to define a semi-formal vocabulary of terms which can be found on or related to results of radiological examinations. We used the version available in April 2009. The FMA provides us with a comprehensive source of formal knowledge about human anatomy. It has a rich representation ranging from the whole body down to macromolecules. It covers about 80,000 anatomical entities and over 2.1 million relationship instances from 168 relationship types which link the various classes together.

Both have individual advantages and drawbacks. RadLex is closely oriented on the needs of radiological practice and therefore most suitable for the audience targeted with our application. On the other hand, with only little more than 5,000 different anatomical terms it is far less comprehensive than the FMA which has more than ten times as many anatomical concepts and both more relationship types as well as instances defined between these classes. However, the comprehensiveness of the FMA introduces scalability problems when integrated into the UI for search-as-you-type and even more for an interactive graph visualization of its structure. Additionally, as we have learned during our user study, radiologists are significantly slowed down during

[1] `http://www.rsna.org/`

RadLex anatomy terms	5131	100 %
Lucene phrase matches	2665	51.9 %
matched via ascending the RadLex hierarchy	2412	47.9 %
never matched terms	54	0.2 %

Table 12.1: Results for RadLex - FMA Mapping

annotation when confronted with almost 80,000 different concepts for each anatomical annotation. Thus, the aim of our RadLex FMA mapping was to combine the strengths of RadLex (oriented at radiological practice and vocabulary, lightweight structure) with the comprehensiveness of the FMA. At the same time, this combination should avoid scalability issues and cluttering the user interface with too many different possible annotation concepts. Therefore, we decided to add a RadLex "view" to the FMA. The goal was to present terms and hierarchy from RadLex to the user and map them internally to FMA concepts. This mapping would allow us to leverage the rich semantic modeling of the FMA, e.g., for query expansion.

For our mapping we used only the `anatomic entity` subtree of RadLex. Fig. 12.2 illustrates our general mapping approach. The algorithm is split up into two main steps. Firstly, terms and concept labels are mapped using string matching using the Lucene parser. This resulted in mappings for 2,665 terms (Fig.12.2 (A)). Secondly, for all remaining terms we ascended recursively in the RadLex hierarchy and searched for a RadLex term that already had a mapping to an FMA class from step 1 (Fig.12.2 (B)). We assumed, that all children without direct mappings to FMA classes can inherit this mapping (Fig.12.2 (C)). Parts (B) and (C) present a generalization step. For example, the RadLex term "alar part of nasalis muscle" is mapped indirectly via the RadLex term "muscle of face" to an FMA concept with the preferred name "Muscle of face". During search the query expansion based on the FMA hierarchy performs the inverse operation during retrieval: Each search concept gets expanded into the search concept itself and all children of it by descending in the hierarchy. Eventually, step 2 resulted in mappings for another 2,412 terms. Table 12.1 presents an overview of the mapping results.

A review of the remaining unmapped terms revealed that eight of the 54 terms are

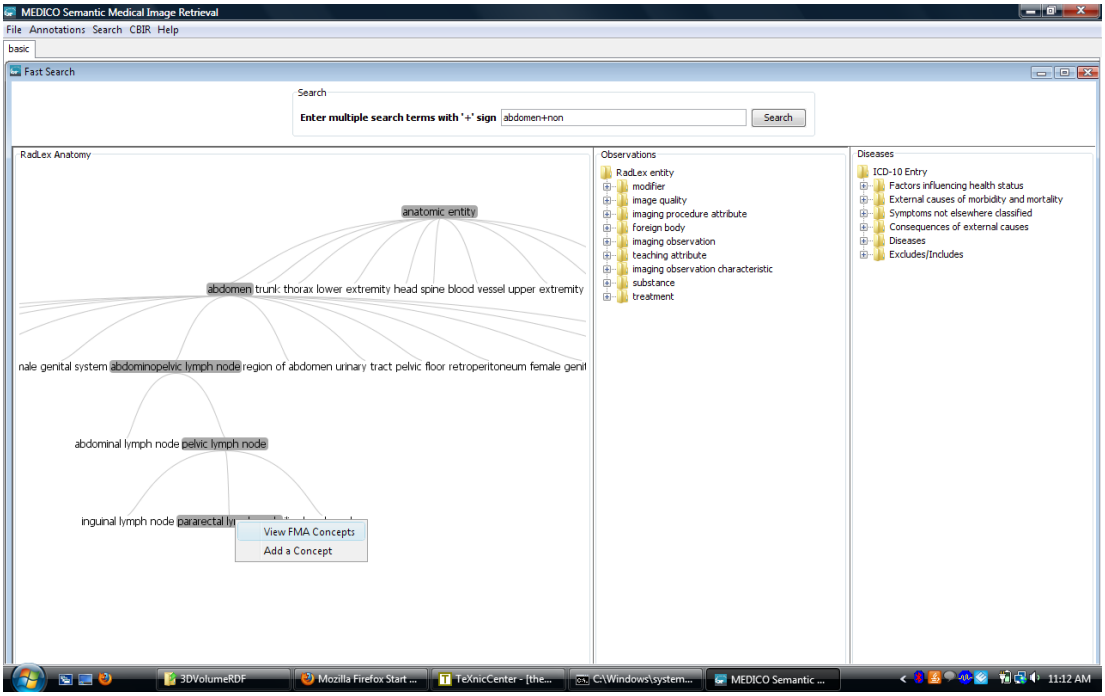

Figure 12.3: RadLex-FMA Mapping in *FastMedSearch*

non-anatomical terms such as `anatomy_metaclass`, `artery_metaclass` etc. Only the remaining 46 unmapped terms currently cannot be used for annotation.

12.3 Subtree Search

The search carried out using the query built by user is not just equality match in *FastMedSearch*. We carry out a subtree search algorithm to get all the images with annotations for the concept that was given in search query as well as all the images with annotations that are in subtree of the selected annotation. E.g., if the user has given anatomy concept as 'heart' then the search will also find images with annotation 'right ventricle' and 'left ventricle' in the search result. This is again very useful for inexperienced users who can provide a general concept in search query and get fine grained results of annotations. For the subtree search we wrote an algorithm that will assign a minimum and maximum index to all the terms of RadLex anatomy hierarchy. This is a one time process and is generally done in offline mode. This way when a concept is used in search, all the results between its minimum and maximum index are retrieved. This is a very simple way of retrieving subtree results.

We also have a facility to add a new concept in the current RadLex ontology if it is not available. This can be done by right clicking and adding a concept to the concept under which you want the new concept to be added. This process adds a new concept to RadLex hierarchy in server and other users can synchronize it when they start application again. This process of adding a new concept is problematic since we assign minimum and maximum index to all terms and now we are required to change the indexes for all the concepts from the place where this new concept was added. Since updating the hierarchy is not frequent for the RadLex hierarchy, we do a simple update of all the indexes when a new concept is added. This makes the process of adding a new concept very time consuming but it is acceptable in terms of the application usability. The whole process of finding a node and a subtree in the worst case will be O(n). The update of the indexes will also be an O(n) operation which is fast considering that the anatomy terms in the whole tree are very less. The data store updating is a bit more time consuming in updating the tree.

12.4 Summary

With the use of standard MEDICO ontology for medical image annotations, we develop a fast search plug-in to retrieve medical images. It is efficient and useful for clinicians to analyze disease or defect and also in for diagnostic purposes. Complex queries can be applied easily in it which is important for expert users with knowledge of standard ontologies for annotations such as RadLex, ICD-10. The visualization provided in the interface facilitates inexperienced users in understanding the ontology structure. Moreover, the functionality of RadLex FMA mapping, subtree search and adding a new concept to existing ontology gives a flexibility of use. Most importantly, $FastMedSearch$ can be used in any imaging annotation systems with semantic web based storage. The next chapter concludes the thesis with possible future work discussion.

Chapter 13

Conclusion and Future Work

The thesis presents two applications based on the underlying ontology to make the overall task easier for the end-user. The first application uses an English dictionary based WorNet ontology to infer privacy policy for the users. we utilize personal and social group annotations to develop automatic tools for managing content sharing. Our *APPGen* is a privacy policy generation framework that enables automatic generation of access control policies for users' contents. Our main approaches are to utilize static and dynamic semantic similarity analysis and social group structures for automatically inferring policies in content-based access control. We show the feasibility of our new approach by experimental evaluation and user study. Our privacy policy generation tool is definitely a step towards *Automatic Privacy Management systems.*

Although promising, our system has several limitations, that we plan to investigate in future. First, our approach on social group analysis needs to be refined to achieve its full potential. One may argue that similar users do not have similar privacy preferences. Hence, inferring from social groups may not always be accurate. We can further explore this issue by carrying out some comparative analysis between social groups that take into account users' privacy inclinations against groups that are purely based on other similarity features. Second, the semantic similarity analysis can be certainly improved which can be done using the Wikipedia based explicit semantic analysis by Gabrilovich and Markovitch [15]. Instead of using synthetic sensitivity scores for clustering, exploring use of social annotation and personal profile to infer the *average* sensitivity scores for clustered words and its impact on the score of the new tag should be studied. Also, how a selection of random policy affects experimental results needs to be measured. The semantic analysis could include multiple tags, rather than a single tag for picture.

Finally, users studies in larger scale are certainly desirable, to confirm our findings on a larger population. As part of this extension, it remains to be investigated whether, for legal purposes, certain levels of privacy are to be guaranteed, regardless of the user's actual input.

The second application is based on formal ontology using semantic web standards for retrieval of medical images. It makes extensive use of formal ontologies to connect (1) medical findings and (2) information about the patient. It demonstrates how distributed and disparate information sources can converge to a comprehensive medical information retrieval tool. Throughout the application the MEDICO ontology hierarchy is used as a common formalism to represent both medical expert knowledge as well as a model of the domain of application using RDF and OWL. This ontology hierarchy it tightly integrated with the user interface and used for all major tasks. We have generated a mapping which creates a RadLex "view" for the FMA. This view allows us to use terms and hierarchy of RadLex in the user interface which is well aligned with the needs of radiologists which are targeted by this research. To ease the task of creating complex search queries we have presented a solution which combines search-as-you-type, interactive ontology hierarchy visualizations and an intuitive query syntax to address the needs of both inexperienced users as well as experts. It allows the user to explore the structure of the ontologies interactively and at the same time construct queries by selecting concepts from the displayed hierarchies. On top of this the search interface can be applied to many other imaging applications available so that the annotation data can be shared easily and used because of the standard storage method.

This search application if used and shared between hospitals or clinicians can pose a privacy concern as it also contains patient metadata which gets shared. Since, the storage method is in semantic web standards, access control mechanisms can be easily applied on the data. Moreover, anonymization can also be considered useful. These are some of the aspects that should be considered further. In both the applications described here, the use of ontology in the framework adds a usability aspect into it. The use of such ontology in social networks for annotations purposes will be an interesting avenue of research. Currently, the annotations in Web 2.0 are free-form and does not

follow any pattern. It might be possible to create a small hierarchy with visualization by extracting terms from the blog and show users to help them annotate the content. However, a proper user study of this technique can reveal actual user preferences in using standardized annotations on Web.

References

[1] Apache lucene, apache foundation, 2008. `http://lucene.apache.org/java/docs/`.

[2] The prefuse visualization toolkit, berkeley institute of design, 2008. `http://prefuse.org/`.

[3] International classification of diseases, tenth revision (icd-10), center for disease control and prevention (cdc), 2009. `http://www.cdc.gov/nchs/icd.htm`.

[4] Stanford center for biomedical informatics research, stanford university school of medicine, 2009. `http://protege.stanford.edu/`.

[5] A. Acquisti and R. Gross. Imagined communities: Awareness, Information Sharing, and Privacy on the Facebook. In *In Proc. of Privacy Enhancing Technologies*, pages 36–58, 2006.

[6] A. Acquisti and J. Grossklags. Privacy and rationality in decision making. *IEEE Security and Privacy (Jan/Feb)*, pages 26–33, 2005.

[7] S. Banerjee and T. Pedersen. Extended gloss overlaps as a measure of semantic relatedness. In *In Proceedings of the Eighteenth International Joint Conference on Artificial Intelligence*, pages 805–810, 2003.

[8] S. Bao, X. Wu, B. Fei, G. Xue, Z. Su, and Y. Yu, editors. *Optimizing Web Search Using Social Annotations*, 2007.

[9] D. Beaver. 10 billion photos, October 2008. `http://www.facebook.com/note.php?note_id=30695603919`.

[10] M. Blaze, J. Feigenbaum, and A. D. Keromytis. KeyNote: Trust management for public-key infrastructures. In *Proceedings of Security Protocols International Workshop*, 1998.

[11] O. Bodenreider. Biomedical ontologies in action: role in knowledge management, data integration and decision support. In International Medical Informatics Association (IMIA), editor, *IMIA Yearbook 2008*, pages 67–79. Schattauer, 2008.

[12] B. Carminati, E. Ferrari, and A. Perego. Rule-based access control for social networks. In R. Meersman, Z. Tari, and P. Herrero, editors, *OTM Workshops (2)*, volume 4278 of *Lecture Notes in Computer Science*, pages 1734–1744. Springer, 2006.

[13] T. Finin, A. Joshi, L.Kagal, J. Niu, R. Sandhu, and B. ThuraiSingham, editors. *ROWLBAC - Representing Role Based Access Control in OWL*, 2008.

[14] C. Finkbeiner. Estimation of the multiple factor model when data are missing. pages 409–420, 1979.

[15] E. Gabrilovich and S. Markovitch. Computing semantic relatedness using wikipedia-based explicit semantic analysis. In M. M. Veloso, editor, *IJCAI*, pages 1606–1611, 2007.

[16] A. Gangemi and J. Euzenat, editors. *Knowledge Engineering: Practice and Patterns, 16th International Conference, EKAW 2008, Acitrezza, Italy, September 29 - October 2, 2008. Proceedings*, volume 5268 of *Lecture Notes in Computer Science*. Springer, 2008.

[17] C. Gates. Access control requirements for Web 2.0 Security and Privacy. In *IEEE Web 2.0 Privacy and Security Workshop*, 2007.

[18] K. K. Gollu, S. Saroiu, and A. Wolman. A social networking-based access control scheme for personal content. In *Proceedings of the 21st ACM Symposium on Operating Systems Principles (SOSP '07), Work-in-Progress Session*, 2007.

[19] M. Hart, C. Castille, R. Johnson, and A. Stent, editors. *Usable Privacy Controls for Blogs*, 2009.

[20] M. Hart, R. Johnson, and A. Stent. More content - less control: Access control in the web 2.0. In *In Proc. of Web 2.0 Security and Privacy (in conjunction with IEEE Symposium on Security and Privacy)*, 2007.

[21] P. Hayes. RDF Semantics. Recommendation, World Wide Web Consortium, February10 2004. See `http://www.w3.org/TR/2004/REC-rdf-mt-20040210/`.

[22] J. Huai, R. Chen, H.-W. Hon, Y. Liu, W.-Y. Ma, A. Tomkins, and X. Zhang, editors. *Proceedings of the 17th International Conference on World Wide Web, WWW 2008, Beijing, China, April 21-25, 2008*. ACM, 2008.

[23] J. Jiang and D. Conrath. In *Proceedings of ROCLING X*, Semantic similarity based on corpus statistics and lexical taxonomy.

[24] D. Jiménez, E. Ferretti, V. Vidal, P. Rosso, and C. F. Enguix. The influence of semantics in IR using LSI and K-means clustering techniques. In *Proc. of Workshop on Conceptual Information Retrieval and Clustering of Documents*, pages 286–291, 2003.

[25] A. Kulkarni and T. Pedersen. Senseclusters: Unsupervised clustering and labeling of similar contexts. In *ACL*. The Association for Computer Linguistics, 2005.

[26] C. P. Langlotz. Radlex: A new method for indexing online educational materials. *RadioGraphics*, 26:1595–1597, 2006.

[27] K. Lerman, A. Plangprasopchok, and C. Wong. Personalizing image search results on flickr. *CoRR*, abs/0704.1676, 2007.

[28] X. Li, L. Guo, and Y. E. Zhao. Tag-based social interest discovery. In Huai et al. [22], pages 675–684.

[29] Y. Li, D. McLean, Z. A. Bandar, J. D. O'Shea, and K. Crockett. Sentence similarity based on semantic nets and corpus statistics. *IEEE Transactions on Knowledge and Data Engineering*, 18(8):1138 – 1150, 2006.

[30] D. Lin. An information-theoretic definition of similarity. In *Proceedings of Conference on Machine Learning*, page 296 to 304, 1998.

[31] M. Mannan and P. C. van Oorschot. Privacy-enhanced sharing of personal content on the web. In Huai et al. [22], pages 487–496.

[32] A. Mathes. Folksonomies: cooperative classification and communication through shared metadata, 2004.

[33] D. L. McGuinness and F. van Harmelen. OWL Web Ontology Language overview. W3C recommendation, World Wide Web Consortium, February 2004.

[34] J. L. Mejino, D. L. Rubin, and J. F. Brinkley. FMA-RadLex: An application ontology of radiological anatomy derived from the foundational model of anatomy reference ontology. In *Proc. of AMIA Symposium*, pages 465–469, 2008.

[35] P. Mildenberger, M. Eichelberg, and E. Martin. Introduction to the DICOM standard. *European Radiology*, 12(4):920–927, April 2002.

[36] A. Mojsilovic, J. Gomes, and B. Rogowitz. Semantic-friendly indexing and querying of images based on the extraction of the objective semantic cues. *International Journal of Computer Vision*, 56:79–107, 2004.

[37] K. Möller, O. Ambrus, L. Josan, and S. Handschuh. A visual interface for building sparql queries in konduit. In C. Bizer and A. Joshi, editors, *International Semantic Web Conference (Posters & Demos)*, volume 401 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2008.

[38] M. Möller, S. Regel, and M. Sintek. Radsem: Semantic annotation and retrieval for medical images. In *Proc. of The 6th Annual European Semantic Web Conference (ESWC2009)*, June 2009.

[39] M. Möller, M. Sintek, P. Buitelaar, S. Mukherjee, X. S. Zhou, and J. Freund. Medical image understanding through the integration of cross-modal object recognition with formal domain knowledge. In *Proc. of HEALTHINF 2008*, volume 1, pages 134–141, Funchal, Madeira, Portugal, 2008.

[40] M. Möller, N. Vyas, M. Sintek, S. Regel, and S. Mukherjee. Visual query construction for cross-modal semantic retrieval of medical information. In *Malaysian Joint Conference on Artificial Intelligence (MJCAI), 14th July 16th July 2009*, 2009.

[41] A. W. Moore and D. Pelleg. X-means: Extending k-means with efficient estimation of the number of clusters. In *Proceedings of the Seventeenth International Conference on Machine Learning*, page 727734. Morgan Kaufmann, 2000.

[42] A. Mueen, R. Zainuddin, and M. S. Baba. Automatic multilevel medical image annotation and retrieval. *Journal of Digital Imaging*, 21(3):290–295, 2008.

[43] A. Mykowiecka, M. Marciniak, and T. Podsiadly-Marczykowska. 'data-driven' ontologies for an information extraction system from polish mammography reports. In *Proc. of the 10th International Protg Conference*, Budapest, Hungary, July 2007.

[44] E. J. S. Norm O'Rourke, Larry Hatcher. *A step-by-step approach to using SAS for univariate and multivariate statistics*. SAS.

[45] T. Pedersen. WordNet::Similarity. `http://www.d.umn.edu/~tpederse/similarity.html`.

[46] T. Pedersen and A. Kulkarni. Selecting the "right" number of senses based on clustering criterion functions. In *EACL*. The Association for Computer Linguistics, 2006.

[47] G. Pirro' and N. Seco. Design, implementation and evaluation of a new semantic similarity metric combining features and intrinsic information content. In *Proc. of On the Move to Meaningful Internet Systems*, 2008.

[48] A. Plangprasopchok and K. Lerman. Exploiting social annotation for automatic resource discovery. *CoRR*, abs/0704.1675, 2007.

[49] E. Prud'hommeaux and A. Seaborne. SPARQL query language for RDF. Technical report, W3C, March 2007.

[50] M. M. Rahman, B. Desai, and P. Bhattacharya. Supervised machine learning based medical image annotation and retrieval in imageclefmed 2005. In Gangemi and Euzenat [16], pages 692–701.

[51] D. Raicu, E. Varutbangkul, J. Furst, and S. Armato III. Modeling Semantics from Image Data: Opportunities from LIDC. *International Journal of Biomedical Engineering and Technology*, 2007.

[52] P. Resnik. Using information content to evaluate semantic similarity in a taxonomy. In *In Proceedings of the 14th International Joint Conference on Artificial Intelligence*, pages 448–453, 1995.

[53] D. Rosenblum. What anyone can know: The privacy risks of social networking sites. *IEEE Security and Privacy*, 5(3):40–49, 2007.

[54] C. Rosse and J. L. V. Mejino. *Anatomy Ontologies for Bioinformatics: Principles and Practice*, volume 6, chapter The Foundational Model of Anatomy Ontology, pages 59–117. Springer, December 2007.

[55] S. Sakurai, H. Tsutsui, and R. Orihara, editors. *Classification of Bloggers using Social Annotations*, 2009.

[56] P. R. Smart, A. Russell, D. Braines, Y. Kalfoglou, J. Bao, and N. R. Shadbolt. A visual approach to semantic query design using a web-based graphical query designer. In Gangemi and Euzenat [16], pages 275–291.

[57] R. Tamassia, D. Yao, and W. H. Winsborough. Role-based cascaded delegation. In *Proceedings of the ACM Symposium on Access Control Models and Technologies (SACMAT '04)*, pages 146 – 155. ACM Press, June 2004.

[58] L. Temal, M. Dojat, G. Kassel, and B. Gibaud. Towards an ontology for sharing medical images and regions of interest in neuroimaging. *J. of Biomedical Informatics*, 41(5):766–778, 2008.

[59] M. Vilares, F. J. Ribadas, and J. Vilares. Phrase Similarity through the Edit Distance. In *Database and Expert Systems Applications*, volume 3180 of *Lecture Notes in Computer Science*, pages 306–317. Springer, 2004.

[60] Wordnet - a lexical database for the English language. `http://wordnet.princeton.edu/`.

[61] X. Wu, L. Zhang, and Y. Yu. Exploring social annotations for the semantic web. In *WWW*, pages 417–426, 2006.

[62] J. Yao, S. Antani, R. Long, G. Thoma, and Z. Zhang. Automatic medical image annotation and retrieval using SECC. In *Proc. of 19th International Symposium on Computer-Based Medical Systems (CBMS 2006)*, Salt Lake City, June 2006.

[63] E. Zheleva and L. Getoor. To join or not to join: the illusion of privacy in social networks with mixed public and private user profiles. In *WWW '09: Proceedings of the 18th international conference on World wide web*, pages 531–540, New York, NY, USA, 2009. ACM.