UNIVERSITY OF CALIFORNIA,
IRVINE

Discovering Real-World Context to Tag Personal Photos

DISSERTATION

submitted in partial satisfaction of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

in Computer Science

by

Arjun Satish

Dissertation Committee:
Professor Ramesh Jain, Chair
Dr. Amarnath Gupta
Professor Nalini Venkatasubramanian
Professor Deva Ramanan
Professor Bill Tomlinson

2013

DEDICATION

And it's whispered that soon, if we all call the tune,
Then the piper will lead us to reason
And a new day will dawn for those who stand long
And the forest will echo with laughter

Jimmy Page & Robert Plant,
from Stairway to Heaven, 1970

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

CURRICULUM VITAE

Arjun Satish

EDUCATION

Doctor of Philosophy in Computer Science **2013**
University of California, Irvine *Irvine, CA*

MS in Computer Science **2011**
University of California, Irvine *Irvine, CA*

BE in Electronics and Communications **2002-2006**
Sir MV Institute of Technology *Bangalore, Karnataka, India*

RESEARCH EXPERIENCE

Graduate Research Assistant **2007–2013**
University of California, Irvine *Irvine, California*

Research Intern **2010**
Microsoft Research *Mountain View, California*

TEACHING EXPERIENCE

Teaching Assistant **2007–2013**
University of California, Irvine *Irvine, California*

SELECTED HONORS AND AWARDS

Google, Ph.D. Fellowship **2012**
Google, Inc.

Hitec Octane Entrepreneurship Competition **2008**
Second Prize *Irvine, California*

REFEREED CONFERENCE PUBLICATIONS

CueNet: a Context Discovery Framework to Tag Personal Photos **Mar 2013**
ICMR 2013

Visualizing Progressive Discovery **Mar 2013**
ICMR 2013

Tolkien: An Event Based Storytelling System **Aug 2009**
VLDB 2009

A New Approach for Adding Browser Functionality **Jun 2008**
Hypertext 2008

TECHNICAL REPORTS

Context Networks for Annotating Personal Media **May 2013**
ESL.UCI.EDU-TR 2013-May/01

Tagging Personal Photos Using Contextual Information **Aug 2012**
Whitepaper

Lives: A System for Creating Families of Multimedia Stories **Aug 2010**
MSR-TR-2011-65

Tolkien: Weaving Stories from Personal Media **Jun 2010**
ESL.UCI.EDU-TR 2010/04/11

SOFTWARE

CueNet `https://github.com/wicknicks/cuenet/`
Polyglot implementation of the CueNet ecosystem to tag faces in photos.

ABSTRACT OF THE DISSERTATION

Discovering Real-World Context to Tag Personal Photos

By

Arjun Satish

Doctor of Philosophy in Computer Science

University of California, Irvine, 2013

Professor Ramesh Jain, Chair

Automatic annotation algorithms assign one or more labels from a candidate search space to a given input object. The collection of these labels is either constructed manually, or derived from a set of sources. It is known that the accuracy of annotation algorithms is typically a decreasing function of the number of labels in the search space. In annotating real-world multimedia objects, this search space quickly explodes if not carefully curated. In this work, we present a technique to prune search spaces yet retain a very large number of correct tags for a multimedia object. Specifically, we shall address pruning of search spaces for the problem of tagging faces in the photo media.

Photos capture the momentary state of real-world entities. Any knowledge of the state of entities in the real-world at the time of photos capture can provide very useful context to eliminate candidates in the search space. For example, people in Japan cannot appear in photos taken at Paris, and photos taken at an academic conferences will largely contain people who work in the same field.

In the real world, relations between entities change rapidly. It is not possible to use a static model of entity relationships to reason over faces in photo instances taken across a wide range of time. Thus, it is imperative that any system first *discover* the set of relationships which are true at the time of photo capture before pruning the search space.

In this dissertation, we adopt a **relation-centric view** of real-world contextual information to design and implement a framework, **CueNet** to discover real-world context. The primary contribution of this dissertation is a **progressive discovery algorithm** to identify relationships between real-world entity relationships. We refer to such representations as **context networks**. We also present a **method to rank** entities in the case of incomplete contextual information. We shall extend the concepts behind the PageRank algorithm, used to rank webpages, to order real-world entities, and design score propagation techniques based on event semantics. We present **experiments** on both real-world photos and simulated events. We implement a system using CueNet to tag thousands of personal photos captured by multiple users, to study the efficacy of using context discovery in pruning search spaces. We also present simulation experiments to measure the time and space complexity of large scale discovery operations. We present the convergence characteristics and results on our real-world datasets to show its utility.

Finally, we conclude the dissertations by presenting challening problems for future research and ideas for new applications which can be designed using the foundations of context discovery techniques.

Chapter 1

Introduction

Annotating the content of unconstrained photos is a challenging problem. The primary complexity of photo annotation problems lie in their large search spaces and the diversity of feature-based representations of semantically similar images. Contextual information about a photo provides knowledge to prune this large search space, and simplify the task of feature-based classification techniques. The main intuition behind this work is that photos capture the state of the real world, and if we can reconstruct the various real-world entities and their relations at the time of photo-capture, this real-world context can prune the candidate search space. Ideally, the reduced search space will be orders of magnitude smaller, and will contain all the correct tags. In this work, we consider the problem of annotating faces in photos, where the search space consists of all possible people who can appear in a photograph.

Recently, the following changes in the technology climate motivate us to view real world information as context. First, with mobile phones becoming the primary mode of photo taking, the nature of context has evolved from providing cues about tags, to describing the world around a photo taking moment when a person was clicking the camera. Second, with the ever increasing amount of personal, social and public information, it is becoming harder

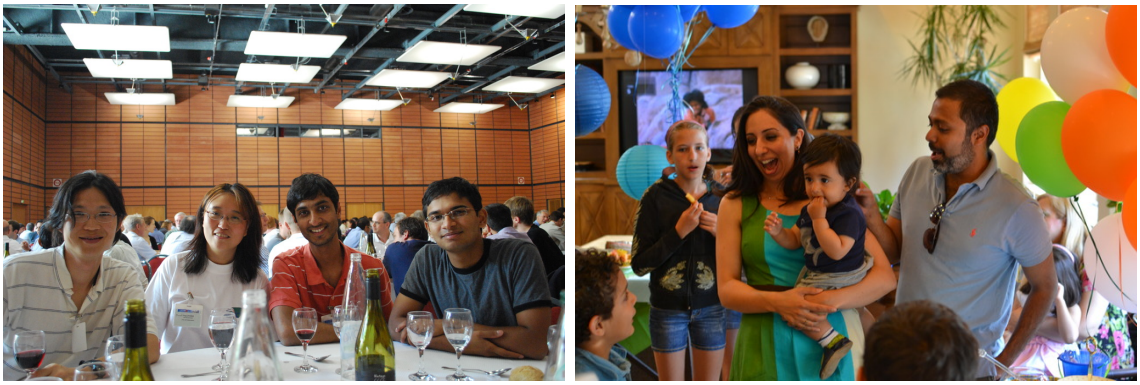

Figure 1.1: Photos can be captured at a variety of different events.

to specify which subset of these would constitute the most interesting context for a given picture. Thus, it is not clear what source will provide context, and how do we combine them to form models of the real world which will allow photo annotation algorithms to reason what tags to assign various regions in a given photo. Finally, a large number of photos to be tagged today are *personal* in nature. Tagging them is very different from tagging photos in datasets like LFW [62] or Pascal-VOC [46]. Photos are created alongside other media (tweets, calendar entries, webpages), and a tagging system must take advantage of their inter-relationships.

The most relevant context for a photo is contained in the relationships between the various real-world entities at the *time of photo capture*. Given the various data sources and sensors, each of which contain partial information about these entities, it is non-trivial to meaningfully combine this information to produce the context for a photo. In this dissertation, we construct computational representations of such dynamic real-world entities and their relationships from the partial information in various heterogeneous data sources. We refer to such a representation as the **Context Network** of the photo. The network describes real-world events occurring with the photo-capture event, the objects participating in them, and their semantic inter-relationships. Given such a network for a photo, we can reason which parts of the search space can be pruned, and which of the remaining candidates are most likely present in the photo.

1.1 Importance of Context Discovery

Annotation of multimedia artifacts such as images, audio and/or video clips [49, 84, 88, 104, 109, 113, 111] is a very active area of research. A technique to discover relevant context will play a large role in rendering such technologies more effective and allow them to operate in broader application domains.

Computer science is becoming largely utilized to address social [94], environmental [7] and economic real-world problems [33]. Social network analysis, data-driven diagnosis in medical sciences, philanthropic engineering, monitoring public interests through real time communication networks, and situation-based advertising are some of the emerging applications of computer science. The common requirement for all of them is to construct real-world context networks. A technology to construct such unified representations from various data sources available today will play a key role in the scope and architecture of such systems.

For the purposes of this dissertation, we propose context discovery techniques in the light of photo annotation problems, but the technology and ideas are not tightly coupled with any singly media, and can be ported to assist in solving any problem which requires models of real world information.

1.2 Context in Multimedia

Context has been used to address many multimedia problems [58, 73, 79, 82, 101]. For example, time and location information or social network information from Facebook to address the face recognition problem in personal photos. We refer to such a direct dependency between the search space and a data source as **static linking**. Although these systems are meritorious in their own right, they suffer from the following drawbacks: they are tightly

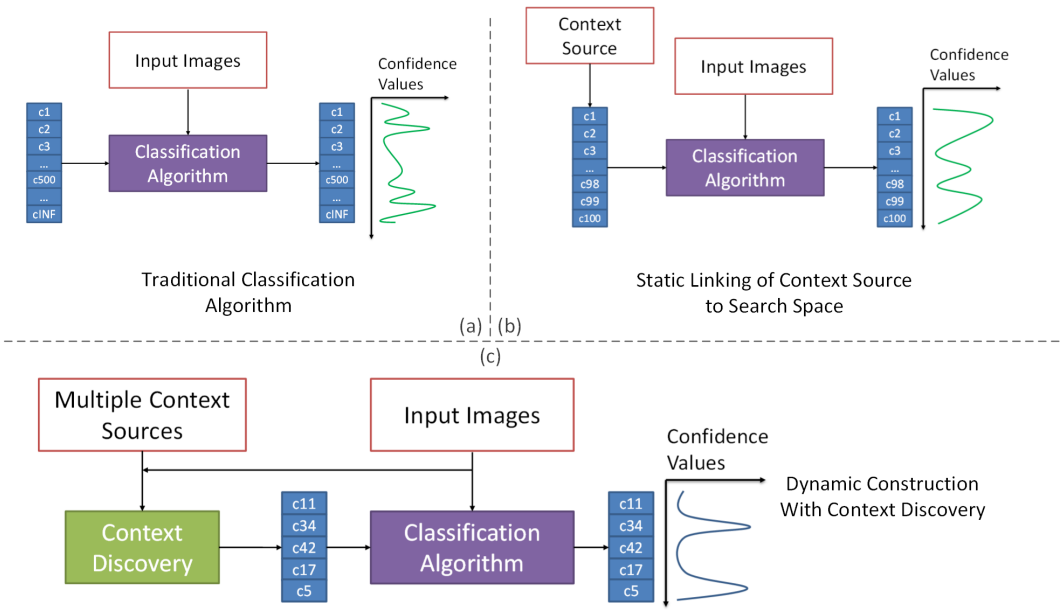

Figure 1.2: The different approaches in search space construction for a multimedia annotation problem.

coupled with a few data sources, the unavailability of any of which would reduce the efficiency of the system. They do not employ multiple sources synergistically, and therefore undervalue the **relations** between them. By realizing that these sources are interconnected in their own way, we are able to treat the entire source topology as a network, and traverse it to discover the relevant context.

Figure 1.2 shows three approaches to building photo annotation systems. Figure 1.2(a) is one of the earliest system setups where the search spaces were manually constructed, and the focus was mainly on constructing smarter features to correctly classify image regions [15, 102]. Figure 1.2(b) shows systems such as [101] which use a single source, to populate their search space based on some attributes of the input problem (in this case, the identity of the user to query Facebook to discover relevant candidates). This restricts the search space but it assumes that all relevant tags will be supplied from the context source, which may turn out to be an expensive restriction. In contrast, our approach **dynamically links** context sources to the photo. We assume the availabilty of a large number of sources and sensors, but extract context from only a subset of them for a photo taken by a known user.

Figure 1.2(c) shows how properties of input photo is used to reduce the search space. The properties are used to discover context, and prune the search space. This reduced space is considered by a face tagging algorithm to tag faces. Once the tagging is complete, any new information (new face tags) can be used to discover further context, and repeat the process till all the faces are tagged.

1.3 Approach

This primary contribution of this dissertation is a ***progressive discovery*** algorithm to ingest information from various real world data sources to construct context networks containing relevant information for pruning the search space of annotation problem. Examples of data sources include personal data sources such as personal calendar, mobile phone sensors (for example GPS which inform applications of the location of a person is present at any given point in time), personal emails, personal checkins on location services; social media web services to provide information about events and entities such as Facebook, Twitter; services which can be queried to find information about places such as Yelp; and finally public data sources such as Wikipedia, Upcoming, DBLP or websites on the world wide web which provide information about specific events (conferences or concerts).

In contrast to current static linking techniques, progressive discovery **dynamically links** context sources to the photo. It uses all available knowledge about a given problem (the input photo and related properties) to associate a subset of data from various sources as context. This allows us to decouple the technique of gathering context from the properties of the photo data. Because of this decoupling, there is no direct dependency on any specific set of sources, freeing the system developers to plug-and-play with various sources. With this decoupling, we are able to select different sources for different photos, and therefore discover only the relevant context for a given photo.

1.4 Examples

In this section, we present two examples of progressive discovery to demonstrate how different context is discovered for different photos. Figure 1.3 shows a photo of a person giving a presentation. Progressive discovery is done in three steps. First, the discovery algorithm proceeds to find the EXIF parameters and the camera device's owner with the photo. In our discussions we will use the term `photo-capture` event to signify the event in which a person clicks his/her camera and records a photograph. The event is assumed to contain spatio-temporal attributes which specify when and where the photo was taken. In the next step, these spatio-temporal attributes are used to discover *what events, if any, is this object (the owner) participating in, at this time*? Only those data sources are searched which can provide results to such a query (for example, facebook's social network will not be searched, as it does provide temporal information about people). We obtain a response from a conference database saying that the owner was attending the ICMR conference at Dallas, Texas. Third, given this new conference event, the algorithm discovers what conference subevents (keynotes, talks or break sessions) were occurring at this time. Finally, it finds that Mor Naaman and John Smith were speaker and host respectively, for the keynote talk going on that time. Given the two candidates, the face tagging algorithm proceeds to identify Mor Naaman as the person in the photo, as shown in figure 1.4.

Figure 1.3: Who is in this photo?

Figure 1.4: Mor Naaman at ICMR.

Now, let us attempt to discover context for photo shown in figure 1.5. Context discovery initiates the same way as in the above example, but after searching for events related to the owner in data sources, finds nothing. It proceeds to rank all known contacts according to location, and given that this photo was taken to the owner's workplace ranks colleagues higher than friends. The top 20 (an arbitrary constant) ranked candidates are passed to the face tagging algorithm which finds Ramesh Jain in the photo. But not the person to his right in the photo. Since the `photo-capture-event` has an additional participant, Ramesh, his personal information (calendar or social network) can be queried to find events in which he was participating. The calendar returns the entry "Kasturi". The algorithm uses this term to find all Ramesh's contacts to find all people with first or last name "Kasturi", and finds his long time friend and colleague "Rangachar Kasturi". The face tagging algorithm is invoked with one candidate. In this case, it is simply checking if Prof. Kasturi is present in the photo or not.

Figure 1.5: Who is in this photo?

Figure 1.6: Kasturi and Jain.

In this first run, the algorithm links only to a conference database, whereas in the second case, it used spatial information, personal calendar and contact information. For the first image, we did not link social networking or contact information to the image, and similarly, conference databases were not linked to the second image. This variation in linking to sources is an example of dynamic linking. In the later chapters, we will present techniques to represent and link context in a systematic manner.

1.5 Overview

This dissertation is organized into the following chapters. Chapter 2 provides an overview of context, how context has been used to address problems in various scientific disciplines and how we use context in our specific personal photo tagging application. Chapter 3 describes the related work in computer science, specifically the main ideas in photo annotation and tag propagation techniques, and presenting the techniques prevalent in context modeling and processing communities: how context is used to solve problems in different domains. Chapter 4 describes our context discovery framework, how it models various data sources, and how our progressive discovery algorithm constructs models for real world problems. We facilitate this discussion with an example real world application to tag faces of people in personal photos. Chapter 5 analyzes the algorithmic complexity of different parts of the system, and provides experiments to verify the competence and performance of the system. We also present experiments to confirm the efficacy of our approach in the light of the real world application. Chapter 6 presents a technique to use previously tagged photos and their context networks to rank candidates in context networks for new photos. Finally, chapter 7 describes the known issues of our approach and the future possibilities of using context discovery in various applications.

Chapter 2

Real World Context

The existing definitions of the word context are usually tied to the applications they are supporting. In the photo tagging problem, photos can be taken in a variety of situations, and appropriate context can be obtained from a multitude of data sources. Existing definitions of context do not provide enough clues as to what context is relevant when many data sources are present. Should we consider all available objects and relations across time? Or, does context consist of co-occurring events only? Or, the entities in a social network? Or, a combination of both? In this chapter, we look deeper into these definitions, and extract a relation-centric perspective of contextual information. This perspective will equip us to associate all relevant objects, irrespective of their type, as context, and at the same time, provide a reasoning framework to evaluate which source is beneficial and which is not.

Our justification for the use of context in a personal face tagging applications begins with the observation: *A photo captures the state of the real world at an instant of time.* Hypothetically, if there existed a database of all real-world entities, the various events occurring at different parts of the world, the roles objects played in these events, then we can simply query such a database to obtain the context for a photo. All the people who were physically closest

to the camera at that time are going to be in the photo, and those who were distant from it will not be present. Such a database does not exist, but with the World Wide Web and mobile phones being omnipresent in our daily lives, a number of data sources and sensors have emerged which capture different facets of the real world. To a point, where it can be argued that all their content can be semantically merged to construct our 'real-world entity' database. The accurate and scalable merging of all the data sources in a scalable manner is an extremely challenging one. Since we are tagging only a specific set of photos, we will restrict our problem to selectively merge content which is relevant to photos.

In the following sections, we will look at some previous definitions of context, and why they are insufficient to selectively merge the content of various data sources. We will present our relation-centric view of context to help reason about which data sources are relevant and which are not, for a given photo. Finally, we present the model of context for the personal photo tagging problem.

2.1 Previous Definitions

One of the earliest studies on context was reported by Bill Schilit et al. in [96]. The focus of this study was how to build software in dynamic environments. The dynamics of the environments were largely due to people requiring different computational services at the different times, the modality of request (through a mobile device or through a workstation), and the environment of the device (are there cameras and projectors nearby if the task requires video conferencing?). This software-centric view of context highlights the importance of two things. One, context is always described with respect to an object. In this case it is the software which runs on processors distributed in a real world environment. Second, context is used to determine how this object interacts with the entities near it? For example, Schilit uses the example that a workstation should automatically load his favorite text

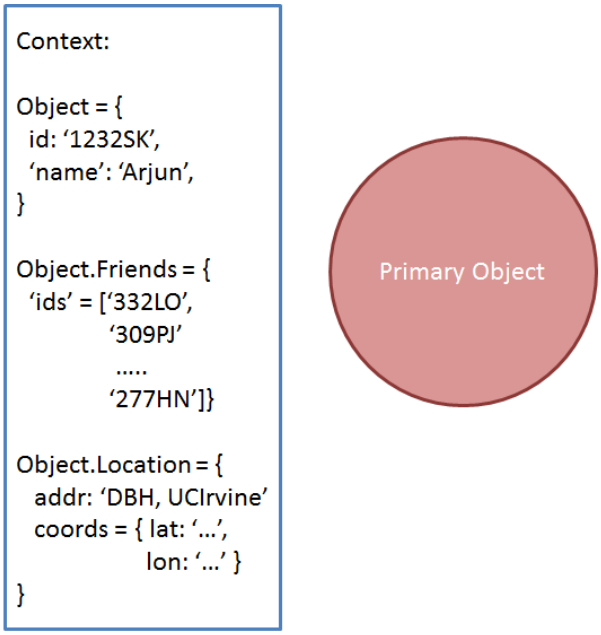

Figure 2.1: Information related to the situation of an object.

editor when he approaches it; and an rooster music sample must be played whenever fresh coffee is prepared. Both very different and precise interactions even though they might share common background (environment or participating entities). We would not expect a text editor to be shown when coffee is prepared, and the rooster music to be played when an employee walks to a workstation.

In his seminal paper, Anind Dey [38] describes context *as any information that can be used to characterize the situation of an entity. An entity is a person, place or object that is considered relevant to the interaction between a user and an application, including the user and applications themselves*, as shown in figure 2.1. He proceeds to explain this definition with the example of an "indoor mobile tour", arguing that there are there are two additional pieces of information which can be used: *weather* and *presence of other people*. if the user is present with his friends, they might visit sites that are of interest to everybody. There the presence of other people is important context. Because the tour is indoor, weather does not affect the application. It is true that the weather has no direct affect on the application but what about the following scenarios:

- Could we use the weather information to serve different drinks in the cafeteria to boost the experience of the visitors? On a cold day, placing the hot chocolate kiosk next to the entrance and the ice cream kiosk closer on a warmer day might boost some sales.

- If the tour is similar to Alcatraz, where a ferry ride takes people to the island, and back from it, a storm brewing in the ocean could lead to disrputed ferry services. Should the application warn its users who are liesurely touring at this time? Or should they continue the tour at the same pace, miss the last ferry and spend the night at Alcatraz?

They proceed to define Context-Aware computing as follows: *A system is context-aware if it uses context to provide relevant information and/or services to the user, where relevancy depends on the user's tasks.* But, we need to ask ourselves why a system which uses this "additional information" should be considered a context-aware system? There are numerous systems which would simply consider these "additional information" as regular inputs. What is different between a system which takes in these inputs and processes them as regular data, and one which processes them as context?

Karen Henricksen et al. [57] make the following interesting observation: Context information exhibits a range of temporal characteristics. Some context information can be static, for example the attributes of people using a system (for example, the date of birth of a person). But a large amount of information is dynamic. For example, the current geo position of a person or her social network, as shown in figure 2.2. There is no straightforward way to obtain this dynamic information other than through sensors. But, such a approach tightly couples the application logic to the types of sensors used, and requires the system to convert the input data to usable representations. For example, the application requires explicit modules to convert GPS coordinates to readable addresses. The problem with such an approach is that there are many ad-hoc modules built to tackle the sensors, and therefore causing the context-awareness to be tied to a specific application.

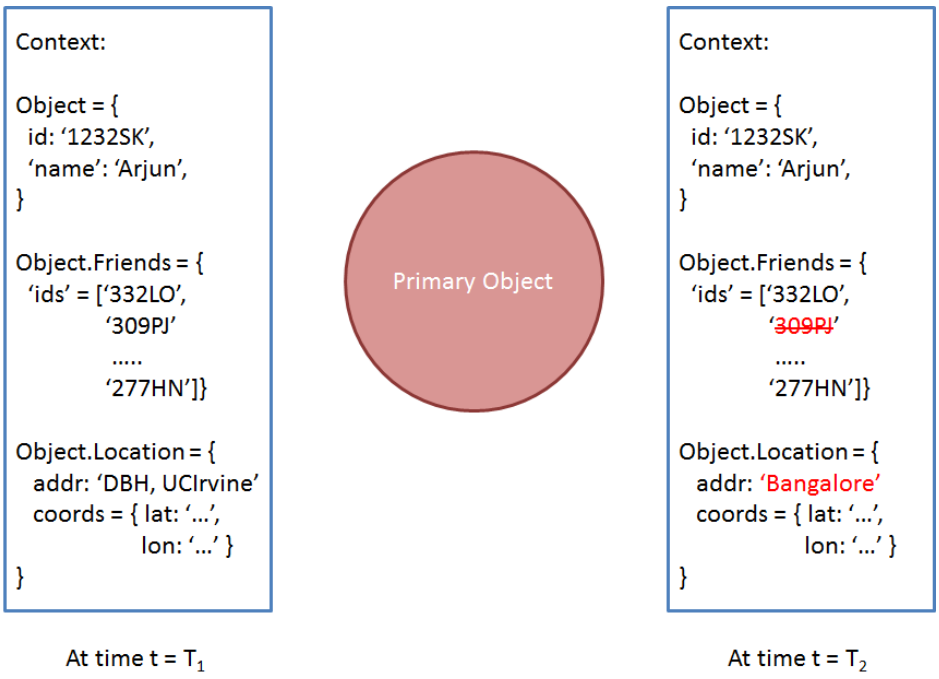

Figure 2.2: Henricksen's observation about temporality of Context.

More recently, Vaninha Vieira et al. [103] uses a rule centric view of context to design their context sensitive system, Cemantika. Vaninha defines a contextual element as any piece of data or information which can be used to characterize an entity in an application domain, whereas the context of an interaction between an agent and an application is the set of instantiated contextual elements that are necessary to support the task at hand. Context awareness, for them, is to explicitly switch the task the system is executing under different conditions. For this they explictly model the *context sources* which includes heterogenous and external sources like sensors, user dialog interfaces and databases. Figure 2.3 shows various data sources providing context. Some data sources are preferred over others depending on certain conditions pre-defined in the system. This allows the various processes to operate independently of the type of sources. It should be noted that the use of ontologies in describing knowledge and context sources is becoming increasingly popular (more similar systems are described in chapter 3).

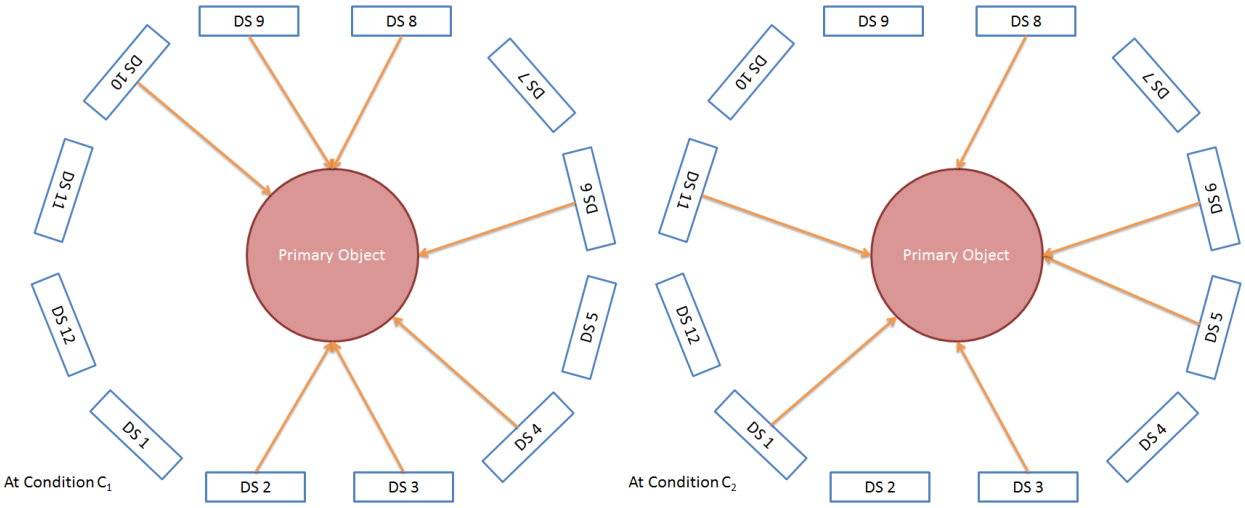

Figure 2.3: Modern context aware systems obtain data from different sources.

2.2 Relation-Centric View of Context

The common ground behind these definitions is their object centric view of context. Context is largely a set of objects that "surround" a primary object (whose context is in question). This view is insufficient while addressing applications which are broader in scope, the photo tagging application for example, where users can take photos in very diverse environment, and a large number of sensors and sources of context exist. Specifically, it is non-trivial to identify which subset of available data qualifies to be relevant context for the given photo, and which is not. The two examples from chapter 1, are shown in figures 2.5 and 2.6. In order to tag the photo on the left, we exclusively used conference schedule information. Whereas, to tag the photo on the right, we used personal information. Our motivation in this chapter is to extend the above definitions of context to allow context aware systems to better extract relevant context from these various sources.

Relations between entities change with time. The most relevant context for a photo are the relationships between entities at the time of photo capture. We define context for the photo as the **"set of relationships between real-world entities at the time of photo-capture"**. The entities include events, objects such as people, their pets or organizations

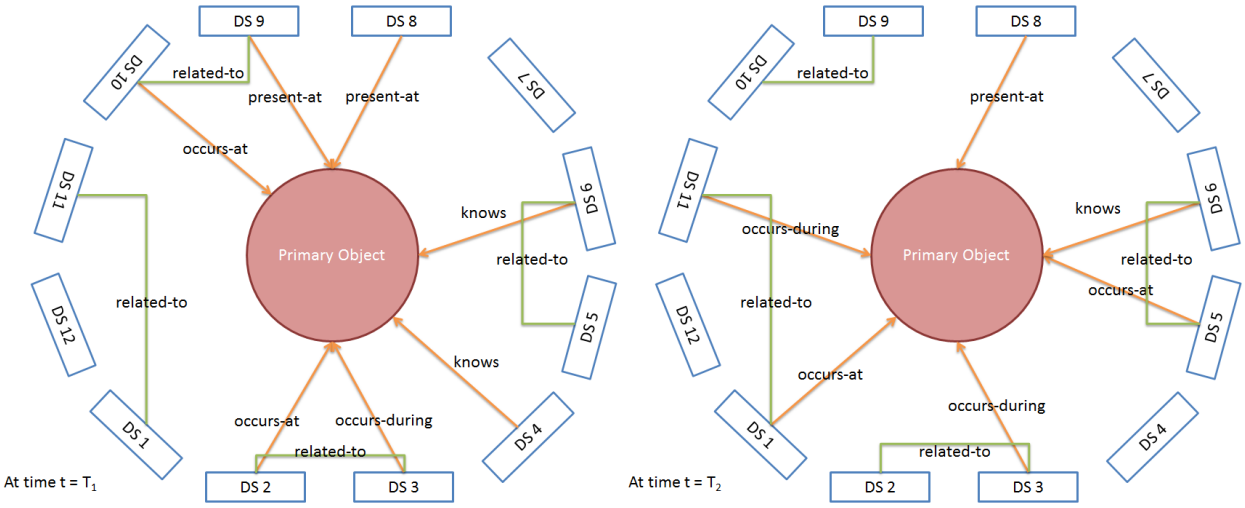

Figure 2.4: Utilizing relations to define context for the primary object.

they are affiliated to and places such as landmarks (Eiffel Tower, Niagra Falls), restaurants, stadiums or highways. This definition must not seen as tightly coupled to photos, but can be applied to any multimedia object (such as audio, videos or tweets) or any real-world object at a given time. Figure 2.4 shows the relations between the primary object and other real world objects contained in the sources change from time T_1 to T_2.

Relationships can be of different types. They can be simple labels such as `friend-of` or `father-of` signifying a social relationship. Or they can impose constraints with relations such as `located-at` or `participant-of`, which relates an object to a location or event respectively, and asserts a constraint on its spatial attribute. In this dissertation, we will see that such relations, which impose property constraints are critical in algorithmically determining which information is relevant context.

Using this relation centric view, we now look at how the examples from chapter 1 can be formulated in a systematic process to discover context. A system to construct context networks must establish a set of relationships to connect its nodes. For the two photos, we choose **participant-of**, which indicates that an object is a participant in an event, and **subevent-of** which indicates that an event is occurring within another super-event. Thus,

Figure 2.5: Mor Naaman at ICMR.

Figure 2.6: Kasturi and Jain.

any object can be related to other events through a `participant-of` edge, and events can be related to others events as well as entities through `subevent-of` and the `participant-of` edge respectively. Let us also assume that we have available to us the four types of data sources: event sources, place databases (like yelp.com), weather information sources and social networking information.

Figures 2.7 through 2.10 show how the two relationships can be used to gather context for the photo in figure 2.5. Figure 2.7 shows the initial graph created using the `photo-capture-event` and the photographer, `entity:ent81`. Figure 2.8 shows the result of adding context by associating an event with `entity:ent81`. Given this new graph with three nodes, a context aware system can find more context by trying to find objects which can be related through the two edges. Since one of the nodes is a `conference` event, it proceeds to find events occurring within it, and adds the keynote event, which also happens to be the super-event of the `photo-capture-event`. The result is shown in figure 2.9. Finally, the keynote event is extended with relations to associate subevents or participants. In this case, the only new context available are the participants of the keynote event. These two entities are associated with the event as shown in figure 2.10.

In the above example, given a graph containing primary objects, we grew it by relating objects from the real-world using a fixed set of relationships {`participant-of`, `subevent-of`}.

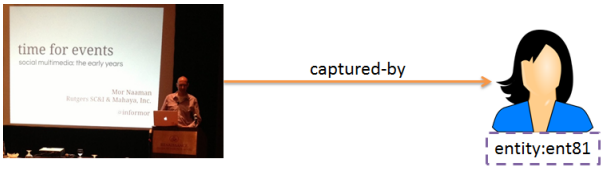

Figure 2.7: Primary objects.

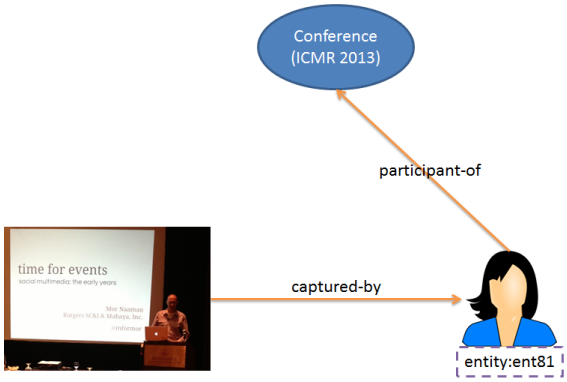

Figure 2.8: Associating `conference event` using the `participant-of` relation

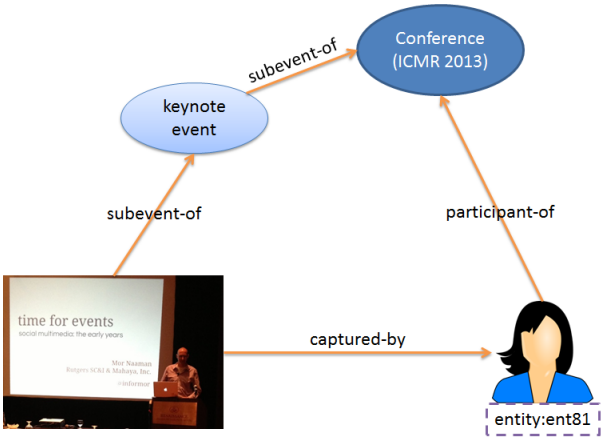

Figure 2.9: Associating the keynote event.

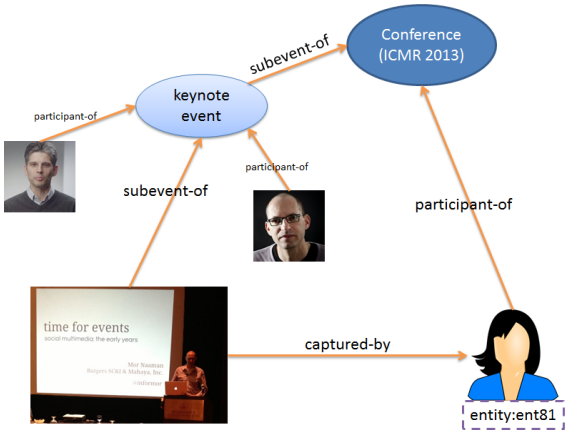

Figure 2.10: Associating the participants using the `participant-of` relation

No matter how many different sources or object types are included, we can use the relations and their semantics to reason which source is suitable, and extract context from it, if any, to construct and grow context networks.

Because of the type of relationships chosen, some information which was readily available (weather, place or social networking, for example) was not associated. But if we extend the relationship set to contain another relation `occurs-at`, then the place where the conference was held will be included in the context network. Similarly, the inclusion of a `friend-of` relation can relate entities with each other. If the social networking source reports that

`entity:ent81` was a friend of Mor Naaman, an additional edge would be introduced between these nodes in the context network. Thus, relations are key in determining which objects are context and which are not, and how they are related to the primary objects. Figure 2.11 shows one such context network for Mor Naaman's photo taken at ICMR 2013 with the additional `friend-of` edges.

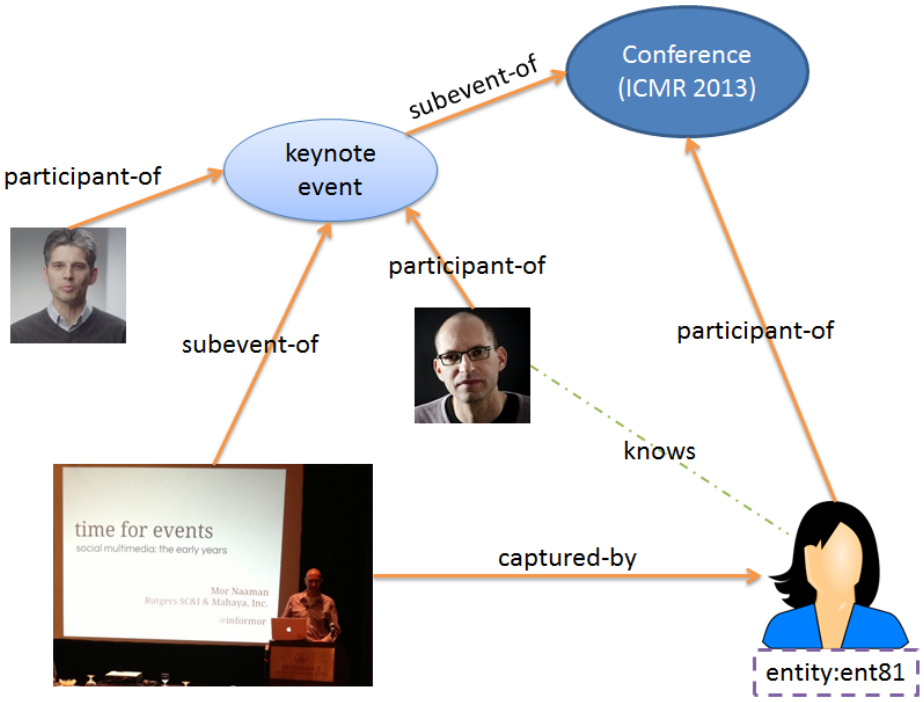

Figure 2.11: Context Network for Mor Naaman's photo with the additional `friend-of` edge.

It must be noted that asserting a relation-centric view of context is not discounting the importance of objects. Rather, it is a way of saying that relationships are atleast as important as objects are. Keeping this in mind, we will see what are the primitives required to model context for real-world problems in this section.

2.3 Modeling Context for Real-world Problems

Traditionally, context has been modeled for specific domains. Henricksen et al. [57] model context using context graphs to represent these models, and [92] provides a technique to con-

struct petri nets for given context graphs to implement context-aware behavior. Although, these techniques work for specific domain (for example, [92] presents a domain of class presentations, and [57] presents a case study in context-aware personal communications), creating context graphs **at system design time** for each and every case which might occur in the real world is impossible. Also, with the rapidly changing relations in the real-world, it is not clear what are the good principles and practices to build scalable systems around context graphs.

If we are to represent context networks computationally, we first need to understand what are the basic building blocks of such a network. The following sections list these blocks, and present their properties which are essential in constructing such networks. These blocks are fairly generic, and can be utilized to model context for various application.

2.3.1 Object Types and Semantics

Context is always specified with respect to a real world object at a given time. This **primary object** must be uniquely identifiable in the computational system, and must be an instance of one of the known classes. This object must have some real world attributes. For example, if the object is an event, then the interval during which it occurs, and location of occurrence are two real world attributes. In the tour guide application, the visitor is one possible primary object whose context needs to determined. In the photo application, possible types of interest include the `photo-capture-event`, the `photographer`, various possible events which can occur in the world, people and places where photography can occur.

Different entities bring with them very specific real-world semantics. While incorporating them into the context model, these semantics must be preserved. For example, an event exists only for a fixed time interval. People can only be present at one place at one point in time. The `photographer` is a person, and therefore inherits the property of being at only

one place at a time. In chapter 4, we will see that these semantic properties play a very important role in the context discovery algorithm.

2.3.2 Relationships and Constraints

Context is the set of relationships between real-world entities at a point in time. Instead of finding entity instances of a specific type, and qualifying them as context, in this work the relationship structure between entities is considered to be context. These relationships impose constraints on the instances they relate. For example, the subevent relation constraints the super-event to spatio-temporally contain its child event. Such constraints are very common in real-world relationships. Another example of a real-world relation can be seen in naturally occurring chemical reactions, where it is not sufficient to have just two reactants at the same time and place, but the reaction could demand very specific environmental factors (like temperature or pressure). Thus, asserting constraints in terms of environmental factors. The context modeling tool should permit the specification of such specific constraints.

The relationships and constraints are not black boxes. In our work we decouple the context discovery logic from the sources and type of entities, but cannot do the same with relationship types. Their semantics must be **declared** or **tightly-coupled** to the modules of a context-aware system, and must be used to reason which entities are part of a context network and which are not. Pushing our example of chemical reactions, given all the materials and conditions which can be present at a given region, the relationships and constraints between different materials must be 'machine-readable' to simulate the various ensuing reactions. Because we are fixing the number of relation types in our system, care must be taken to ensure that these relations are generic enough, at the same time are rich enough in terms of semantics and constraints.

In our application of photo tagging, social relations and presence of people in a particular

event are used to rule them out of other events. Why? Because, a person can be present only at one place at one point in time, and if the events are too far apart or contain a very distribution of people than what a person commonly co-participates with, the chances of that person being present in that photo is very low. Such a **participation** and **co-occurrence** relation, which are generic yet semantically rich, must be exclusively utilized in a context discovery algorithm for photos. More concretely, consider a 10 level deep class hierarchy of entities in an application's ontology. The relationships which relate instances of classes at the top of hierarchy are more generic than the ones below, but the relations which only relate instances at the bottow of the hierarchy are richer. For example the `participant` relation only asserts the presence of an object at an event. But the `keynote-event-host` is a very specific relation which can make more assertions about the two objects in question (a keynote event at a conference and the host here is an academic person who has obtained a PhD and continues doing research at research establishment). The drawback being that the relationship is used in very few objects, and therefore cannot be extensively used in reasoning about entities in different domains.

2.3.3 Temporal Semantics

Modeling context requires the ability to express relationships which assert temporal constraints. For example, during a lunch break at a conference, there are no co-occurring session events; the workshops at a banquet is always the last event at a conference, which could be followed by one or more workshops. Temporal relations have been studied in literature [9, 106], and can be reused for this purpose. Temporal relationships defined in Allen's Interval Algebra are shown in figure 2.12. Figure shows these relationships. In order to relate events with entities, we use the `occurs-during` relation defined in [53]. For example, the event X is said to `occur-during` the interval I iff X's temporal extent is equal-to I.

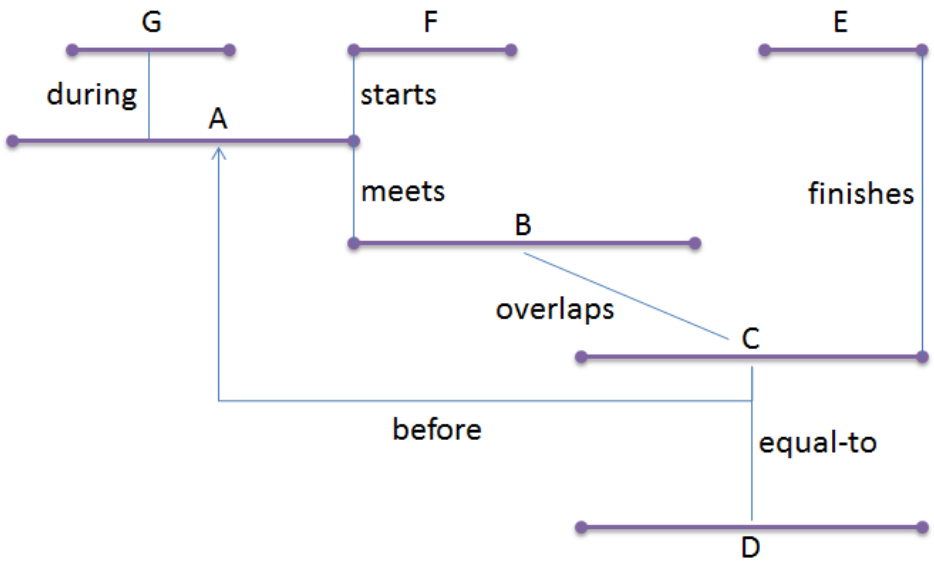

Figure 2.12: Intervals and their relations in Allen's Interval Algebra.

Additional relations `occurs-sometime-during` and `occurs-during-n` are also defined in [53]. An event is said to `occurs-sometime-during` another interval I if its interval is completely within I, but not `equal-to` the extent of I. Look at the `during` relation in figure 2.12. The `occurs-during-n` is used to represent events which occur multiple times within a given interval. This relation is accompanied by an arithmetic function $\phi(n)$ which asserts constraints on the number of occurrences within the interval.

2.3.4 Spatial Semantics

Real-world events play an important role in context, and therefore, we need to pay special attention to the spatial relationships between entities and events. A large amount of literature is available on spatial representation and reasoning framework originally known as RCC-8. Some of its common relationships are shown in figure 2.13. The important relations in the figure are `disconnected` (B and D), `externally-connected` (B and C), `tangential-proper-part` (A and B), `contained-in` (E and D) and `partial-overlaps` (C and D). In order to represent spatial relations between events and entities, we use the

`occurs-at` relation defined in [53]. An event E is said to `occur-at` region S if the spatial extent of E completely lies within S. The relations between the different regions in figure 2.13 are in table :

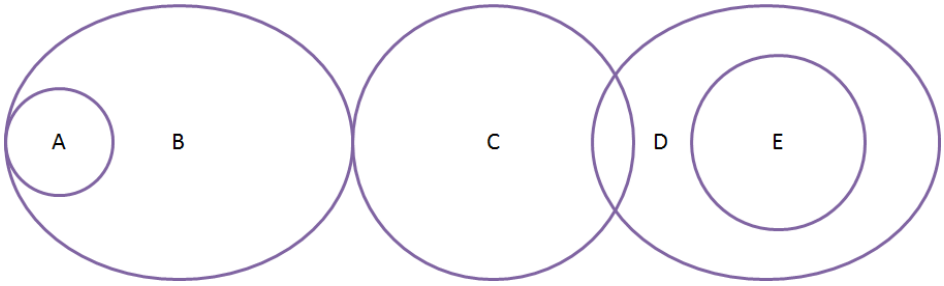

Figure 2.13: Spatial regions which can be represented in the RCC-8 framework.

Relation	**Symbol**	**Example**
Disconnected	DC	B DC D
Externally Connected	EC	B EC C
Tangential Proper Part	TPP	A TPP B
Completely Contained-In	CI	E CI D
Partially Overlaps	PO	C PO D

Table 2.1: RCC-8 relations in figure 2.13

Additional relations `occurs-somewhere-at` and `occurs-at-n` are also defined in [53]. An event is said to `occurs-somewhere-at`, if the actual region of occurrence is completely contained the given region, but not fully overlapping with it. The relation `occurs-at-n` is used when the event occurs at multiple places within the given region. This relation is accompanied with a function $\Psi(n)$, which asserts a constraint on the number of occurrences.

2.3.5 Real-World Knowledge

Modeling real world knowledge is critical in context based systems. Examples of knowledge are: An academic conference has atleast one keynote talk; or Sodium reacts with atmospheric oxygen at room temperature; or water expands when it freezes. By explicitly modeling such networks of knowledge, we provide the opportunity to influence extraction of context

from different sources. For example, if an object is associated with a keynote talk, data about the co-occurring conference is definitely going to be relevant context. Effectively, knowledge is the relations between real-world entities that have been documented so far. The primary motivation behind knowledge bases in context-aware systems is to transfer human experiences to the system. In a context discovery system, such knowledge bases play a pivotal role in reasoning about which relations need to be discovered. Given partially known information about an entity, knowledge indicates what general relations it shares with other entities, provides insight about what data sources must be searched for, and what can be ignored. In the above example, given the partial information about keynotes, and knowledge that keynotes occur with conferences, we can progress to discovering related conferences. At the same time, ignore all data sources relevant to social networks, sports games and weather information.

2.3.6 Source Agnosticism

Context is obtained from data sources or sensors. Therefore, it must be imperative to separate the sources from context. Context must be modeled irrespective of the nature of sources or their query abilities [110]. This provides the following advantage: context is now entirely defined in terms of objects, their possible inter-relationships and real world knowledge decides how objects are related to each other, irrespective of what data actually can be obtained. Second, with the growing number of data sources, personal and public sensors, this allows system engineers to plug-and-play different sources with ease, without disturbing the model and discovery algorithms.

This design requirement also brings a problem. How does context-aware system know what data sources are at its disposal, and how does it interact with them? The short answer to this question is to utilize data integration technology [41, 55, 70]. Data integration techniques

provide a uniform query interface to a multitude of autonomous data sources, irrespective of their native query capabilities or data storage formats. This might not be most ideal in the long run as these technologies were built to support RDBMS-like analytics operations, and we might benefit by adding more "context operations" into the integration layers, thereby allowing more fruitful grounds for query optimization. But for our current needs, it is advisable to re-use the ideas presented in these techniques to realize a fully working context-based system. The later chapters will describe in detail the data integration framework used to obtain data for the photo application.

There are many tools which partially support encoding models. For example, the ontology specification (OWL 2) supports specification of various entities, their class inheritance structure (`is-a`), their inter-relationships (where the inter-relationships can have their own hierarchy structure). But encoding time and spatial semantics is a relatively recent capability [59] in such frameworks, and not ready for prime-time for applications like real-world context discovery. In our work, relations expressed in OWL are simply labels with domain and range restrictions and sometimes cardinality constraints, their semantics being entirely declared and maintained within the discovery engine.

2.4 Context for Personal Photos

In this section, we list the specific objects, relationships which will be used in constructing context networks, as shown in figure 2.14, for tagging personal photos. Once such a context network is constructed, the search space is derived from it. The types used in the contextual domain, but not limited to, are the following:

- **Events**: events such as real world conferences, parties, trips or weddings, and their structure (for example, what kind of sessions, talks and keynotes are occurring within a

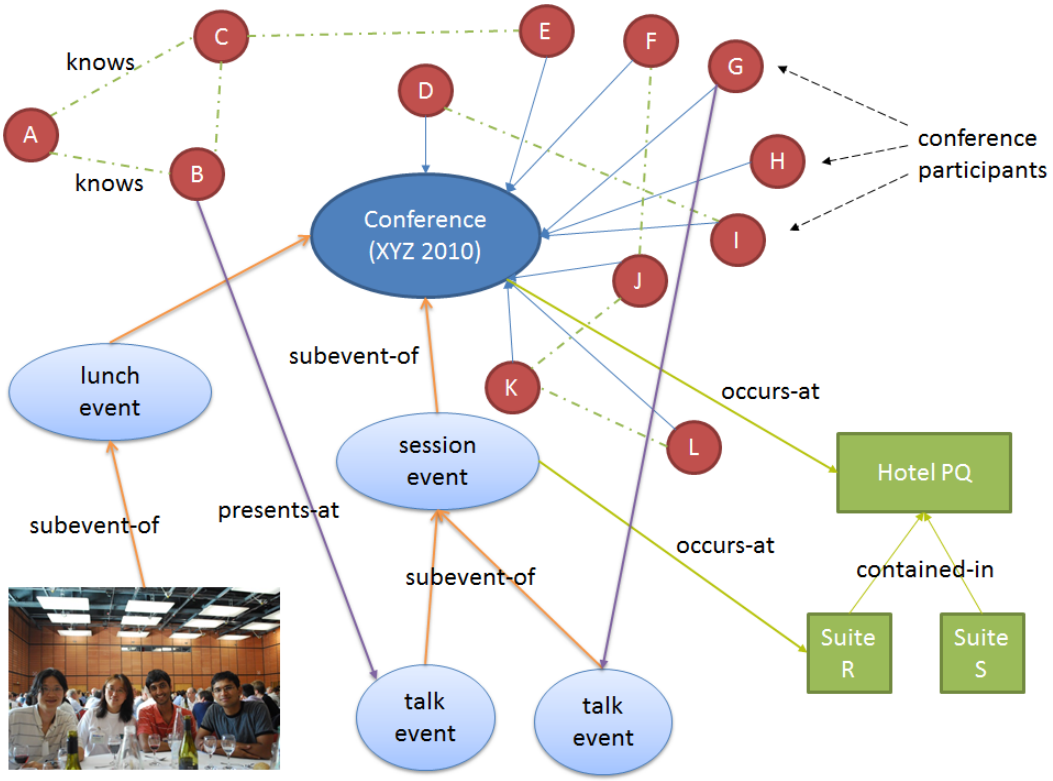

Figure 2.14: A context network representing real-world entities and their relationships.

particular conference). We also model the photo as media facet of a `photo-capture-event`, which signifies the moment where a person took a photo with his/her camera. Events in our framework are modeled according to the modeling primitives described in [53, 105].

- **Objects**: entities model users, people in their social graph, people with whom he/she corresponds with using email and other messaging services. The different organizations which can participate in events mentioned above. For example, a company Acme Corp where the user and person X are colleagues working in the same team.

- **Geographical Objects**: Places play an important role. A place can referenced by a geo position, for example (33.643036, -117.841911), or an address for example "Donald Bren Hall, University of California, Irvine".

In order to relate the above objects, we use the following types of relationships, along with their constraints and semantics:

- **Spatio-Temporal Relations**: Events occur during specific time intervals, and at some location. We use the relations `occurs-at` and `occurs-during` to model these properties [53]. They are formulated using recursion as follows: if `E` is an event and `I` is a time interval, `E` occurs during the interval `I`, if it also occurs during all subintervals of `I`. Similarly, if `R` is a region, `E` occurs at `R`, if it occurs at all subspaces of `R`.

- **Subevent Relation**: An event spatiotemporally contains all its subevents. The subevent relationship is an irreflexive, asymmetric and transitive relationship between a pair of events [53]. For example, a talk event is a subevent of the conference during which it happens. A subevent relation between two events `S` and `C` entails the following spatio-temporal constraint: If `C` is the subevent of `S`, the location of occurrence of `C`, `C.location`, is completely contained within `S.location` and its interval of occurrence, `C.duration` is completely contained within `S.duration`.

- **Participation Relation**: The relation `participant-of` is a relation between an event and object. If an person is performing a functional role in an event, s/he is said to be a participant-of that event. Note that this relation constraints the spatio-temporal boundary where the entity could be present during the interval of the event. If a person `P`, participates in an event `E`, the location of `P`, `P.location` must be contained with the space of `E`'s occurrence, `E.location` at the time of its occurrence, `E.duration`.

- **Social Relation (knows)**: This relation is defined in the FOAF ontology [26] as "the `knows` property relates a Person object to another Person object that he or she knows". This is used to model relations in social networking sources, such as Facebook or Twitter, between people.

The above classes of contextual data can be obtained from a variety of data sources. Examples of data sources range from mobile phone call logs and email conversations to Facebook messages to a listing of public events at upcoming.com. We assign each of our data sources

into one of the following classes:

- **Metadata**: Photos contain very valuable metadata in the form of EXIF tags. Prior research in multimedia systems has used them extensively to annotate photos with tags such as `landscape, portrait, indoor, outdoor` with very high accuracy [22, 99]. In our work, we use time and location tags from EXIF to populate the attributes of `photo-capture-event`s. Although we do not use them in our current work, the high importance of sensors embedded inside mobile phones deems their mention. Sensors like gyroscope, inclinometer, accelerometers and ambient light sensors can provide interesting cues about the environment [85, 98].

- **Personal Data**: include all sources which provide details about the particular user whose photo is to be tagged. Some examples of personal data sources include Google Calendar, Email and Facebook profile and social graph. Details include common personal attributes like name, date-of-birth, place of residence, place of work, the various they are attending (personal calendar), their personal preferences (type of food, music concerts, activity preferences).

- **Social Data**: include all sources which provide contextual information about a user's friends and colleagues. For example, LinkedIn, Facebook and DBLP are some of the commonly used websites with different types of social graphs. The information includes the personal information of the user's friends, their past or future activities where their participate together, their common friends (communities in social networks [12, 67]),

- **Public Data**: include all sources which provide information about public organizations (for example restaurants, points of interest or football stadiums) or about public events (for example conferences, fairs, concerts or sports games). Some sources include Yelp, Upcoming, DBLP and Factual.

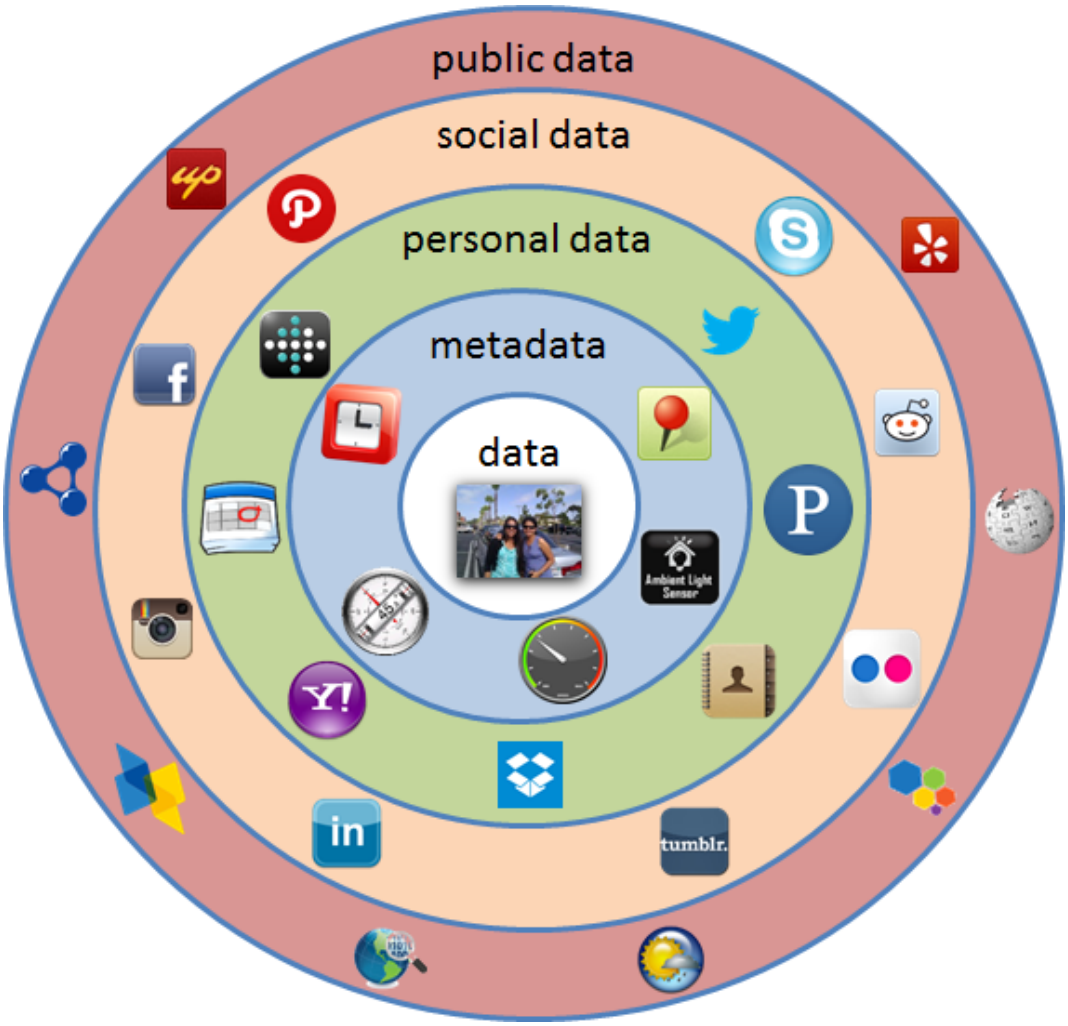

Figure 2.15: Personal, social and public data sources which can contribute relevant context.

Figure 2.15 shows some commonly available personal, social and public sources. Social and public data sources are enormous in size, containing information about billions of events and entities. Trying to use them directly will lead to scalability problems similar to those faced by face recognition and verification techniques. But, by using personal data, we can identify which parts of social and public sources are more relevant. For example, if a photo was taken at Staples Center, Los Angeles, CA, an indoor stadium, we only need consider public events which such as concerts or sports games in the area. Thus, the role of personal information is twofold. **Firstly**, it provides contextual information regarding the photo. **Secondly**, it acts as a bridge to connect to social and public data sources to discover interesting people

connected to the user who might be present in the event and therefore, the photo.

At this point we should revisit the **dynamic linking** property of a discovery algorithm. Given a stream of photos taken during a time interval, the source which contributed interesting context for a photo might not be equally useful for the one appearing next. This is because sources tend to focus on a specific set of event types or relationship types, and the two photos might be captured in different events or contains persons with whom the user maintains relations through different sources. For example, two photos taken at a conference might contain a user's friends in the first, but with advisers of these friends in the next. The friends might interact with the user through a social network, but their advisers might not. By using a source like DBLP, the relations between the adviser and friends can be discovered. We say that the temporal relevance of these context sources is ***low***. This requirement will play an important role in the design of our framework, as now, sources are not hardwired to photo, but instead need to be discovered gradually.

In short, we saw the different definitions of the term Context, and the different implications it has on solving problems. This chapter presented a relation centric view of context, and the implications it has in modeling context for building systems to solve problems using context based techniques. We also outlined the different types of contextual information which will be used in our specific application of tagging faces in photos. Chapter 4 will present a technique to construct context networks similar to the one shown in figure 2.14 using metadata, personal, social and public sources as shown in figure 2.15. The next chapter, will present a short survey of various techniques relevant to our problem, and highlight the important ideas which helped develop the algorithms presented in this dissertation.

Chapter 3

Related Work

In this chapter, we will look at the state-of-the-art understanding of different topics presented in a variety of areas, which context relies on. These include techniques to annotate photos using image features, techniques to represent knowledge in ontologies, techniques to query data from a variety of sources.

3.1 Photos and Annotation

Kindberg et al. [65] conducted a study on multiple subjects to analyze photo capture behavior. Specifically, they were trying to understand the different motivations for people to take photos. They found two main motivations – *affective* (sentimental) reasons and *functional* reasons. A significant number of photos are captured to enrich a shared experience, or to share an experience with absent members like friends or family. Much lower, but a significant amount of photos were taken to share photos, in support of a task, with people who were present at the event or with remote family or friends. Personal uses of photo capture include personal reflection or reminiscing or some future task (capturing photos of gift ideas,

whiteboard designs) not involving sharing. Their study shows that photos are mostly used in social context, and lesser in personal context. Ames [10] and Frohlich [48] independently describe a survey conducted to study motivations for people to tag their photos. They noticed two broad motivations: Organization of photos and Communication with photos. Almost orthogonal to the applications observed by Kindberg, Ames and Frolich are brave new world applications for photography described in [50, 43], where life logs were collected in the form of photos, emails, document scans and stored in SQL Server database, and photos were retrieved using SQL queries. The photo content was tagged by the user in this case. The annotated photos are used exclusively for personal *feedback* to improve quality of life. The medical advantages of collecting photos to log personal health are becoming very well known [16].

3.1.1 Spatio-Temporal Annotation

Ever since its release in 1995, EXIF metadata, contained in JPEG photos, has been exploited to organize pictures. Almost all photo management applications use timestamps to order photos in an album, a concept which has also been studied in academia [52, 54]. Apple's iPhoto is the most common example of using both time and space to organize personal photos. Naaman et al. have exploited GPS attributes to extract place and person information [78, 79]. Rattenbury [91] devised techniques to find tags which describe events or places by analyzing their spatiotemporal usage patterns. Sinha [99] and Boutell [23] have used EXIF metadata to predict concepts such as (Indoor, Outdoor, Portrait, Landscape) to further organize these photos. Boll [20] is a very interesting work which aggregates textual and experiential content from web 2.0 communities about places using a set of predefined rules. These works motivate the need for adding annotations to improve the overall experience of viewing a collection of photos.

3.1.2 Computer Vision

The Computer Vision community has contributed extensive work in the area of detecting scenes [108], humans [35] or geo localization [56]. Here we will specifically look at the part of the computer vision research which is relevant to face tagging.

Face Recognition The face recognition problem is one of the primary problems taken on by the computer vision community, the others being action, pose, gesture recognition and human detection. One of the earliest works on recognizing people in faces was attempted by Turk and Pentland in [102]. The essence of the technique put forward here was to extract useful features which can be represented mathematically and compared using some distance measures (such as euclidean, mahalanobis or cosine distance) to features in an image database. This contribution sparked a large interest in the community to find more meaningful and powerful features, which led to Fisherfaces [15] and more recently, local binary pattern [6] features. SIFT features have also been used to identify faces [19, 51, 75]. One of the big difficulties of such feature-based representation is the scalability of the technique. It is now well known that with the increasing number of candidates, these distance based techniques provide in lower performance [107]. A quick alternative is to increase the dimensions in the features to allow for more diversity, but it has been proven that as the number of dimensions increase, the maximum distance between the two points reduces, and if the number of dimensions is as less as 32, the distance is almost negligible [18]. Effectively, for dimensions more than 32, the distance between all points can be considered very small, and the concept of 'nearest-neighbour' loses its meaning. Thus, an pure feature based technique relying on nearest neighbor principles cannot be used to build large scale real world photo tagging frameworks.

Face Verification More recently, face verification techniques have attracted a lot of attention in the computer vision community. One of the earliest efforts towards robust face

verification was undertaken by Huang et al. in [61]. They constructed and annotated a dataset of profile photos of celebrities taken in unrestricted environments. Previous datasets such as FRGC [86], BioID [64] and the color FERET database [87] were criticized to have photos taken at very constrained environments thereby reducing the complexity of the face tagging problem. But the techniques which worked on such photos could not duplicate their performance in real-world photos. Huang's database, commonly known as the *Faces In the Wild* has allowed many researchers to implement various face verification technologies. A recent and important development on top of Huang's work is that of Neeraj Kumar et al. [68], where the confidence of presence or absence of facial attributes such as range, skin color, hair style, gender, eye color features are used to train various classifiers to ultimately test if two given faces are of the same person or not? The technique was able to achieve up to 85.29% accuracy on the LFW dataset. More recently, Berg [17] increased this accuracy to 93% by automatically finding distinguishing features.

The important different in face verification is in its contract with co-existing components in a system. Whereas, face recognition works as a standalone component, face verification doesn't make this assumption, and allows external components to filter candidates. This systemic behavior of face verification will prove to be very useful in future systems.

Probabilistic Techniques Both face verification and recognition techniques seen so far assume that none of the input photos contain any annotation. But what if this assumption could be relaxed? A large number of photos on Facebook, Flickr or Google+ are annotated. If we further assume that these partial annotations are mostly true, the label propagation technique by Cao et al. [27] can be applied to annotate the rest of the dataset. This technique was tested on propagating concept based tags (such as beach, fun, dinner, yard) on personal photo datasets. Also, Barthelmess et al. extract semantic tags from noisy datasets containing discussions, speeches about a set of photos in question[13].

Miscellaneous Techniques Collaborative games also have been evaluated as a possible

way to tag photos[39]. Systems like Picasa, iPhoto and [52] organize photos based on time, GPS coordinates and sometimes faces in the photo. These attributes of the photo do not capture event semantics [95]. Events are a natural way of categorizing photo media. Events also allow large number of photos captured during a single event be organized hierarchically using subevents.

3.2 Context

The use of context in the sciences has been continuously increasing. It finds applications in various fields, starting from its use in holistic thinking to better understand biological and ecological phenomenon in [29], to associating the right external data to form coherent stories about economic phenomenon [71], to associating plausible connections between history and geography in [40]. The advances proposed in these works can be loosely characterized as "utilizing external information" or "out-of-the-box thinking". The reason they are included in this section is because of their common trait of relating multiple pieces of external data, to create coherent stories which allows the scientists to gain valuable insights into the problem they are solving.

3.2.1 Uses in Computing

In Computer Sciences, the main interest towards context has been largely from the Pervasive Computing community and the Human-Computer Interaction community. Their interpretation of the word context is mainly inspired from the definition set by Anind Dey [38], i.e. *Context is any information which describes the situation of an entity.* The role of context in mobile computing has been studied in [31]. It shows the growing importance of contextual thinking in reasoning about networking problems in the mobile computing era.

More recently, information retrieval communities are showing interest in context based representation of data and context-based techniques. One of the most important works in information retrieval is the PageRank algorithm [83] developed by Google co-founders Larry Page and Sergey Brin in 1996. In this work, Page and Brin define context as the links between different web pages, and the anchor text of these links, and argued that this context is more descriptive of a page than the contents of the page itself. PageRank was derived by combining this insight with Jon Kleinberg's HITS [66] algorithm.

3.2.2 Definitions

There are many definitions of the word context in academic works. Notable among them are those of Schilit [96], Dey [38], Viera [103], Zimmermann [114] and the work of Patrick Brezillon, a summary of which can be found in [77]. One of the problems with the term Context is the overloaded use of it [57]. To this effect, Brezillon compiled a list of 150 definitions, and did to this list what lots of scientists do with large collections of text – text analysis using natural language processing techniques, and derived the essential components which should make up a holistic definition of context. Their conclusion as described in [14] was that *context acts like a set of constraints that influence the behavior of a system (a user or a computer) embedded in a given task.* In spite of this promising "big-data" analysis, many questions are unanswered. For example, is context internal or external? Is it a phenomenon or an organized network? The most accepted definition of context is the one propsed by Dey: *as any information that can be used to characterize the situation of an entity. An entity is a person, place or object that is considered relevant to the interaction between a user and an application, including the user and applications themselves.* As we have seen in chapter 2, this is a object centric definition. The information about relations between objects is implicit. In our work, we make this information explicit by taking a temporal relation centric view: a notion where context is determined by the relations between entities

at a particular point in time.

3.2.3 Role in Photo Annotation

In [44] Dumitrescu and Santini argue that "images are a node in a complex network of signification that goes beyond their content and includes other form of textuality that go around them, as well as the cultural practices of the community that creates or receives images". Such philosophies became more commonplace after the web went contextual with PageRank. Image retrieval, specifically, obtained a remarkable opportunity to index and rank images on the web by using the textual content which accompanies it [32, 47]. Later, with Flickr the amount of user contributed tags has resulted in gigantic datasets which allow researchers to train sophisticated machine learning models to produce one or more relevant tags [24, 72, 74]. Some of early advocates of using external context to annotate photos are [36, 63]. Context information and image features are used in conjunction by [82, 27, 23, 28] identify tags. The semantic web community is using linked data technologies to annotate and query photographs [76, 80].

3.2.4 Modeling Context

Context is represented using three components: knowledge, external context and proceduralized context by [25]. These ideas are represented using a contextual graph (CxG) representation of knowledge and reasoning. Henricksen et al. [57] also use a graphical notation to represent their context. The former approach uses the graph to model the flow of reasoning between different objects within an environment, whereas the latter approach uses graph to only represent the various objects and their inter-relationships. Their modeling framework allows representation of static and dynamic associations, which is very critical in modeling real world relationships. For our work, we rely on primitives like the one mentioned in this

framework as well as event relations mentioned in [53]. [92] presents a technique to transform contextual graphs to create concrete situation handling implementations using Petri Nets. [60] presents a detailed survey of various other context-aware systems.

Probabilistic and statistical frameworks are used to incorporate context in problem solving too. One example of using a probabilistic framework is [27], where high confidence tags from a few contextual images are propagated to other images which were taken during the same event. [101] uses potential functions in conditional random fields to model social relations and photo co-occurrence strengths. More recently, Zhang [112] used cues from photos like clothing, human attributes like face complexion or gender, people co-occurrence and scene annotations to cluster similar faces in photos together.

3.2.5 Industrial Momentum

The tech industry is immersed in the so-called "big-data" revolution. In order to make sense of such large collections of data, context is becoming increasingly popular. One of the biggest works in this area is the blog and upcoming book "Age of Context" by Robert Scoble and Shel Israel[1]. The reasons for this interest is cited on the increasing number of smartphones and sensors, personal and public. The increasing number of wearable devices like Google Glass, Oakley AirWave Goggles, Plantronics, Smith I/O Recon Heads-up Display, FitBit, Basis Watches, Nike FuelBand and Jawbone Up's personal health/activity monitoring gadget. The continuously increasing social data from companies like Facebook, Twitter, Flickr, Instagram and Pinterest, and information about locations through services like Foursquare, Google/Bing Maps, Waze and Factual. Lesser known startups like Tempo and Cueup (formerly Greplin) which are entirely focused on using personal contextual information to simplify otherwise very hard problems.

[1] http://scobleizer.com/2013/02/12/a-new-way-to-fund-a-books-development-sponsors-heres-ours/

3.3 Knowledge Representation

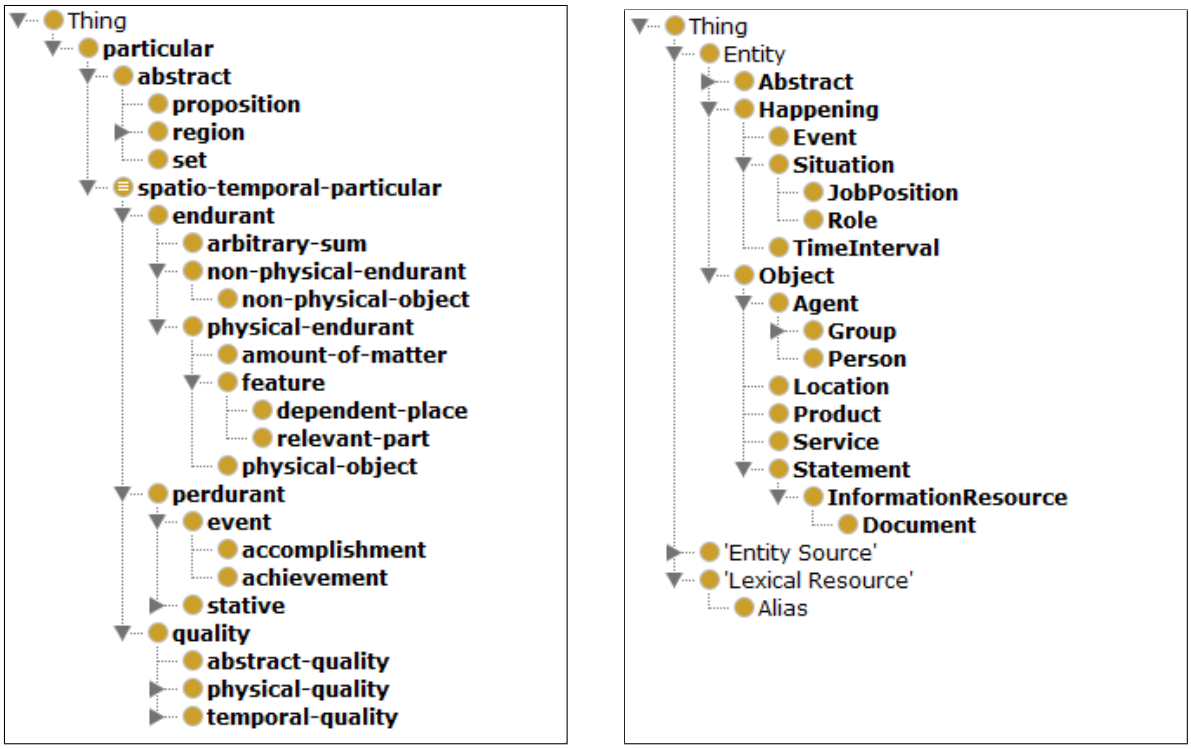

Figure 3.1: Dolce and Proton Class Hierarchies

In our work, we use ontologies to model key pieces of knowledge. This includes the type of objects and relationships they are expected to have. OWL is the current standard language in authoring ontologies. The current standard being OWL 2.0. Some of the work in this dissertation is in using ontologies for integrating autonomous data sources [100, 81, 11]. To achieve this, we borrow features of the declarative mapping language [42] to express the structure of relations stored in various data sources. Specifically, we use the s-expression like syntax to declare the relations and the mapping axioms which relate items in the expression to objects in our ontology. SPARQL [89] is the current standard for querying RDF documents. The event and object queries to discover context from various data sources use SPARQL. We use the Jena [30] SPARQL query processor (ARQ) to evalaute these queries and convert them to the native query language of the data source.

We use the terminologies followed in upper level ontologies. Figure 3.1 shows the class

hierarchies for the Dolce Upper Ontology (on the left) and the Proton Top Level Ontology (on the right). Similar to these taxonomies, we will use the term 'entities' to collectively refer to all real-world events and objects (In DOLCE, the `Particular` class is a synonym for 'Entity'). Events, are perdurants, and have temporal or spatial parts. Events in our context discovery will include Academic Conferences, Concerts, Photo-Capture Events or Meetings. Objects, such as Persons, on the other hand, are uniquely identifiable wholes. They do not need temporal or spatial attributes for their description. It must be noted that we extend the DOLCE-Lite upper ontology to construct our knowledge base for CueNet.

Chapter 4

Context Discovery Framework

In this chapter, we shall look at the the CueNet framework and its components: a *data integration module* to model and query the various data sources and sensors, a *discovery algorithm* to construct queries agnostic to what the sources are themselves, a *knowledge representation module* to store relationships about the various real world objects, and finally mechanisms to interact with a *face verification algorithm*, which computes the confidence score of a person being present in a photo or not.

4.1 Pruning Search Spaces with CueNet

Automatic media annotation algorithms essentially assign one or more labels from a search space to a given input image. Figure 4.1 shows the various approaches of constructing search spaces for such algorithms. The traditional approach is shown in 4.1(a). These spaces were limited to a set of labels chosen by an expert, with no way of pruning the search space in case it got very large. The focus was instead on extracting the best features from images, to obtain high overall classification accuracy[102].

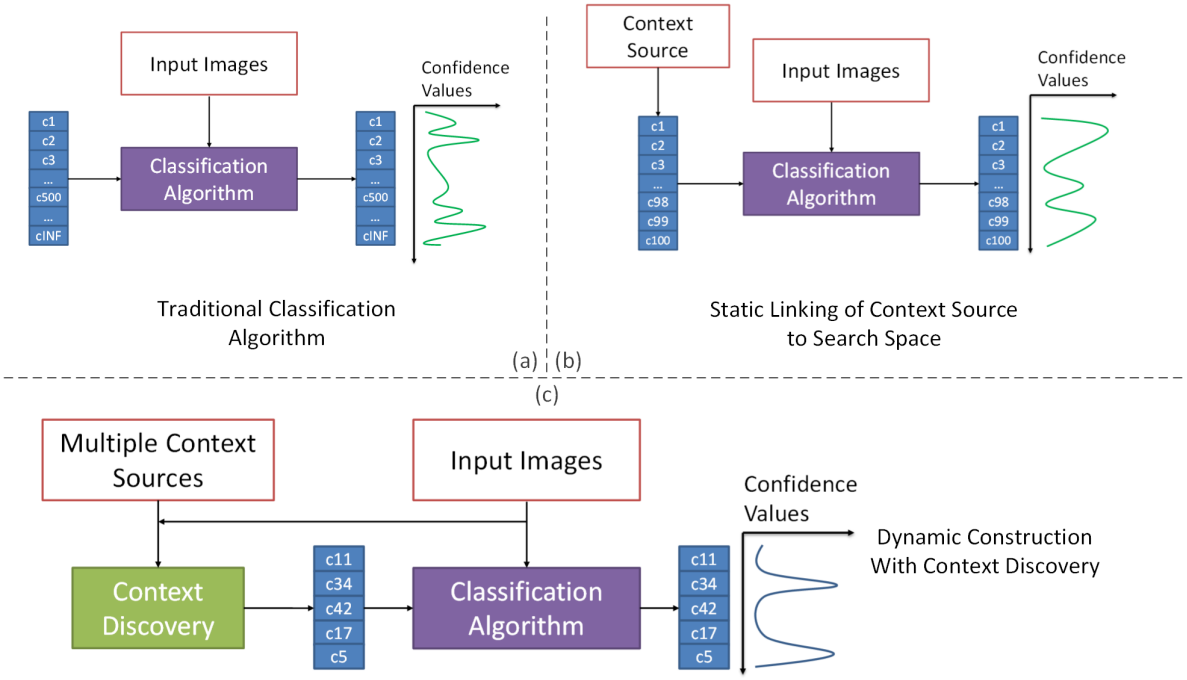

Figure 4.1: The different approaches in search space construction for a multimedia annotation problem. A traditional classifier setup is shown in (a) where the search space candidates are manually specified. Context is used to generate large static search spaces in (b). The desired framework is shown in (c), which aims to produce small search spaces with many correct annotations.

With the popularity of global social networks and proliferation of mobile phones, information about people, their social connections and day-to-day activities becoming available at a very large scale. The web provides an open platform for documenting many real world events like conferences, weather events and sports games. With such context sources, the search space construction is being delegated to one or a few sources [58, 73, 79, 82, 101] (figure 4.1(b)). These approaches rely on a single *type* of context. For example, time and location information or social network information from Facebook to solve the face recognition problem. We refer to such a direct dependency between the search space and a data source as **static linking**. Although these systems are meritorious in their own right, they suffer from the following drawbacks: they do not employ multiple sources, and therefore the **relations** between them. By realizing that these sources are interconnected in their own way, we are able to treat the entire source topology as a network. Our intuition in this work is to navigate this network

to progressively discover the search space for a given media annotation problem. Figure 4.1(c) shows how context discovery can provide substantially smaller search spaces for a set of images, which contain a large number of correct tags. A small search space with large number of true positives provides the ideal ground for a classification algorithm to exhibit superior performance.

The CueNet framework provides access to multiple heterogeneous autonomous data sources containing event, social, and geographical information through a unified query interface to extract information from them. CueNet encapsulates our **Context Discovery Algorithm**, which utilizes the query interface to discover the most relevant search space for a media annotation problem. To ensure a hands-on discussion, we show the use of context discovery in a real world application: face tagging in personal photos. As a case study, we will attempt to tag photos taken at conference, trip and party events by different users. These photos could contain friends, colleagues, speakers giving very interesting talks, or newly found acquaintances (who are not yet connected to the user through any social network). This makes the conference photos particularly interesting because no single source can provide all the necessary information. At the same time, by studying the efficacy in trips and parties, we will ensure that discovery can be done across different kinds of events. It emphasizes the need to utilize multiple sources in a meaningful way.

Here is an **example** to illustrate CueNet's discovery process. Let's suppose that Joe takes a photo with a camera that records time and GPS in the photo's EXIF header. Additionally, Joe has two friends. One with whom he interacts on Google+, and the other using Facebook. The framework checks if either of them have any interesting event information pertaining to this time and location. We find that the friend on Google+ left a calendar entry describing an event (a title, time interval and name of the place). The entry also marks Joe as a participant. In order to determine the category of the place, the framework uses Yelp.com with the name and GPS location to find whether it is a restaurant, sports stadium or an apartment complex.

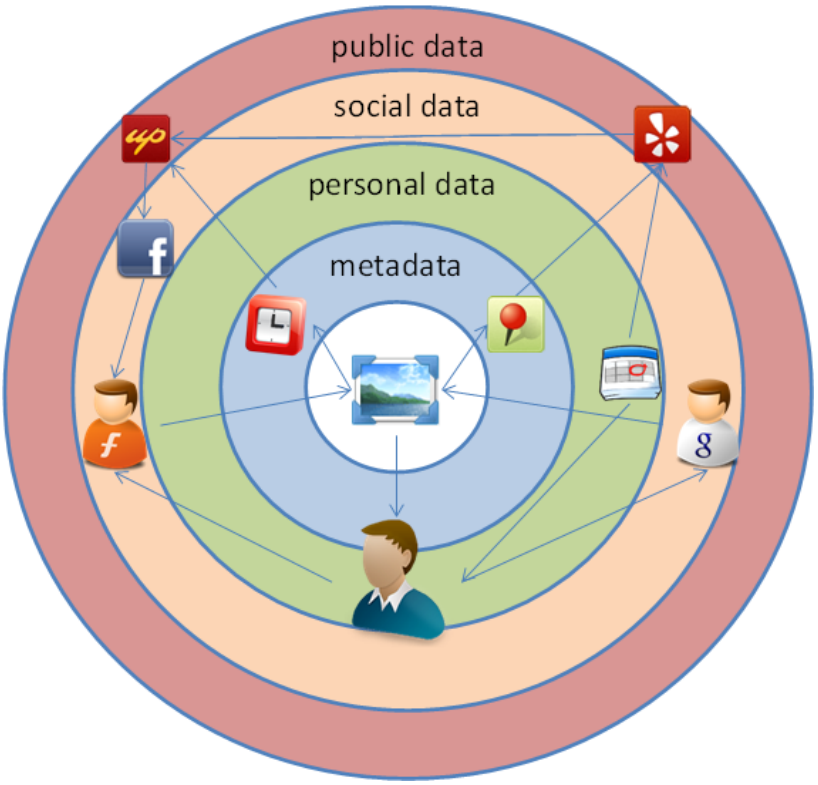

Figure 4.2: Navigation of a discovery algorithm between various data sources.

If the location of the event was a sports stadium, it navigates to upcoming.com to check what event was occurring here at this time. If a football game or a music concert was taking place at the stadium, we look at Facebook to see if the friend "Likes" the sports team or music band. By traversing the different data sources in this fashion, the number of people, who could potentially appear in Joe's photograph, was incrementally built up, rather than simply reverting to everyone on his social network or people who could be in the area where the photograph was taken. We refer to such navigation between different data sources to identify relevant contextual information as **progressive discovery**. The salient feature of CueNet is to be able to progressively discover events, and their associated properties, from the different data sources and relate them to the photo capture event. We argue that given this structure and relations between the various events, CueNet can make assertions about the presence of a person in the photograph. Once candidates have been identified by CueNet, they are passed to the face tagging algorithm (as in [69]), which can perform very well as

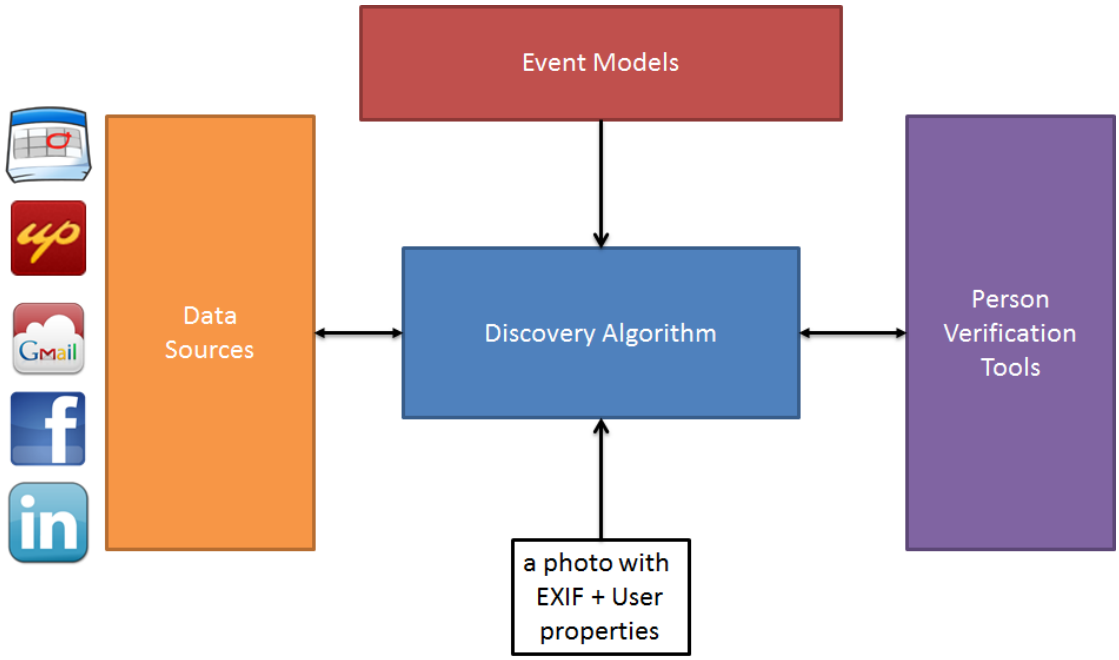

Figure 4.3: The Conceptual Architecture of CueNet.

their search space is limited to two candidates.

Figure 4.3 shows the different components of the CueNet framework. The Ontological **Event Models** specify various event and entity classes, and the different relations between them. These declared types are used to define the **Data Sources** which provides access to different types of contextual data. The **Person Verification Tools** consist of a database of people, their profile information and photos containing these people. When this module is presented with a candidate and the input photograph, it compares the features extracted from the candidate's photos and the input photo to find the confidence threshold. In this section, we describe each module, and how the context discovery algorithm utilizes them to accomplish its task.

4.2 General Approach

Figure 4.3 shows a high level architecture of CueNet. The data integration module of CueNet provides a uniform query interface to a multitude of autonomous data sources, which may reside within an a personal device (such as mobile phone or the PC) or on the World Wide Web [55]. The event model describes the various real world events that are relevant to the problem, and their relations and constraints with other entities. Any relevant axioms are specified in the model. We will describe axioms for the face tagging problem in section 4.6. For the face tagging problem, CueNet needs to closely work with tagging algorithms to check if a person is present in a photo or not. For this purpose, we shall assume verification semantics in such a tagging algorithm, where given an input photo and a candidate person, the algorithm returns true or false (possibly, with a confidence score). Face recognition models would have to be retrained when the candidate set changes, which is not necessary in verification algorithms. Also, as described in chapter 3, the state-of-the-art techniques for face verification perform much more reliably than their recognition counterparts. Alternatively, an web based service such as the former face.com could be utilized.

At the heart of CueNet, lies the context discovery algorithm. Given a photo the algorithm constructs a context network with all the known information. Using the knowledge base, the algorithm constructs queries to be executed on the interface provided by the data integration layer. A query is constructed in the following way: the current structure of the network is examined to list all nodes. For each node, we check the ontology to see what are the general relations this node usually shares with other types of objects. For example, if the node is a Person, the ontology will respond by saying that Persons `participate-in` events. Using this node and possible relation information, algorithm creates a query to find all entities which can satisfy these relations. These queries are executed on different sources by the data integration module. Contextual information returned from the data sources are merged into existing context network. New entities in the network are passed to the face tagging algorithm to

check for their presence in the photo. If they are present, the context network is updated to reflect this information. The query and merge operations are iteratively performed until all the faces are tagged, or the data integration module is unable to furnish any new context. In the later sections, we will see a more detailed specification of these algorithms.

4.3 Execution Trace

In this section, we will trace the execution on two different photos, to see how the different modules interact to produce context networks, and how they are used to tag faces. The first example will be a relatively simpler one, requiring only 2 data sources, whereas the second photo will require multiple sources to sucessfully tag all photos.

4.3.1 Simple Case

Consider the photo shown in figure 4.4. For the purposes of this trace, we assume that we have access to the sources shown in figure 4.5 through the data integration module. Given, an input photo, the knowledge base is queried to find what other objects can be associated with an photo object. The KB stores the information that every photo consists of an EXIF header, which stores timestamp and location coordinates and a fact which states that every photo is owned by a user object, where **owner-of** is a relationship described in the KB. This knowledge is used to construct the context network shown in figure 4.6.

Now, the algorithm traverses the graph to list all the possible queries it can execute on the data integration layer. Given the knowledge that entities participate in events, and events can contain participants, it generates the following queries:

- Does any data source contain participant information related to the photo capture

Figure 4.4: Input Photo.

Figure 4.5: Available Data Sources.

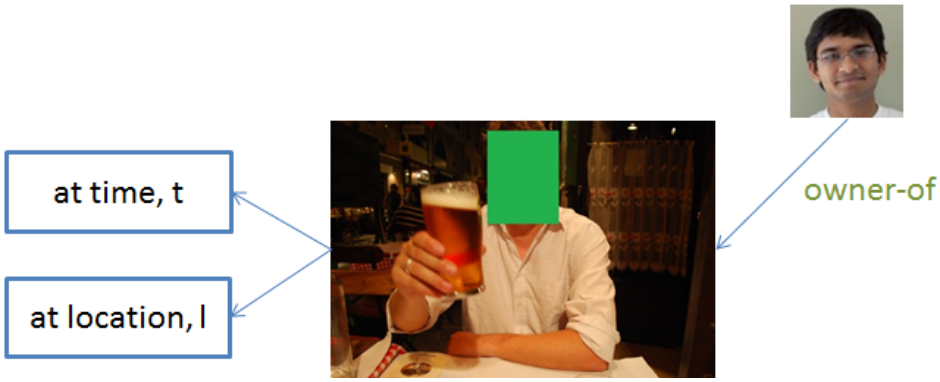

Figure 4.6: Context Network built after User Information and EXIF sources are queried and merged.

event?

- At time t, and location l, which events contain the owner (entity:ArjunSatish) as a participant?

The data integration system looks at the different sources and says that none of them store information about photo capture events, and skips executing the first query. But many sources describe events, and store their participants too (Google Calendar, Facebook, Conference Calendar). These query is converted to their native formats (API calls or relational database queries) and results, are sent back to the data integration module. We see that there was a calendar entry returned be the Google Calendar source, as shown in 4.7. This information is now merged with the existing context graph to produce a context graph similar

to that shown in 4.8.

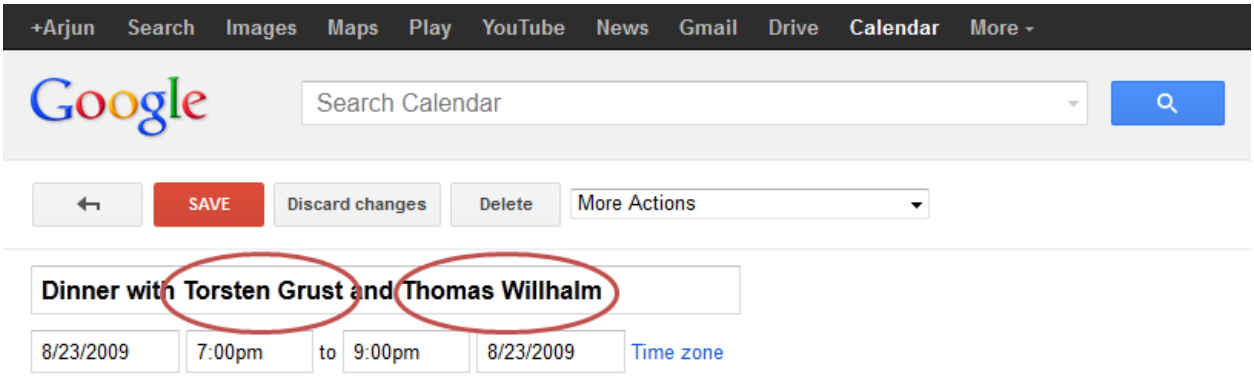

Figure 4.7: Calendar Event.

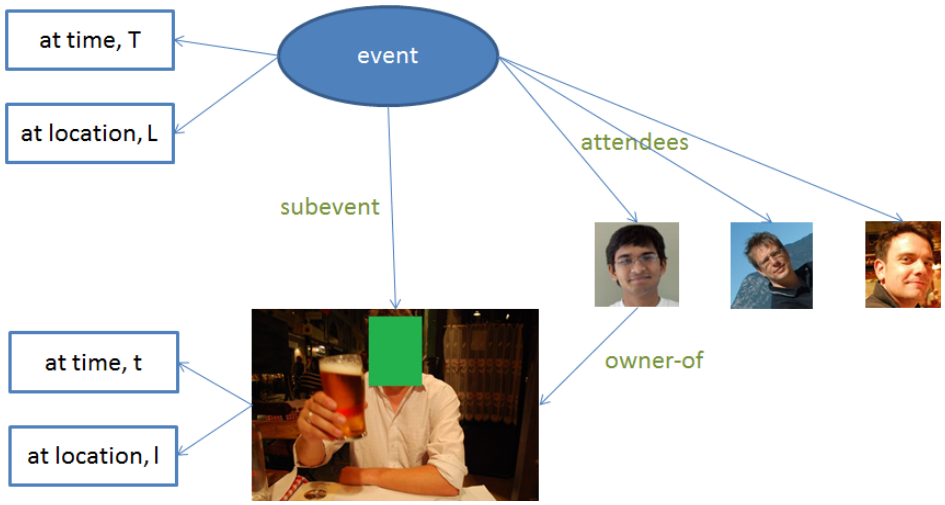

Figure 4.8: Context Network after integrating calendar information.

Now, we have new entities related to the photo. The face verification algorithm is invoked with the new set of candidates. It must be noted that this verification problem is much easier than trying to verify out of many thousands of candidates. Once the correct entity is identified, the photo is annotated as shown in figure 4.9.

One last look at the source diagram in figure 4.10 shows which data sources revealed interesting information related to this photo. In this case, EXIF provided some relevant context on when and where the photo was taken. The owner's personal calendar provided information on what event was occurring during the time of photo capture, and who else was involved in it.

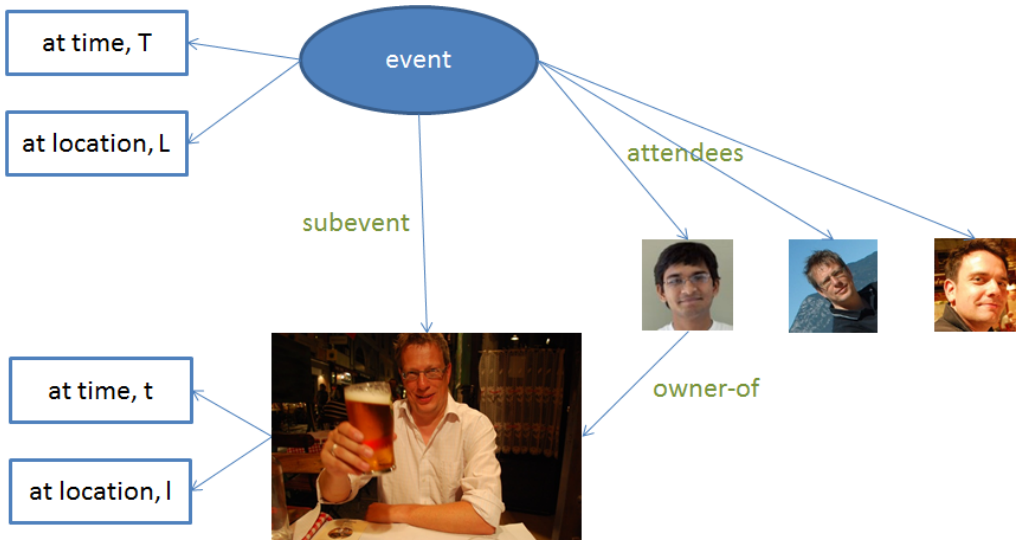

Figure 4.9: Face tagged with the correct person.

Figure 4.10: Highlighted Sources Provided Relevant Context.

4.3.2 Complex Case

Now, we will consider a more complex case which requires more than just metadata and personal sources for successful tagging. The photo under consideration is shown in 4.11. We will use the same set of data sources, shown again in 4.12.

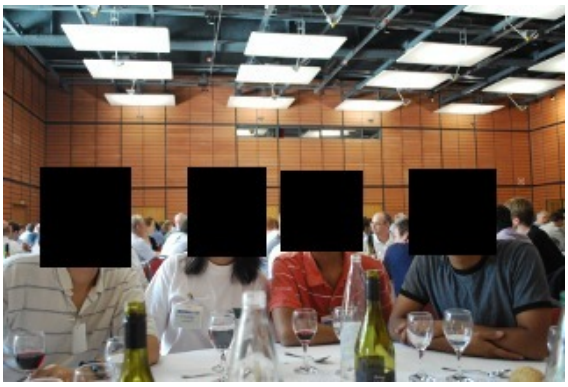

Figure 4.11: Input Photo.

Figure 4.12: Available Data Sources.

Using metadata sources and personal information, we arrive at the context network shown in figure 4.13. The procedure until here is exactly same as that for the previous scenario.

Figure 4.13: Context Network built after User Information and EXIF sources are queried and merged.

Now, given the known state of the world, if we invoke the face verification tools, we discover that the owner is actually present in the photo (figure 4.14). In this case, the candidate set contains just one entity, and therefore reduces the complexity of the tagging algorithm.

Figure 4.14: Context Network built after User Information and EXIF sources are queried and merged.

The next query generated by the system is to discover what the (entity:ArjunSatish) was doing at this time? But, this time we find that the conference calendar holds the answer.

At this point, the conference event is known to our knowledge base to have a definite structure, in terms of keynote, session and talk events with lunch/coffee breaks interleaved, and having many attendees. So it immediately queries the conference source to find and merge all of these objects. It discovers that the photo was taken during a break event, and that the conference (VLDB 2009) has many hundreds of participants, as shown in figure 4.16.

Figure 4.16 shows the various attendees discovered by the algorithm from the conference source. But finding 3 candidates from hundreds is an equally challening task. Before invoking

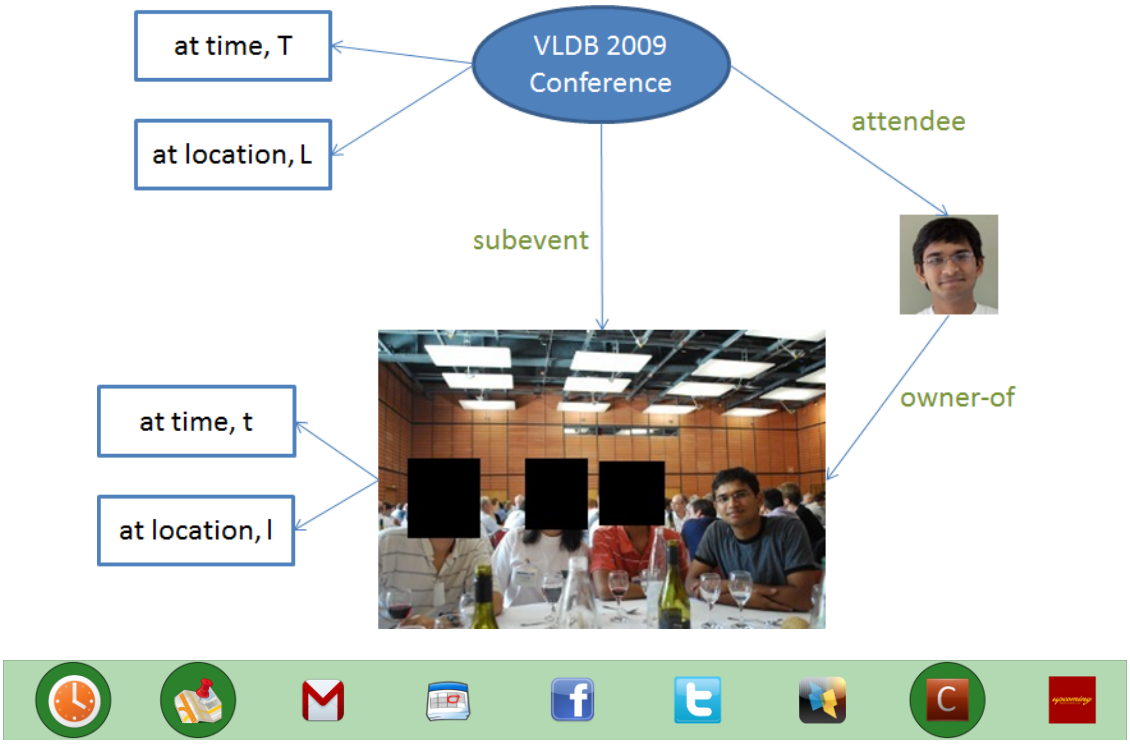

Figure 4.15: Context Network after querying conference sources.

the face tagging algorithm, we want to see if we can discover any more relations between the objects in the context network. So the discovery algorithm consults the knowledge base to find that entities can be related through a **friend-of** relation. So it queries all known sources to find friend relations, and finds that Facebook, Gmail and Twitter are sources which store data containing this relation. Querying it reveals that a few of the entities who were present at the conference were related to the user, and therefore have a bigger chance of appearing in the photo. The face verifier is invoked only with these candidates, for potential true positives. By doing this we tag two more faces in the photo. The context network is shown in the figure 4.18.

Since we have more candidates tagged in the photo, we can repeat the above procedure to discover more relation between the entites related to the photo and those who are present in the conference. This time results are returned from Gmail, and none from Facebook and Twitter (because these people had sent emails to each other during the conference, but did

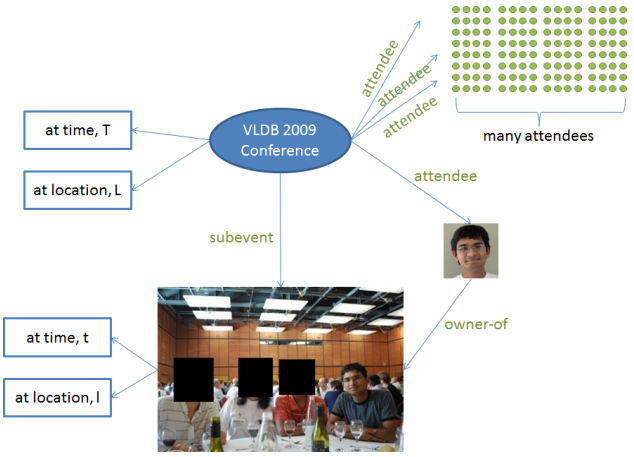

Figure 4.16: Context Network after discovering conference attendees.

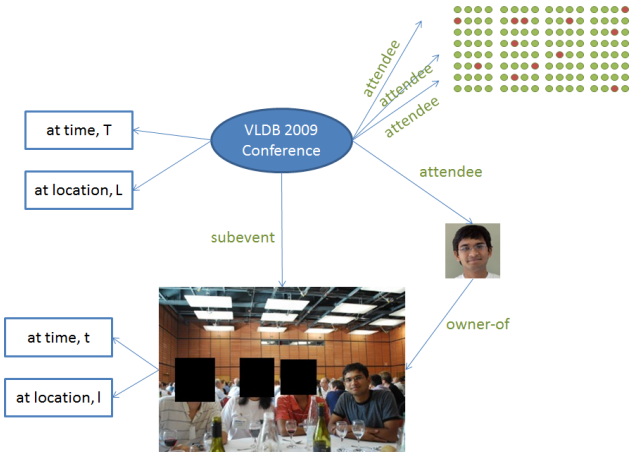

Figure 4.17: Context Network after discovering relations between attendees and owner.

not connect through Facebook or Twitter). The changes in the context network are shown in 4.20 and 4.21 (the final face is identified). Figure 4.22 highlights all sources which returned relevant context for this trace.

4.4 Event Model

Our ontologies extend the E* model[53] to specify relationships between events and entities. Specifically, we utilize the relationships "**subevent-of**", which specifies event containment. An event $e1$ is a subevent-of of another event $e2$, if $e1$ occurs completely within the spatiotemporal bounds of $e2$. Additionally, we utilize the relations **occurs-during** and **occurs-at**, which specify the space and time properties of an event. Also, another important relation between entities and events is the "**participant**" property, which allows us to describe which entity is participating in which event. It must be noted that participants of a subevent are also participants of the parent event. A participation relationship between an event and person instance asserts the presence of the person within the spatiotemporal region of the event. We argue that the reverse is also true, i.e., if a participant P is present in $\mathcal{L}_P$ during the time $\mathcal{T}_P$ and an event E occurs within the spatiotemporal region $< \mathcal{L}_E, \mathcal{T}_E >$, we say P

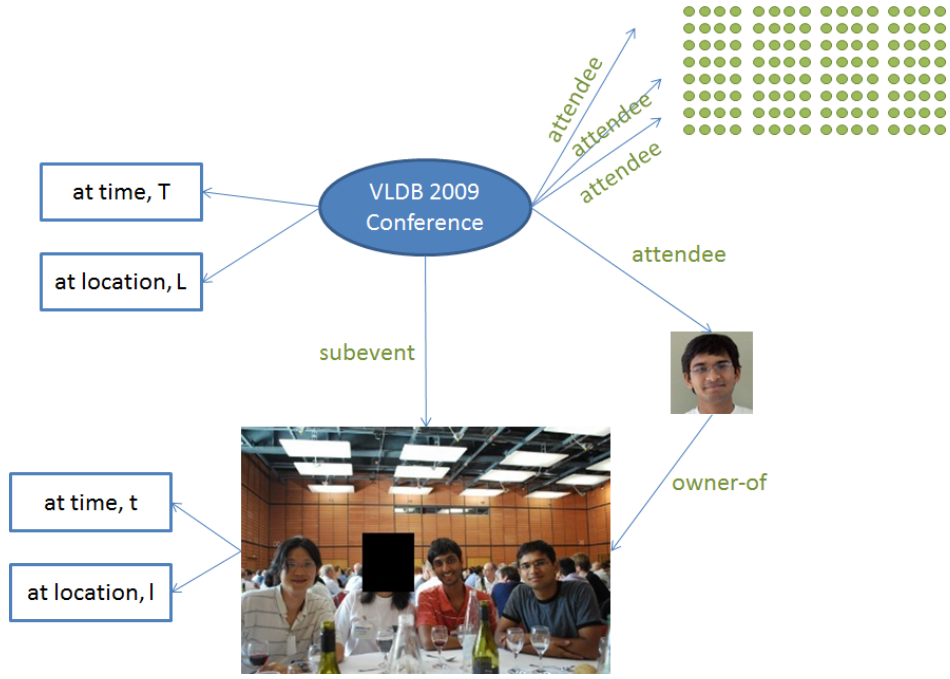

Figure 4.18: Context Network after discovering social relations.

Figure 4.19: Sources used so far.

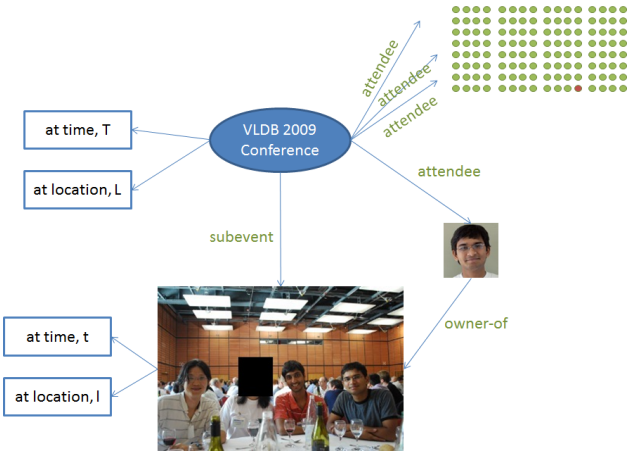

Figure 4.20: Context Network after discovering further social relations.

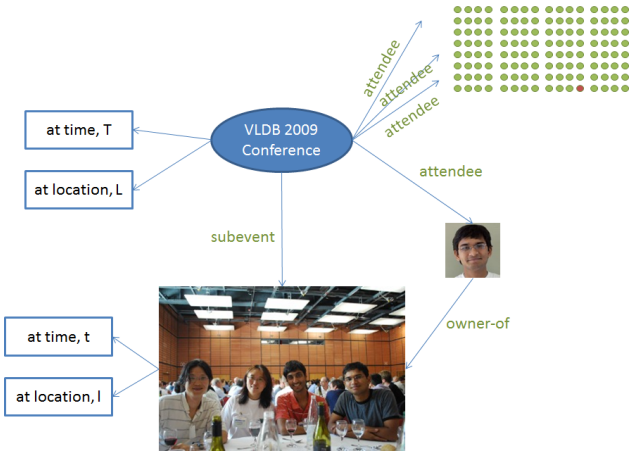

Figure 4.21: Context Network after tagging all faces.

Figure 4.22: Sources used to tag all faces.

is a participant of E if the event's spatiotemporal span contained that of the participant.

$$\texttt{participant}(\text{E, P}) \iff (\mathcal{L}_P \sqsubset_L \mathcal{L}_E) \wedge (\mathcal{T}_P \sqsubset_T \mathcal{T}_E) \tag{4.1}$$

The symbols $\sqsubset_L$ and $\sqsubset_T$ indicate spatial and temporal containment respectively. Please refer to [53] for more details. In later sections, we refer to the location and time of the event, $\mathcal{L}_E$ and $\mathcal{T}_E$ as E.**occurs-at** and E.**occurs-during** respectively.

4.5 Data Sources

The ontology makes available a vocabulary of classes and properties. Using this vocabulary, we can now declaratively specify the schema of each source. With these schema descriptions, CueNet can infer what data source can provide what type of data instances. For example, the framework can distinguish between a source which describes conferences and another which is a social network. We use a LISP like syntax to allow developers of the system to specify these declarations. The example below describes a source containing conference information.

```
(:source conferences
   (:attrs a_url a_name a_time a_location a_title)
   (:rel c-conf type-of cuenet:conference)
   (:rel d-time type-of dolce:time-interval)
   (:rel d-loc type-of dolce:location)
   (:rel d-attendee type-of dolce:person)
   (:rel d-attendee participant-in conf)
   (:rel c-conf occurs-at loc)
   (:rel c-conf occurs-during time)
```

```
(:axioms
   (:map d-time a_time)
   (:map d-loc a_location)
   (:map c-conf.title a_ltitle)
   (:map c-conf.url a_url)
   (:map d-attendee.name a_name)))
```

A source declaration comprises of a single nested s-expression. We will refer to the first symbol in each expression as a keyword, and the following symbols as operands. This above declaration uses five keywords (`source`, `attrs`, `rel`, `axioms`, `map`). The `source` keyword is the root operator, and declares a unique name of the data source. The source mapper can be queried for finding accessors using this name. The `attrs` keyword is used to list the attributes of this source. Currently we assume a tuple based representation, and each operand in the attrs expression maps to an element in the tuple. The `rel` keyword allows construction of a relationship graph where the nodes are instances of ontology concepts. And edges are the relationships described by this particular source. In the above example, we construct individuals *c-conf*, *d-time*, *d-loc* and *d-attendee* who are instances of the *conference* (from cuenet namespace in the OWL ontology), **time-interval**, **location** and **person** class (from the Dolce namespace) respectively. We further say that attendee is a ***participant-of*** the conference, which ***occurs-at*** location loc and ***occurs-during*** the interval time. Finally, the `mapping axioms` are used to map nodes in the relationship graph to attributes of the data source. For example, the first axiom (specified using the map keyword) maps the time node to the time attribute. The third map expression creates a literal called title, and associates it to the conference node, whose value comes from the ltitle attribute of the conference data source.

Formally, we represent the given ontology as O. The various classes and properties in O are represented by C^O and P^O respectively. Since our upper ontology consists of DOLCE

and E*, we assume the inclusion of the classes `Endurant`, `Perdurant`, `Event` and `Person` in C^O. Each source S consists of three parts, a relation graph $G^S(V^S, E^S)$ where the nodes $V^S \in C^O$, specify the various "things" described by the source. The edges $E^S \in P^O$ specify the relations among the nodes. Any graph retrieved from such a source is an instance of the relation graph, G^S. Further, the tuple A_T^S consists of the attributes of the data source. Finally, the mapping $M^S : \{G^S \rightarrow A_T^S\}$ specifies how to map different nodes in the relation graph to the different attributes of the native data source.

4.6 Conditions for Discovery

CueNet is entirely based on reasoning in the event and entity (i.e., person) domain, and the relationships between them. These relationships include participation (event-entity relation), social relations (entity-entity relation) and subevent relation (event-event). For the sake of simplicity, we restrict our discussions to events whose spatiotemporal spans either completely overlap or do not intersect at all. We do not consider events which partially overlap. In order to develop the necessary conditions for context discovery, we consider the following two axioms:

Object Existence Axiom: Objects can be present only at one place during a moment. The entity cannot exist outside a spatiotemporal boundary containing this place.

Participation Semantics Axiom: If an object is participating in two events at the same time, then one is the subevent of the other.

Given, the ontology O, we can construct event instance graph $G^I(V^I, E^I)$, whose nodes are instances of classes in C^O and edges are instances of the properties in P^O. The context discovery algorithm relies on the notion that given an instance graph, *queries* to the different sources can be automatically constructed. A query is a set of predicates, with one or more

unknown variables. For the instance graph $G^I(V^I, E^I)$, we construct a query $Q(D, U)$ where D is a set of predicates, and U is a set of unknown variables.

Query Construction Condition: Given an instance graph $G^I(V^I, E^I)$ and ontology $O(C^O, P^O)$, a query $Q(D, U)$ can be constructed, such that D is a set of predicates which represent a subset of relationships specified in G^I. In other words, D is a subgraph induced by G^I. U is a class, which has a relationship $r \in P^O$, with a node $n \in D$. Essentially, the ontology must prescribe a relation between some node n through the relationship r. In our case, the relation r will be either a **participant** or **subevent** relation. If the relationship with the instances does not violate any object property assertions specified in the ontology, we can create the query $Q(D, U)$.

Identity Condition: Given an instance graph $G^I(V^I, E^I)$, and a result graph $G^R(V^R, E^R)$ obtained from querying a source, we can merge two events only if they are identical. Two nodes $v_i^I \in V^I$ and $v_r^R \in V^R$ are identical if they meet the following two conditions **(i)** Both v_i^I and v_r^R are of the same class type, and **(ii)** Both v_i^I and v_r^R have exactly overlapping spatiotemporal spans, indicated by the $=_L$ and $=_T$. Mathematically, we write:

$$\begin{aligned} v_i^I = v_r^R \iff & (v_i^I.\textbf{type-of} = v_r^R.\textbf{type-of}) \wedge \\ & (v_i^I.\textbf{occurs-at} =_L v_r^R.\textbf{occurs-at}) \wedge \\ & (v_i^I.\textbf{occurs-during} =_T v_r^R.\textbf{occurs-during}) \end{aligned} \tag{4.2}$$

Subevent Condition: Given an instance graph $G^I(V^I, E^I)$, and a result graph $G^R(V^R, E^R)$ obtained from querying a source, we can construct a subevent edge between two nodes $v_i^I \in V^I$ and $v_r^R \in V^R$, if one is spatiotemporally contained within the other, and has at least one common `Endurant`.

$$\begin{aligned} & v_i^I \sqsubset_L v_r^R, \\ & v_i^I \sqsubset_T v_r^R \end{aligned} \tag{4.3}$$

$$v_i^I.\textbf{Endurants} \cap v_r^R.\textbf{Endurants} \neq \{\phi\} \tag{4.4}$$

Here $v_i^I.\textbf{Endurants}$ is defined as a set $\{w|w \in V_i^I \wedge w.\text{type-of=Endurant}\}$. If equation (4.4) does not hold, we say that v_i^I and v_r^R co-occur.

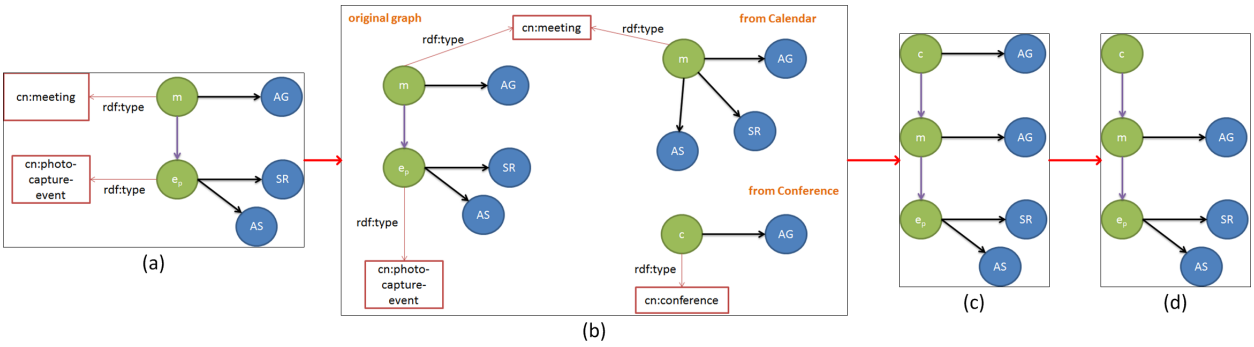

Figure 4.23: The various stages in an iteration of algorithm 1.

Merging Event Graphs: Given the above conditions, we can now describe an important building block for the context discovery algorithm: the steps needed to merge two event graphs. An example for this is shown in figure 4.23(b-d). Given the event graph consisting of the photo capture event on the left of (b) and a meeting event m and conference event c, containing their respective participants. In this example, the meeting event graph, m is semantically equivalent to the original graph. But the conference event, c is telling that the person AG is also participating in a conference at the time the photo was taken. The result of merging is shown in (d). An event graph merge consists of two steps. The first is a `subevent hierarchy join`, and the second is a `prune-up` step. Algorithm 2 presents a detailed algorithm discussion of merging event graphs. Here, we will look at only at it briefly.

Given an original graph, O_m, and a new graph N_m, the join function works as follows: All nodes in N_m are checked against all nodes in O_m to find identical counterparts. For entities, the identity is verified through an identifier, and for events, equation (4.2) is used.

Because of the entity existence and participation semantics axioms, all events which contain a common participant are connected to their respective super event using the subevent relation (equations (4.3) and (4.4) must be satisfied by the events). Also, if two events have no common participant, then they can be still be related with the subevent edge, if the event model says it is possible. For example, if in a conference event model, keynotes, lunches and banquets are declared as known subevents of an event. Then every keynote event, or banquet event to be merged into an event graph is made a subevent of the conference event, if the equation (4.3) holds between the respective events.

It must be noted that node AG occurs twice in graph (c). In order to correct this, we use the participation semantics axiom. We traverse the final event graph from the leaves to the root events, and remove every person node if it appears in a subevent. This is the `prune-up` step. Using these formalisms, we now look at the working of the context discovery algorithm.

4.7 Context Discovery Algorithm

Algorithm 1 below outlines the discovery algorithm, denoted as the **discover** function. The function is tail recursive, invoking itself until a termination condition is reached (when at most k tags are obtained for all faces or no new data is obtained from all data sources for all generated queries). The input to the algorithm is a photo (with EXIF tags) and an associated owner (the user). It must be noted that by seeding the graph with owner information, we bias the discovery towards his/her personal information. An event instance graph is created where each photo is modeled as a photo capture event. Each event and object is a node in the instance graph. Each event is associated with time and space attributes. All relationships are edges in this graph. All EXIF tags are literals, related to the photo with data property edges. Figure 4.23 graphically shows the main stages in a single iteration of the algorithm.

The event graph is traversed to produce a queue of event and object nodes, which we shall refer to as DQ (discovery queue). The algorithm consists of two primary functions: **query** and **merge**. The behavior of the query function depends on the type of the node. If the node is an event instance, the function consults the ontology to find any known sub-events, and queries data sources to find all these subevents, its properties and participants of the input event node. On the other hand, if it is an entity instance, the function issues a query to find all the events it is participating in.

Results from data source wrappers are returned in the form of event graphs. These event graphs are merged into the original event graph by taking the following steps. First, it identifies **duplicate** events using the conditions mentioned above. Second, it identifies subevent hierarchies using the graph merge conditions described above, and performs a **subevent hierarchy join**. Third, the function **prune_up** removes entities from an event when its subevent also lists it as a participant node. Fourth, **push_down** is the face verification step if the number of entities in the parents of the photo-capture events is small (less than T).

The **push_down** step will try to verify if any of the newly discovered objects are present in the photo and if they are (if the tagging confidence of this object, obtained from the face verification algorithm, is higher than the given threshold), the objects are removed from the super event, and linked to the photo capture event as its participant. In other words, they are pushed down the subevent hierarchy. Alternatively, if the number of new objects is larger than T, the algorithm initiates the **vote-and-verify** method, which ranks all the candidates based on social relationships with people already identified in the photo. For example, if a candidate is related to two persons present in the photo through some social networks, then its score is 2. Ordering is done by simply sorting the candidate list by descending order of score. The face verification runs only on the top ranked k candidates. If there are still untagged faces after the termination of the algorithm, we vote over all the remaining people, and return the ranked list for each untagged face.

Algorithm 1: The Context Discovery Algorithm

Data: A photograph H, with a set of detected faces F. Voting threshold, T. The owner O of the photo.

Result: For each face f ∈ F, a set of atmost k person tags.

```
1  begin
2
3      function discover(): {
4          while (DQ is not empty): {
5              node = DQ.deque()
6              results = query (node)
7              E ← merge (E, results)
8              if (termination_check()):
9                  return prepare_results();
10         }
11         reconstruct DQ ← E
12         discover()
13     }
14
15     function merge(O, N): {
16         remove_duplicates()
17         M ← subevent_hierarchy_join(O, N)
18         prune_up(M)
19         if (less than T new candidates were discovered):
20             push_down(M)
21         else:
22             vote_and_verify(M)
23         return M;
24     }
25
26     E ← construct event graph with H and O
27     construct discoverable nodes queue, DQ ← E
28     return discover()
29 end
```

Figure 4.23 shows the various stages in the algorithm graphically. (a) shows an example event graph describing a photo taken at a meeting. The meeting consists of three participants AG, SR and AS. The photo contains SR and AS. (b) shows two events returned from the data sources. One is a meeting event which is semantically identical to the input. The other is a conference event with AG. (c) shows the result of merging these graphs. (d) The `prune-up` function removes the duplicate reference to AG. A live visualization of these steps for different

photos can be found at http://cuenet.site44.com.

4.8 Merging Context Networks

In this section, we look more closely at the merge function. Algorithm 2 presents the pseudo-code for merging a secondary context network, S, into a primary context network P. A terminology of primary and secondary is to signify that all data instances from a secondary network will be merged into a primary network. While merging networks, we also assume that events have atmost one super-event. Thus, no diamond structures are found in either network. Algorithm 2 shows the steps needed to merge two networks each with a single root. A root event is one which has no super-events. The symbol $\forall_{se}$ stands for "for all subevents".

Once the two root nodes, Pr and Sr have been identified, we descend the subevent trees of the two context networks, and do one of the following operations. For each subevent of Pr, we check if any subevent is equal to Sr, then merge the information from Sr to this subevent, and continue recursively merging the children of Sr into the children of the subevent. If a subevent of Pr contains the new instance, Sr, we simply continue recursion. But, if Sr can contain the subevent node, then we add of the siblings sub-events which can become subevents of Sr to a list `containedSubevents`, and add Sr as a subevent of Pr, remove all the children from Pr in `containedSubevents`, and add them as subevents of Sr. Recursion is continued on the children of Sr from the newly connected Sr node in the primary network. Any new literal properties, spatio-temporal attributes and participant information in events which exist in the primary networks are copied using the `mergeInformation` function.

Let's consider an example to understand the merge function. Assume we have to merge network 4.24(b) into 4.24(a). To simplify the example, we will assume that all the events are occurring at the same location. The time intervals occupied by the events are provided

Algorithm 2: The Merge Algorithm

Data: Context Network P and S, where S will be merged into P.
Result: All event instances in S will be merged into P.

```
1  begin
2
3      function merge(P, S): {
4          [Pr] = root event of P
5          [Sr] = root event of S
6          recursive_merge(P, Pr, S, Sr)
7      }
8
9      function recursive_merge(P, Pr, S, Sr): {
10         addAsSubevent = true
11         containedSubevents ← {φ}
12         ∀se (ps of P) {
13             if (equals(ps, Sr)) {
14                 addAsSubevent = false
15                 mergeInformation(ps, Sr)
16                 ∀se (s of Sr): recursive_merge(P, ps, S, se)
17             }
18             if (contains(ps, Sr)) {
19                 addAsSubevent = false
20                 recursive_merge(P, ps, S, Sr)
21             }
22             if (contains(Sr, ps)) {
23                 addAsSubevent = false
24                 containedSubevents.add(ps)
25             }
26         }
27
28         if (addAsSubevent) {
29             addSubeventEdge(Pr, Sr)
30             if (containedSubevents.size() > 0) {
31                 removeSubevents(Pr, containedSubevents)
32                 createSubevents(Sr, containedSubevents)
33             }
34             ∀se (s of Sr): recursiveMerge(subtree, Sr, other, s);
35         }
36     }
37
38 end
```

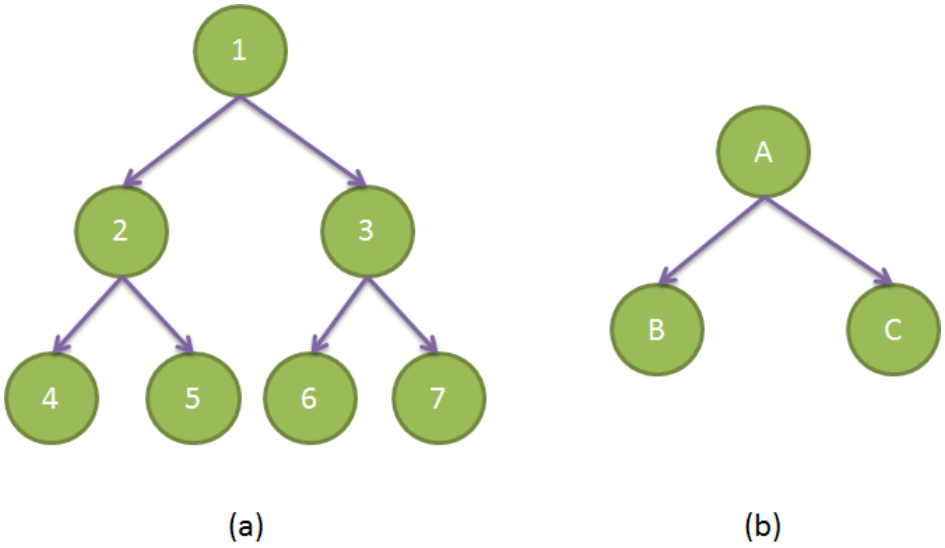

Figure 4.24: (a) Primary and (b) Secondary Example Networks.

in the table 4.1 below.

Event	Time Interval
1	0 - 120
2	10 - 45
3	70 - 100
4	10 - 15
5	28 - 37
6	70 - 80
7	90 - 100

Event	Time Interval
A	10 - 105
B	10 - 45
C	80 - 90

Table 4.1: Time Intervals for Events in Primary and Secondary Networks shown in 4.24

We first check if the root of the primary would contain the secondary. Since, event `1`'s time interval completely contains event `A`, this is true, and we initiate the merge function where `Sr=A, Pr=1`. The subevents `2, 3` of event `1` are both contained within A, the lines 23-24 from the algorithm populate the `containedSubevents` list. This list is used to remove the edges between `1` and `2, 3` and create subevent edges between `1` and `A` and A and `2, 3` in the lines 28-36. The new primary network is shown in figure 4.25(a). Now, the algorithm recursively proceeds to add events `B, C` as subevents of A in the primary network `Sr=A, Pr=B`. Since the temporal extents of `2` and `B` are identical, the event `B` is merged into `2` as

shown in 4.25(b). All participant information which exists in `B`, and not in `2` is copied. This is done in lines 13-17 of the algorithm. Finally, the event `C` is compared with event `3`, and found to be contained within it. The recursion proceeds to the subevents of `3` in lines 18-21. Here, the algorithm realizes that none of the events `6, 7` could contain `3`, and hence it is made a subevent of `3` itself. This is done at the line 29. Figure 4.25(c) shows the final primary network with all events from the secondary merged into it.

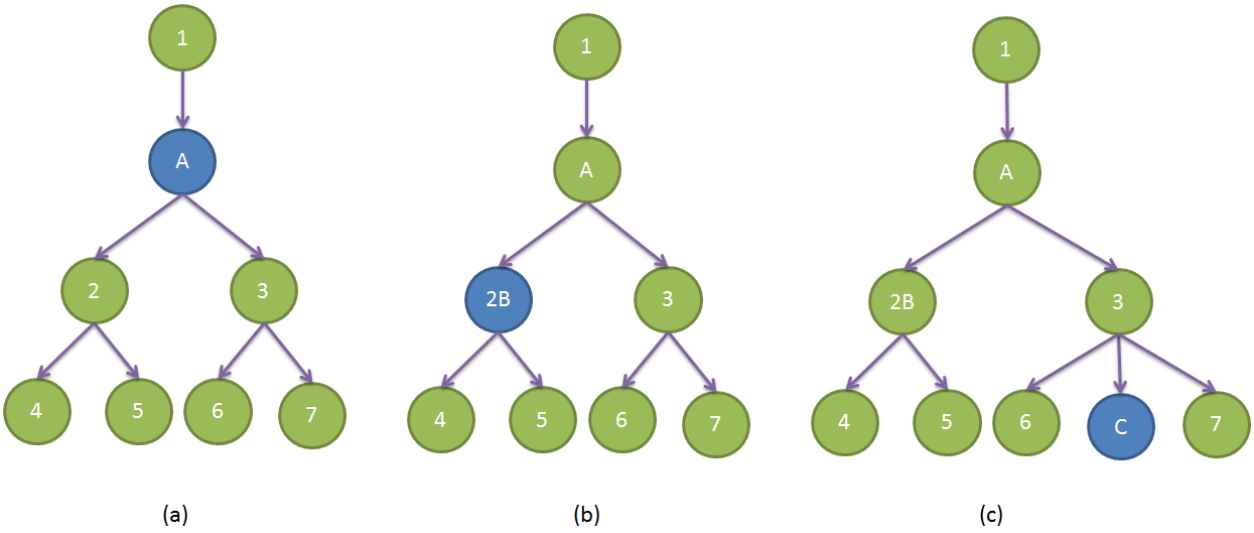

Figure 4.25: The three steps taken to merge the different nodes of the secondary into the primary network.

4.9 Implementation

In this section, we will look at the current implementation of the CueNet framework. Figure 4.26 shows a detailed conceptual architecture of the framework. It consists of three verticals: the data integration modules on the left side, the discovery algorithm and its supporting components in the middle, and the candidate management and face verification tools on the right side. Let's look at them one by one. The data integration module is configured using a source script. This script contains the schema declarations of the various sources, and specifications of how the objects in these sources are related to the objects in the ontology.

The complete script listing is provided in Appendix A. We use JavaCC to construct the compiler for this script. The schema and the relations are stored as graphs which can be queried by other modules. Any discover query to be executed on the multitude of sources is checked upon each graph to see if can respond to the type of query. For example, exclude social networks when queries are requesting event information. Each source requires mediator code to align its content with that of the tuples requested in the query engine. This approach is more reminiscent of GAV designs in data integration literature. We chose this because of simpler query processing designs, and the fact that although we required to plug-in and out sources easily, these sources themselves changed very rarely. At the bottom of the data integration stack, we find some common utilities to make HTTP requests and DB query requests, which are used by the source mediators. The discover queries are created in SPARQL. The query engine uses Jena's SPARQL query processing framework, and extracts triple patterns and predicates using the framework's visitor pattern capabilities.

Now, lets look at the various components making up the discovery algorithm. Our knowledge base is specified using OWL ontologies. This is parsed using the open source Jena library, and the objects and relations specified are loaded into in memory DAGs. When a tag discovery request is made by a user, the discovery algorithm is initialized with a photo-capture-event, EXIF information is extracted through perl-based exiftool, and the seed context network is constructed and provided to the discovery algorithm. The algorithm initiates the discover and merge operations using the query engine, and the candidate manager. As new relations are discovered, the related entities are then checked for their presence in the photo using verification tools. We used face.com until its termination in mid-2012, after which we use the system based on [68] until early 2013, after which we resorted to manual verification upon the top ranked candidates.

When a tag request is initiated, a message is passed to the candidate manager, which scans the user's database to create a list of all entities. This is done by invoking scan queries on

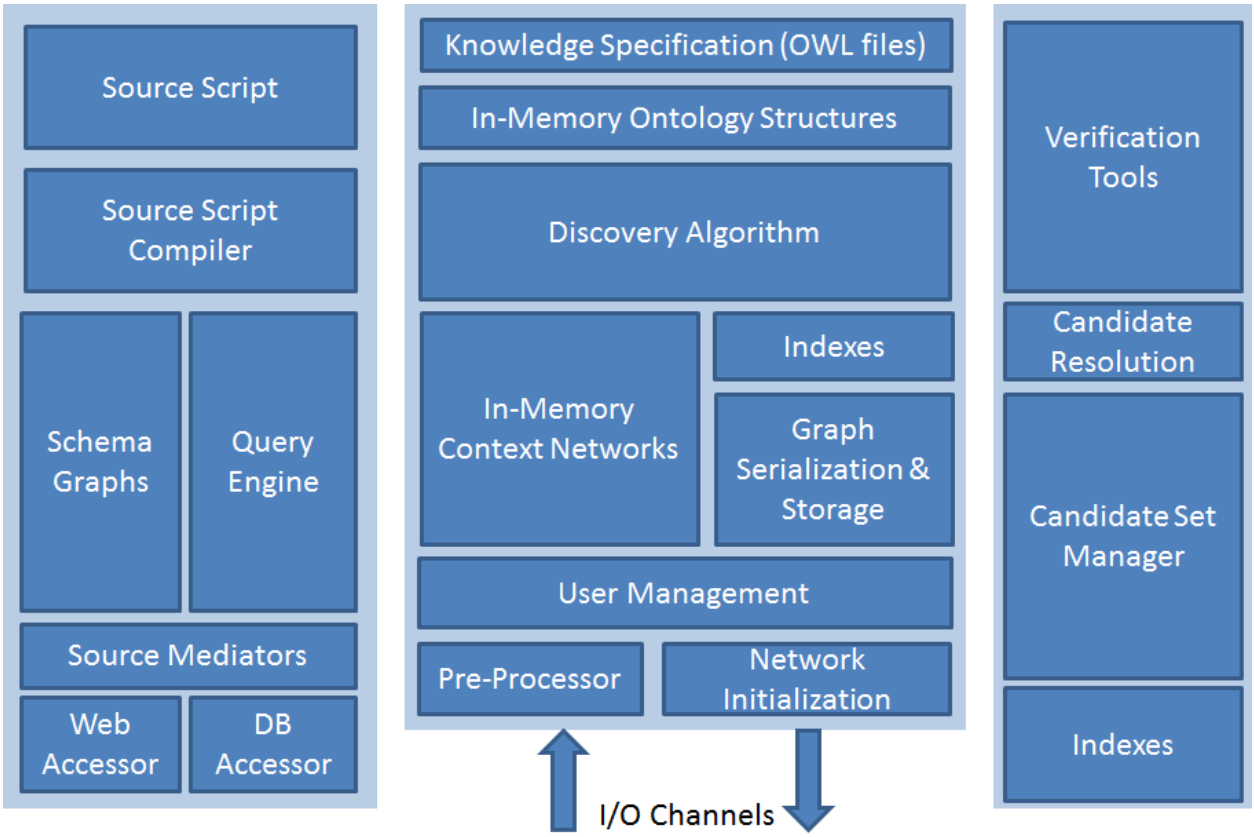

Figure 4.26: Main components of the current CueNet implementation.

all data sources through the query engine. Each data sources has a different representation for the same entity. The job of the candidate manager is to aggregate entities by using cues based on their common names or emails, pass them to an entity resolver (if the data is stored on a web-page), and if needed ask a user to correct any mistakes it made in this process. Ultimately, at the end of its processing, we obtain a list of candidates, each which holds identifiers which are local to each data source. When a context network is created by the source mediator, it looks up the candidate reference through the candidate manager before emitting the network back to the discover algorithm. This step makes the nature of data sources oblivious to the discovery algorithm.

The web interface to the system is shown in figure 4.27. This stack is essential in aggregating data from various sources, and is instrumental in dealing with their intricacies related to

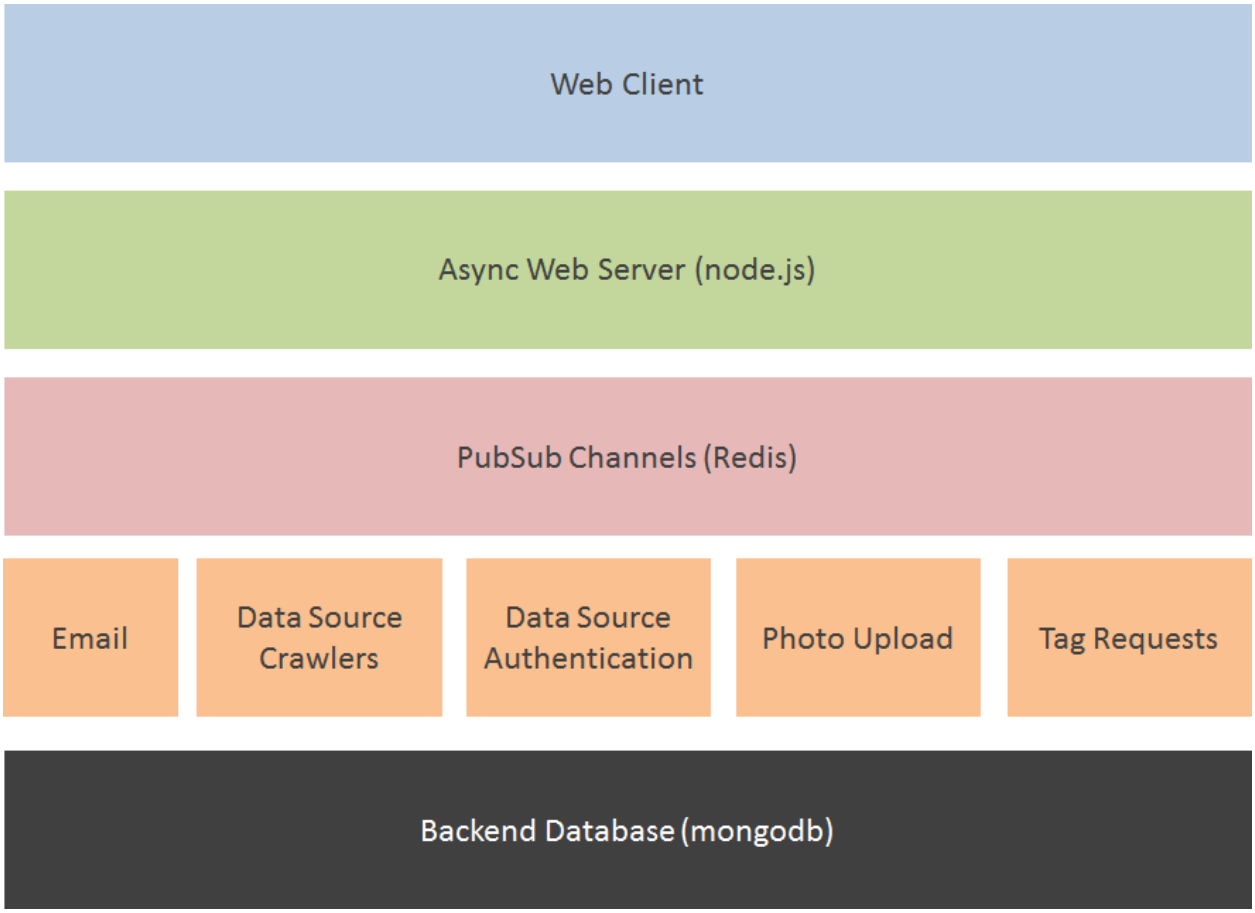

Figure 4.27: Web stack for data aggregation.

authentication, API requests and data formatting. Sources like Facebook and Google have rigid authentication mechanisms which need to be bypassed before any personal data can be accessed. Obtaining user permission needs to be done via a web client for Google APIs, whereas Facebook requires a valid session key from a user. These small details prompted us to implement the data aggregations using Javascript APIs which are served using an asynchronous web server based on node.js. The asynchronous nature of the server allows to develop web clients which execute a large number of HTTP callbacks without worrying about multi-threading issues – a design decision which is becoming increasingly popular for web architecture. Once the data has been aggregated at the client, it is passed back to the server which pushes it to specific process which can deal with processing it. For example, email information is sent to an independent Python process which initiates the IMAP protocol on

the user's message server to aggregate emails. The Facebook events, social network, photo tag and Google calendar information is sent to a script which loads them into appropriate mongodb collections. Unstructured data like conference webpages are sent to a process which initiates the SNER over them to aggregate entities. These entities are stored in appropriate collections. The server communicates through these processes using REDIS, an in memory cache/database which supports PubSub channels. Multiple channels are created, one for each *type* of data source. Any process can listen to it, obtain the data, process it of its own accord.

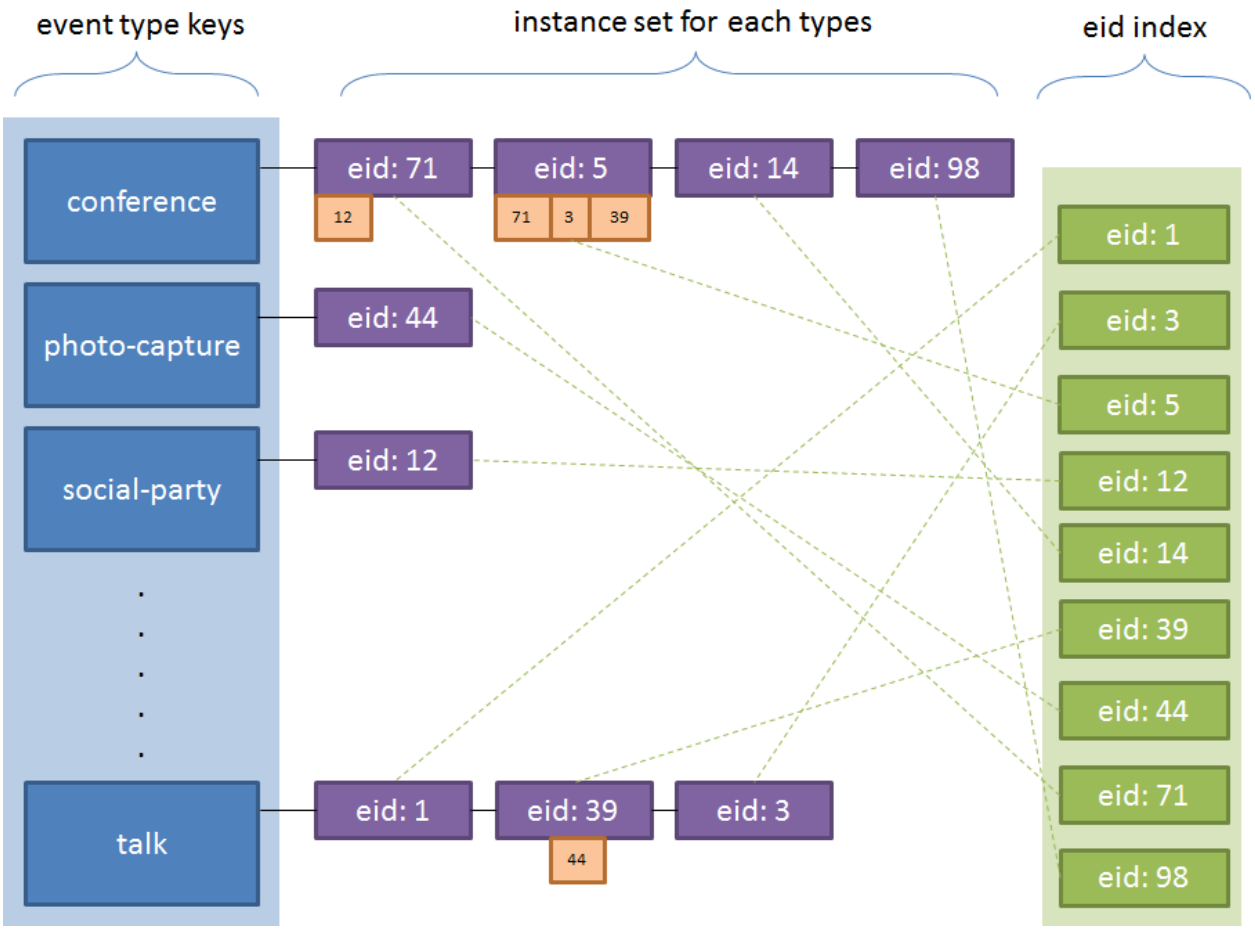

Figure 4.28: The data structure for maintaining Context Network.

Lastly, we will describe the data structures used to hold context networks. A context network is a directed graph containing event, and entity nodes, and their relationships. Any standard directed graph implementation can be used to hold this, but given the operations we perform

on these network, some additional properties can be exploited to achieve faster merges. One operation we need often is to eliminate merging networks which provide no new information. For this purpose we modify the adjacency list structure to look like the one in figure 4.28. A traditional adjacency list holds a hash-table for nodes, and each node is associated with a linked list or another hash-table for the nodes which it connects to. Since IDs don't hold any special meaning between context networks, we cannot rely only on them to distinguish between events in separate context networks. Event identity is established based on type, spatial and temporal information. For this purpose, our primary hash-table contains the **types** of events present in the network. In the example figure 4.28, we see the events of types `conference`, `photo-capture`, `social-party` and `talk` are contained in the network. For each event type, we associate a list of instances to it. This is implemented currently using a HashSet. Each instance in the context network is identified by a unique instance Id. This architecture allows us to quickly compare new events to existing ones and eliminate duplicates.

Edge relations are maintained by a list of (instance-id, edge-type) pairs. Thus traversing a network implies starting from a root node, looking up the instance Ids of the outgoing edges by scanning the HashSet, looking the outgoing nodes in the adjacency list, and repeating the process. In order to avoid multiple lookups, we maintain an additional hash-table over instance IDs. Given an instance ID, this table allows fast lookups to avoid time spent in scanning very large sets which are created when there are many instances of the same type of events.

This architecture suffices because we discover context for a single photo at a time. In the case where an application needs to discover context for many events, we use a separate network for each atomic event under consideration, and assume that the context for one event is independent of the other. Thus, a merge operation reduces to finding the existing context network it can be merged with, and performing the merge only with one such network. This

assumption will not be true if the events are spatio-temporally very close to each other. Here, information about one event will affect events in the other. This is beyond the scope of this dissertation. We also use a R-Tree built using the Java Spatial Index library to index the time intervals of the events of the various context networks (which do not have any super-events themselves), to reduce the search during the merge step. This increases the overall efficiency of discovery algorithm.

In this chapter, we looked at the basic foundations of a context discovery framework. We saw the novel algorithms to discover and merge contextual information into the context network of a given photo, and the different design decisions that were taken to achieve our current implementation. We saw Participation Semantics and Object Existence axioms and the critical role they play in the algorithms. In the next chapter we will evaluate this framework on a set of real world photos and simulated information to test its efficacy and performance aspects.

Chapter 5

Experiments and Analysis

In this chapter, we will present experiments to show the efficacy of the context discovery approach. Experiments are performed on real-world data from real users using a variety of social and personal media applications, public data sources. The motivation behind these experiments is to tag faces in thousands of personal photos taken by various people at different kinds of events. We will also perform experiments on generated data to simulate large scale use cases. Here, the motivation is to analyze the performance characteristics of the merge algorithm.

5.1 Experiments on Real-World Data

In this section, we analyze how CueNet helps tags personal photos taken at different events. For a given users, we will construct a dataset consisting of photos taken at a particular event. For each of these users, we will create a candidate set by aggregating people over personal, public and social data sources. In order to evaluate CueNet, we will attempt to reduce this candidate set, and analyze how many of the faces can be correctly tagged by this

reduced set. We will also compare this performance over metrics like location based ranking, where candidates are ranked according to their last known location, and (if time permits) tie strength. Our final conclusion is that, in order to rank candidates for tagging faces in photos, CueNet provides an event-agnostic platform, where as other techniques perform inconsistently across different types of events.

5.1.1 Setup

We use photos taken during three different **types of events**: Social Parties, Academic Conferences and Trips. This diversity allows for different distributions of face tags across different aspects of a user's life. Social Parties generally tend to have close friends who are spatially co-located. Conferences tend to have people from different parts of the world, but those who are affiliated with the area of the conferences. Trips, cannot always rely on location as a useful metric and can involve people from either social, personal or professional circles from a user's life.

For each event type, we collect multiple datasets from 6 different people. A **dataset** consists of multiple photos during the event, the user's personal information, which contains information from sources like Google Calendar, personal email and profile information from social networks like Facebook and Twitter. We also collect a person's social networking information which consists of tweets written by the user or their friend during the time the event was occurring, the social network itself (friends on Facebook and Twitter, along with their profile information). Conference proceedings are downloaded from DBLP and the conference website. Facebook events are also obtained and stored in our database. Besides, location databases like Yahoo Placefinder were used to geocode addresses and reverse geocode EXIF GPS coordinates. We assume that all photos have a valid EXIF tag, especially the timestamp and GPS coordinates. This assumption is not a very hard one, as almost all photos

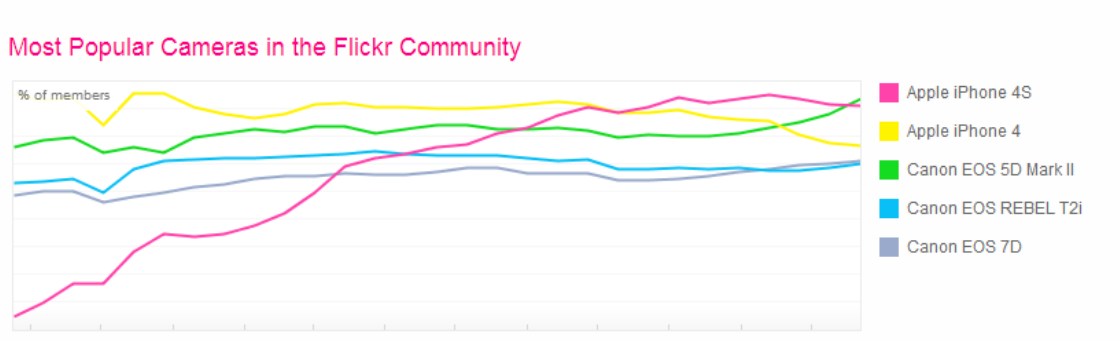

Figure 5.1: Popularity of iPhone at Flickr.com (October 2011).

captured in the last two years are through iPhone or Android smartphones, which add reasonably accurate GPS tags and accurate timestamps (where the phone clock is synced with the cell tower). The domination of iPhone in the photo market is shown in figure 5.1. The ground truth was annotated by the user with our annotation interface. For each photo, an annotation consisted of the ID of the person in the candidate set in it. Face verification was achieved initially with Face.com, but after their website was shutdown, we used the web service, Automatic Face Systems, maintained by Neeraj Kumar. After this service was terminated, we resorted to manual face verification. We use the manual verification for the experiments described below, unless otherwise mentioned.

In order to construct a **candidate set** for each dataset, we construct a so-called `CandidateSet`. A `CandidateSet` data structure performs two functions. It is a persistent set of candidates each of which have a unique identifier. Second, each data source, which has its own identifier for a candidate can look up the candidate set with this identifier to obtain the candidate set identifier. Here is an example of a simple candidate set:

```
[
{csid: 1010-poqs, name: 'Arjun Satish', email: 'arjun@uci.edu',
                                        email: 'arjun.satish@gmail.com',
                                        facebook-id: '656563332'}
{csid: 2010-pasd, name: 'Ramesh Jain', email: 'jain@ics.uci.ed',
```

```
                                        name: 'Dad',
                                        email: 'jain49@gmail.com',
                                        facebook-id: '10004290384'}
{csid: 1255-juen, name: 'Peter Boncz', confid: 'cf-12-candidate-322'}
{csid: 7585-kdye, name: 'Amarnath Gupta', email: 'gupta@sdsc.edu',
                                           twitter: 'aguptasd'}
{csid: 1111-bmel, name: 'Arjun', twitter: '@wicknicks'}
]
```

The above snippet is a part of the candidate set created during the tagging of my personal photos. A `CandidateSet` is a multimap where the keys are the unique IDs (shown as csid above). The values are a list of key-value attribute pairs. The key could be a global key, such as name or email address, which can be added by any data source. In the above example, different sources could contribute different names for the same entity. Same values for the same key are not duplicated. The list could contain keys which are local to a data source, for example the facebook identifier key: `facebook-id`. This primarily helps in querying the candidate set to obtain a reference to the candidate during the discovery phase.

The construction of the `CandidateSet` happens when the system is brought up. For each user, a unique `CandidateSet` is created, or loaded if it exists on disk. In order to create it, each data source is probed individually to provide a list of unique persons. The data source mediator checks if the user already exists in the `CandidateSet` through its local primary key, and if he does, adds additional key value pairs to the attribute list. If a candidate corresponding to its primary key does not exist, a new candidate is created in the set, and the local primary key is added as an attribute pair. This technique will fail to merge different identities of people if they use multiple email addresses or do not store email information on social networks. Thus, we create a merge file, which lists a set of merges between candidates. In the above example, we will use this functionality to combine

the entries with IDs `1111-bmel` and `1010-poqs` into a unified candidate. This gurantees that context discovered for one online identity is propagated to other identities of the same person.

We use 1889 photos taken at 17 different events in our face tagging experiment. Each photo contains one or more faces. We will denote each dataset as 'Di' (where $1 \leq i \leq 17$ for each dataset). Table 5.1 describes each dataset in terms of number of photos, unique annotations in ground truth, the year they were captured and the type of the event.

Dataset	Unique People	No. of Photos	Year	CandidateSet Size	Event Type
D1	43	78	2012	660	conference
D2	24	108	2012	660	conference
D3	6	16	2010	660	conference
D4	7	10	2010	660	conference
D5	36	80	2009	660	conference
D6	18	65	2013	660	conference
D7	7	11	2013	660	conference
D8	12	25	2009	1894	conference
D9	14	65	2011	215	party
D10	13	131	2010	561	party
D11	6	85	2008	656	party
D12	50	74	2012	1049	party
D13	19	330	2009	691	party
D14	14	363	2009	711	trip
D15	2	208	2010	715	trip
D16	4	217	2011	711	trip
D17	7	23	2011	584	trip

Table 5.1: Profile of datasets used in the experiments.

We divide the sources into different categories to facilitate a more general discussion. The categories are "Personal Information" (same as Owner Information in section 4.7), "Event sources", and "Social Networks". Event sources include Facebook events, Yahoo Upcoming web service, our conference events database among other sources. Social networks include Facebook's social graph. Personal information contained information about the user, and a link to their personal calendars. An annotation is considered "Out of Context Network" if

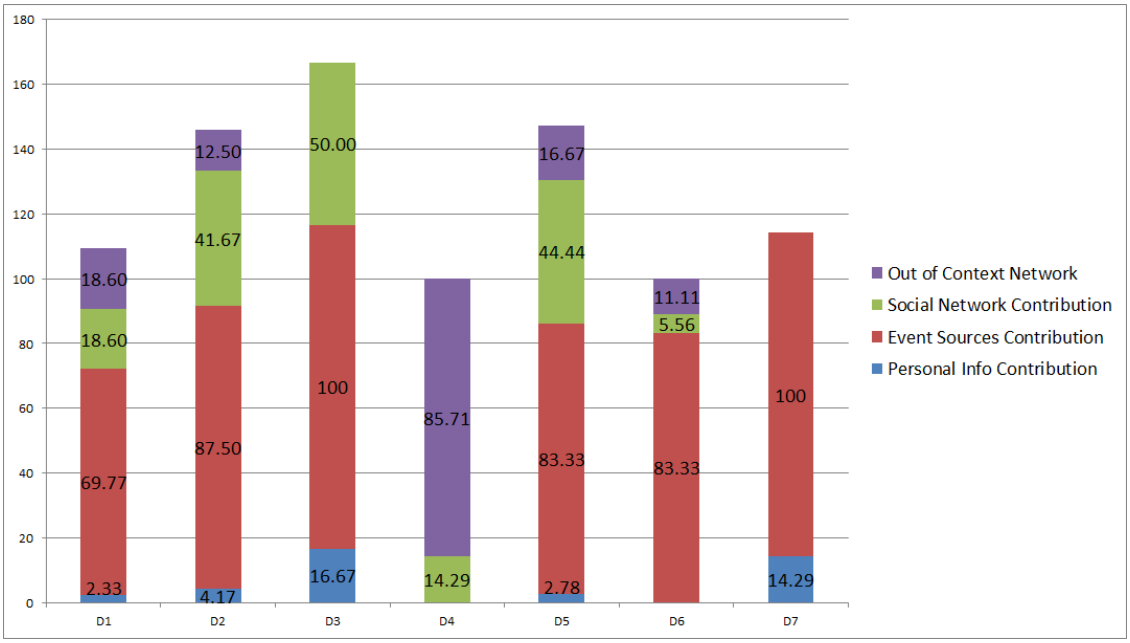

Figure 5.2: The distribution of annotations in the ground truth for conference photos across various sources.

it is not in any of these sources.

Figure 5.2 shows the distribution of the ground truth annotations of the conference datasets across various sources, for each dataset. For example, the bar corresponding to D2 says that 87.5% of ground truth annotations were found in event sources, 41.67% in social networks, 4.17% in personal information and 12.5% were not found in any source, and therefore marked as "Out of Context Network". From this graph it is clear that event sources contain a large portion of ground truth annotations. Besides D4, a minimum of 70% of our annotations are found in event sources for all datasets, and for some datasets (D3, D7) all annotations are found in event sources. The sum total of contributions will add up to values more than 100% because they share some annotations among each other. For example, a friend on Facebook might show up at a conference to give the keynote talk.

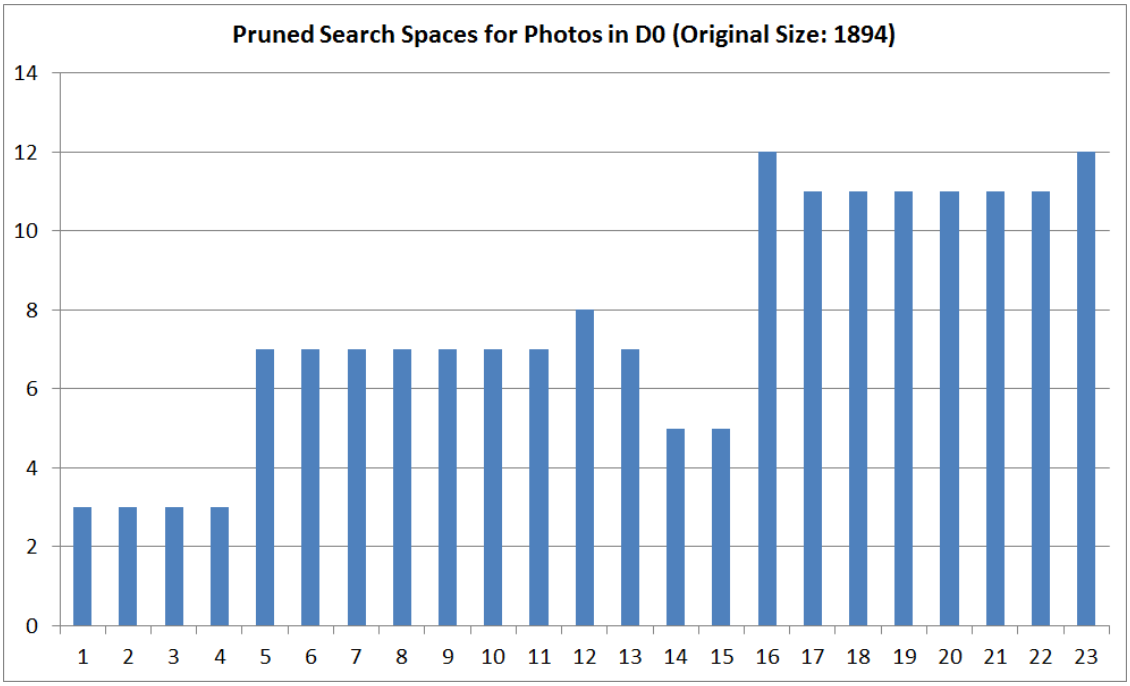

Figure 5.3: Pruned search space for photos in conference dataset D8.

5.1.2 Individual List Sizes in a Dataset

First, we look at how CueNet reduces the number of possible candidates for all photos in a dataset. For this setup, the complete candidate set L, contained 1894 labels (total number of people present at the conference, user's emails and social graph). The figure 5.3 shows various statistics for each photo, which includes the maximum size of the list which was generated by the discovery algorithm, the actual number of people in the photos, the number of true positives and false positives. As it can be seen, the size of the discovered set S, never exceeded 12. This is 0.5% of the original candidate list. Because the total number of possible participants (list size) was low, our False Positive rate (FP) was very low too. Most of the false positives were due to profile orientation of faces or obstructions (this was because the face detector was smart enough to pick up profile faces, but verification worked better only on frontal faces).

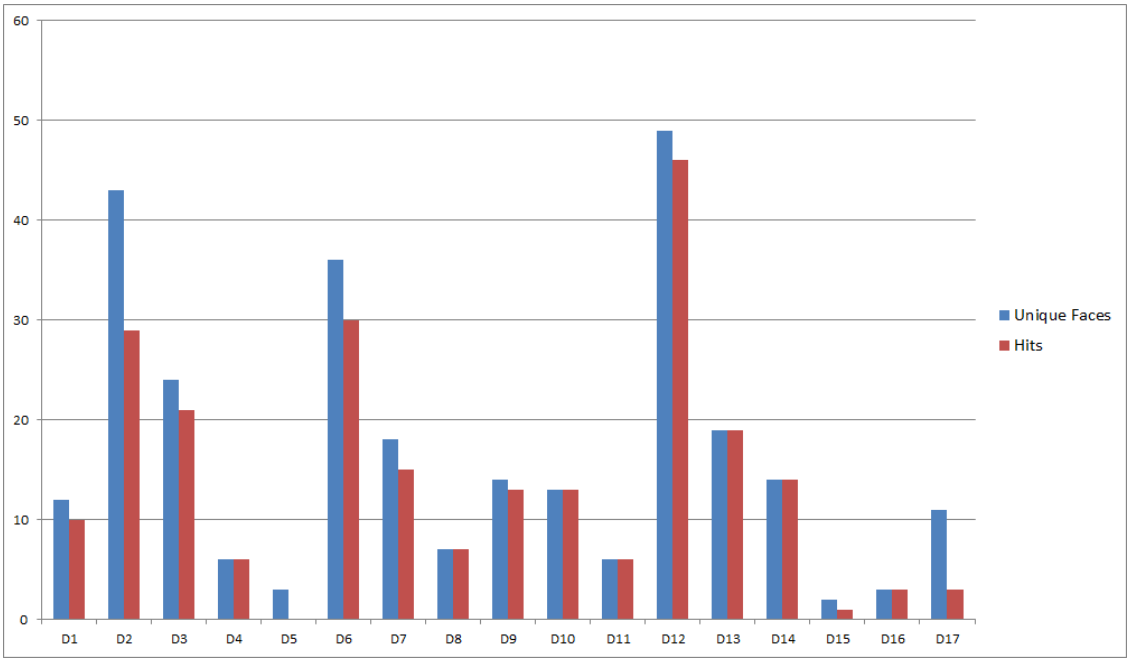

Figure 5.4: Hit counts for all datasets using Context Discovery Algorithm.

5.1.3 Results using Context Discovery

In order to evaluate 17 different datasets, we perform the following modification on each dataset. For a dataset, we select the photo with all annotations, and attempt to tag it. The numbers seen below will reflect the results for such a photo. Alternatively, the numbers below can be interpreted as tagging the event itself, and not individual photos. We do this modification to reduce the number of graphs drawn per experiment. We define a **hit** as a correctly tagged face. Figure 5.4 shows the hits for datasets using the context discovery algorithm. The blue bars show the number of unique faces in the dataset. As we can see the number of hits is equal to the number of unique faces, which implies that our context discovery algorithm was able to find the correct face tags for almost all datasets. Only for dataset 4, this difference is significant. This is because there was no contextual information related to the people in this dataset.

Most of the datasets contain hundreds of candidates. If we have to perform face verification on all datasets, then the discovery algorithm is not performing effectively. The fewer the

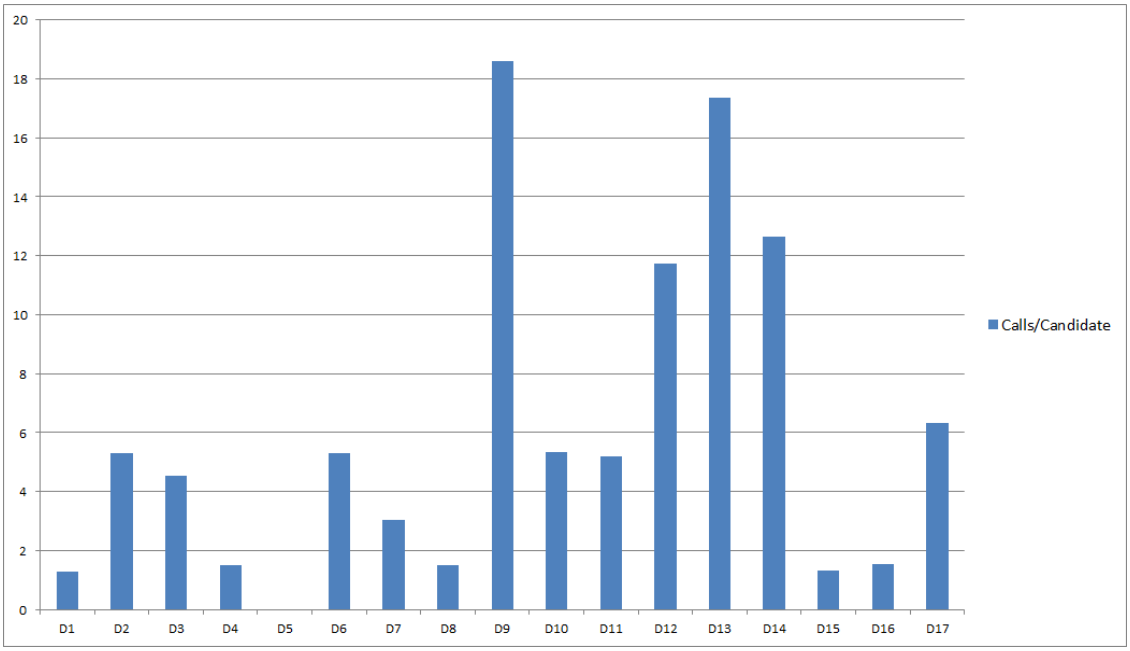

Figure 5.5: Verification Ratio (x 100) obtained using the Context Discovery Algorithm.

number of verifications, the better the accuracy of the algorithm. In order to quantify this, we introduce the metric *Verification Ratio*, which is simply the number of verifications done per candidate in the `CandidateSet`. If this number is equal to 1, that means we have performed a very large number of verifications. The closer it is to 0, the fewer the number of verifications that were done. This ratio also enables us to compare the differences across dataset, where the number of candidates vary. As it can be seen in figure 5.5, the verification ratio never exceeds 19% (the maximum value being 18.3%).

In order to compare these results, we perform a tagging experiment using location information alone. Candidates are ordered according to their last known location, and presented to the user in the same style as the information presented with the context discovery algorithm. In this case, it must be noted that some users do not have location information related to them, and are excluded from the candidate set. The effect of this fact is seen in the reduced hit count seen using location information.

Similarly, we measure the verification ratio using location only. The results in figure 5.7

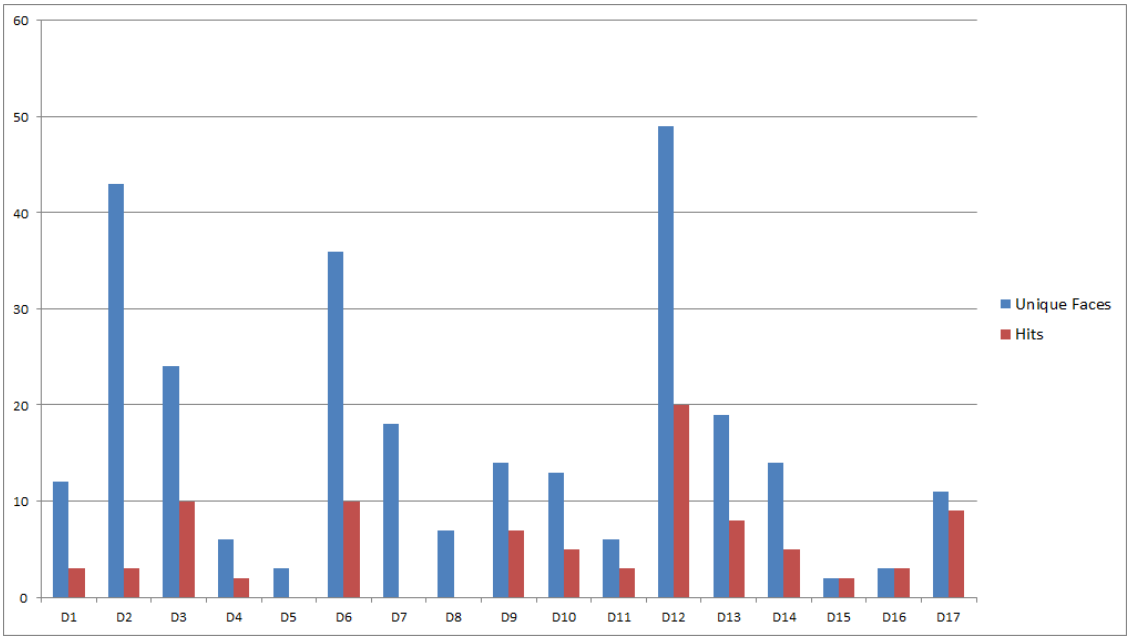

Figure 5.6: Hit counts for all datasets using Location only.

show that the maximum number has now increased to almost 29%.

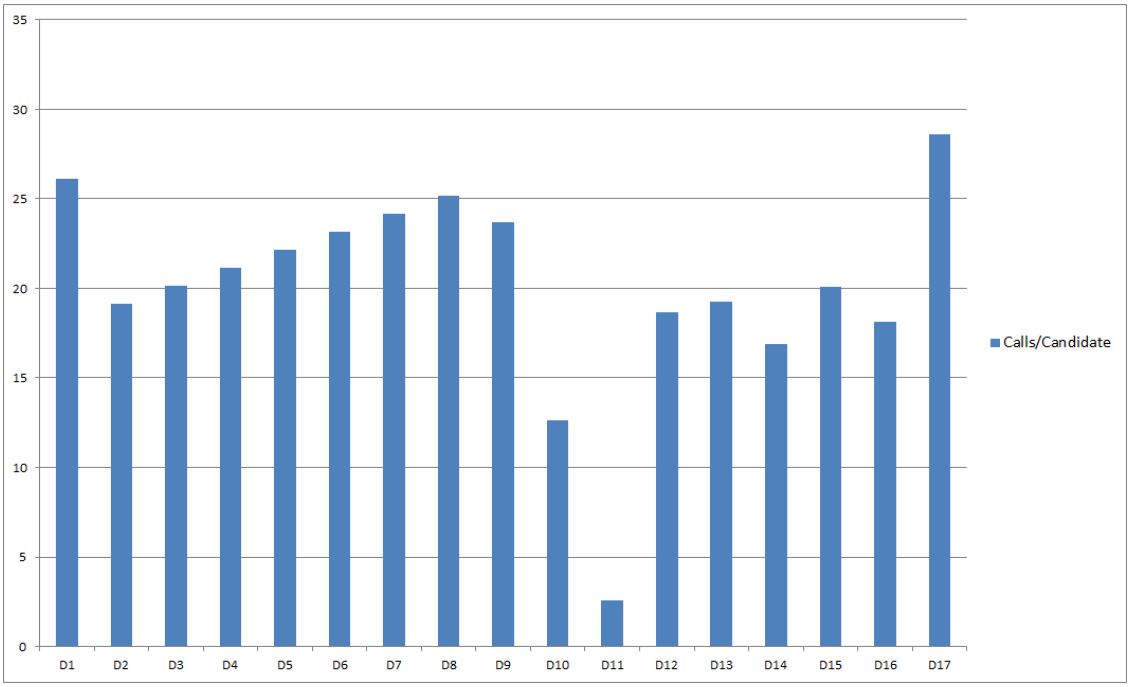

Figure 5.7: Verification Ratio (x 100) for all datasets obtained using Location only.

In figure 5.8, we compare the **hits ratio** of the context discovery algorithm with the locatio based algorithm. The hit ratio is the number of hits per unique face in a dataset.

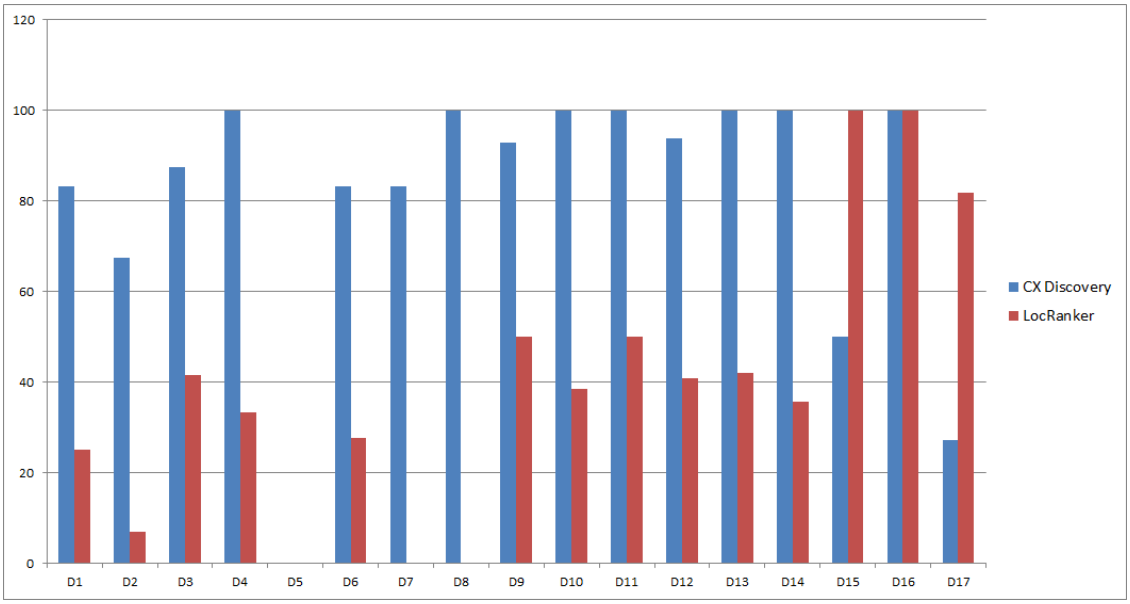

Figure 5.8: Comparing Hits Ratio (x 100) for all datasets.

As it can be seen, the context discovery algorithm outperforms the simple location based ranking algorithm. The more important insight in this graph is that the context discovery algorithm performs **consistently across different types of events**, whereas the location based metric is good only for the party events. This was because in such social gatherings, participants lived closed to the location of the event. And therefore, ranked very highly when only location was used as context. Figure 5.9 compares the different verification ratios. The relatively lower numbers for the context discovery algorithm indicate its superior ranking over the location based algorithm. Again, the important thing is to note its **performance across event types**.

5.1.4 Conclusion

We summarize the results of this experiment as follows. The context discovery algorithm performs consistently well across different types of events. Depending on the type of the event, it discovers the best data sources and ranks the candidate set in the appropriate way. For a single dataset, the ranking of candidates can significantly vary depending on

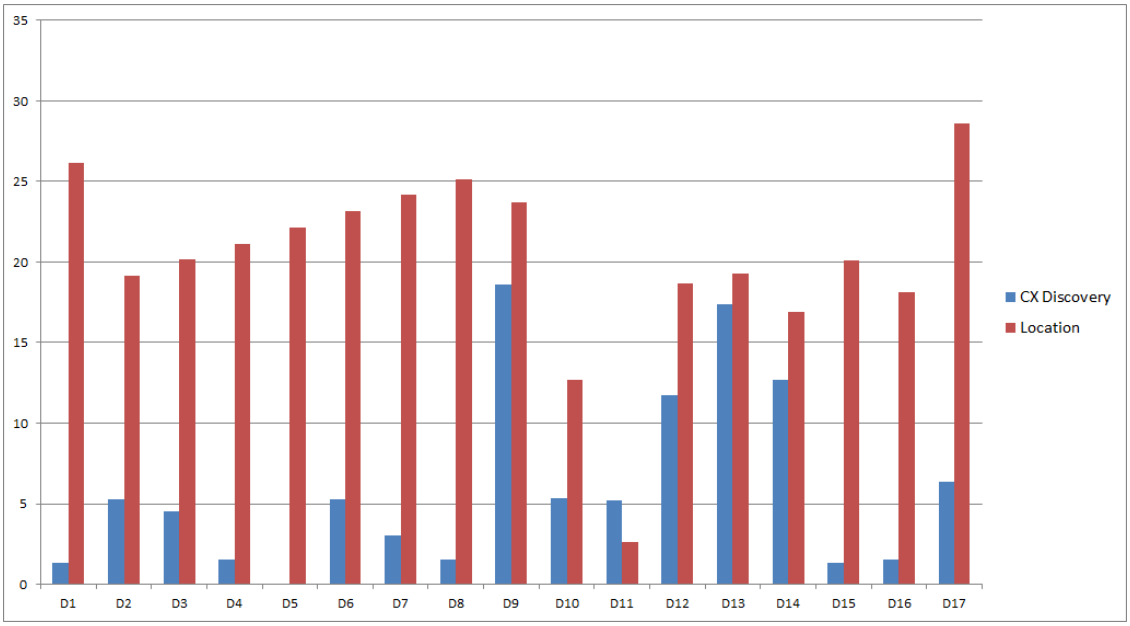

Figure 5.9: Comparing Verification Ratio (x 100) for all datasets.

who has been tagged in the photo. There are many research opportunities to propagate context across photos taken at different events and within the same event. We present such an context propagation algorithm in chapter 6.

5.2 Performance Analysis

In this section, we will look at the performance characteristics of the CueNet algorithm. More specifically, the merge function in the discovery algorithm. In order to achieve this, we will present a generative model to generate large workloads of context networks, which will be merged with each other. Using the generative model, networks of varying sizes will be created so that the time and space complexity of the algorithm can be studied.

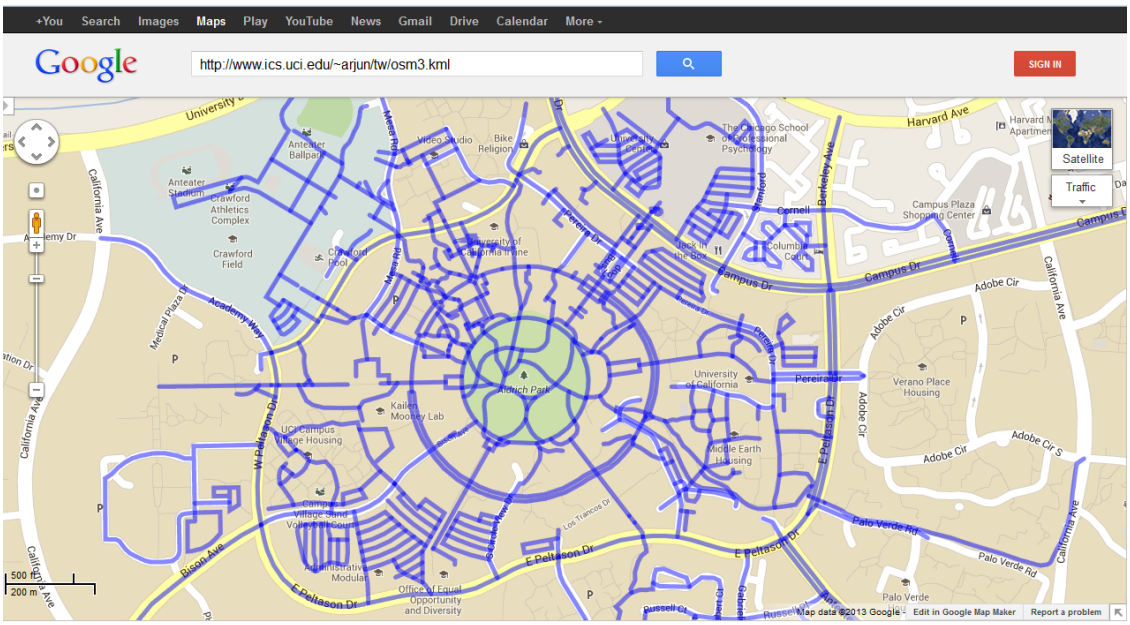

Figure 5.10: Roadmap containing places where events could occur.

5.2.1 Generative Models

The generative model requires three main components. A set of places where events occur. An ontology which determines what events occur, and how certain events are more likely to occur as sub-events of already occurring events and the event instances themselves. We will describe each of the components in turn.

Instead of designing a generative model for places, it is advisable to use one of these real world databases and select a sample of it to obtain a list of places where events can occur. A large number of place databases are available online. Some commonly used databases are shapefiles [5] for road network information about a particular area. Stanford's SNAP database contains the entire road networks of the USA [2]. APIs from websites such as Factual or Yelp.com provide access to large amounts of place information. The FSU dataset available at [4] provides detailed road networks with GPS coordinates from various parts of the country. For our work, we found the spatial database available from OpenStreetMap as the best resource. Their database provides information about road networks, as well as

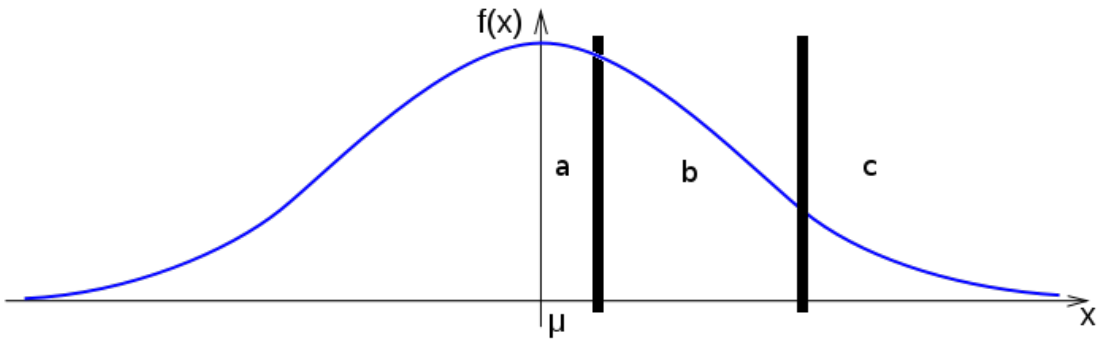

Figure 5.11: Gaussian distribution split to generate randon events in ontology.

the places themselves. At the same time, data can be downloaded from their website using the download tool or maps for entire cities or states or countries can be downloaded from [1]. The place map of UCI downloaded from their website and converted to KML format is shown in figure 5.10. The map data also contains information about different types of objects which can be found on these roads. Please look at their extensive taxonomy [3] regarding what types of objects can be found here. We select a sample of the nodes in such a network to be our place set.

The ontology generation module provides three parameters. O_d, which controls the depth of subevent trees. O_{ms} which is the maximum number of subevents an event can have. D, which is the distribution type. The distribution type decides which events are selected into a hierarchy. A distribution could be Gaussian or Poisson or Uniform distribution. Once the distribution is selected, parameters of the distribution are also provided. We will use a Gaussian distribution as an example to explain the generation of ontologies. But the ideas can be applied to any kind of distribution.

We want to model events into three categories depending on their general rate of occurrence. Some events are very commonly occurring, a majority of events occur average number of times, and events in the third category occur very rarely. Figure 5.11 shows how a gaussian random value generator can be used to bucket events into these three zones. The random values are compared with the values a, b and c. If the $|value|$ is between 0 and a, then

the event falls into the first category. If it lies between a and b, then it lies in the second category. Otherwise it is in the third category.

We consider events in the first category to be atomic events. That is, they occur independently and do not have a subevent hierarchy. An example could be a tweet event or a `photo-capture` event. We want to create two types of hierarchies. One which is short and fat, and the other which is long and tall. We choose category two to correspond to the former, and category three to contain the latter type of hierarchies. A hierarchy is generated as follows: choose an event, if it is from category 2 or 3, a hierarchy with atmost depth O_d and each super event having max O_{ms} subevents is generated. The events are randomly chosen based on their "real-world" frequency. These hierarchies are serialized into files. For events in category 3, we double the value of O_{ms}. This will allow us to get *fatter* instance hierarchies as we will see later.

In order to generate instances, we randomly select an event based on its "real-world" frequency, and load the hierarchy structure from the file, and for subevent class, generate n instances. Thus, if an event x has a subevent y in its hierarchy, then the instance x_i will have subevent instances $\{y_1, y_2...y_{|Y|}\}$ where $|Y| < n$. The number n is decided on the basis of an input metric C, where for events in category 2, $n = c/2$, and for events in category 3, $n = C$. It must be noted that the n is the maximum number of subevents a super event can contain. The actual number of subevent instances is decided by a uniform random number generator. These hierarchies are serialized into files and stored for the merge tests.

5.2.2 Procedure

In order to understand the time and space complexity of the merge function, we load multiple networks from instance hierarchy into in-memory data structures, and perform a sequence of merges. In order to simulate the workload similar to the discovery algorithm, we select

a network, and perform a sequence of merges on it. We will refer to the initially selected network as a **primary** network and the networks which are merged into it as **secondary** networks. Thus, given a primary network and a set of secondary networks, we perform the following set of experiments.

First, we fix the size of the primary networks (in terms of node count), and merge secondary networks with varying size, and measure the time taken for the merges to complete.

Second, we fix the size of the primary networks (in terms of depth), and merge secondary networks with varying depth, and measure the time taken for the merges to complete.

Third, we merge a context network with an exact replica of itself, and measure the time taken for the operation. We also change the size of the heap size, and see if there is any effect on the merge operation.

Fourth, we measure the sizes of the networks once they are in memory.

It must be noted that the event instances being used in the experiments do not have event metadata associated with the in-memory instance object. It is assumed that these are being stored separately. The networks used below are simply events with type, spatio-temporal and identification information and references to their subevents and to entities who participate in the events. Any additional information, for example the image object which constitutes the experiential aspect of an event, is stored separately.

We use the following system setup in all our experiments. We use a 8 core Intel(R) Xeon(R) E5504 processor clocking at 2.01Ghz with 8 gigabytes of main memory. We use a single thread to load networks from the file on the disk and proceeds to merge them. The memory footprints of the networks are measured by the classmexer library [34]. The `Xmx` and `Xms` JVM parameters were set to be 6000m and 256m respectively.

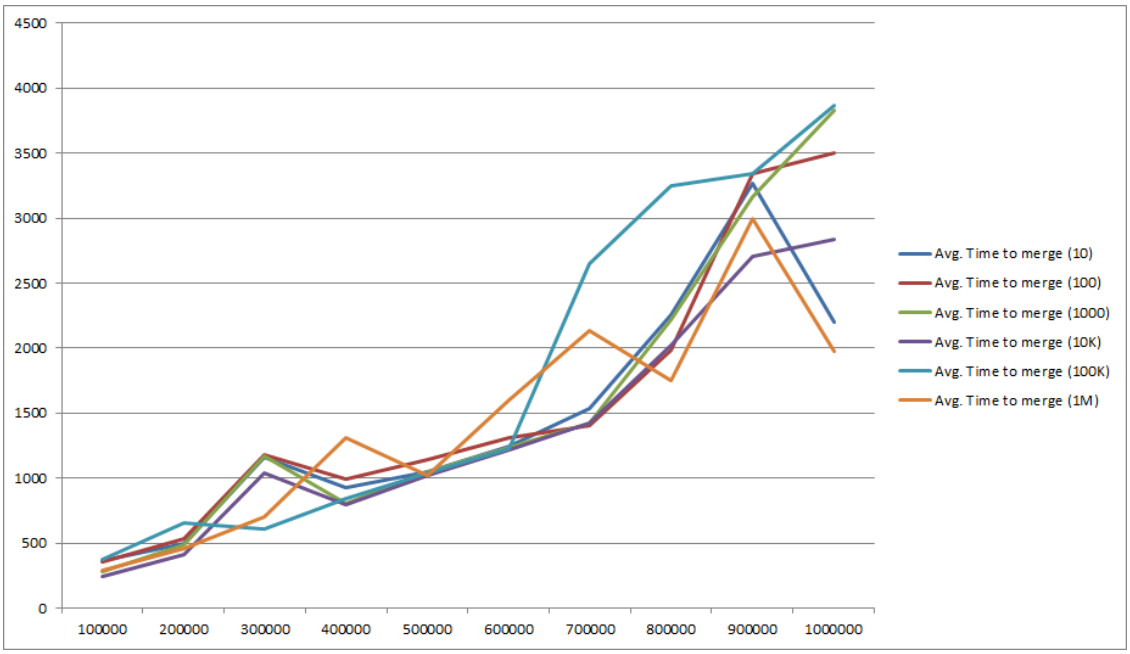

Figure 5.12: Graph showing merge times for different sizes of primary networks.

5.2.3 Results

The graph shown in figure 5.12 shows the time taken for merging secondary networks of increasing sizes to a primary network. The size of the primary network is increased from 10 nodes to a million nodes. For each such primary network, a number of secondary networks are merged. As we can see, the size of the primary network does not contribute to the time complexity of the algorithm. This is because the merge algorithm effectively searches the subevent structure to find the node which can best accommodate the secondary network. This is a walk down a single path of the tree. Thus, as long as small secondary networks are obtained from the sources, the merge operations will be less expensive. This is the reason why we tailor our discover queries to provide only very specific information. On the other hand, as the size of the secondary network increases, we find that the performance increases linearly upto a point, after it which it exponentially increases. This is due to a combination of tree traversal on the secondary network, and the JVM's garbage collector removing outdated objects to accommodate the new objects being created in the primary

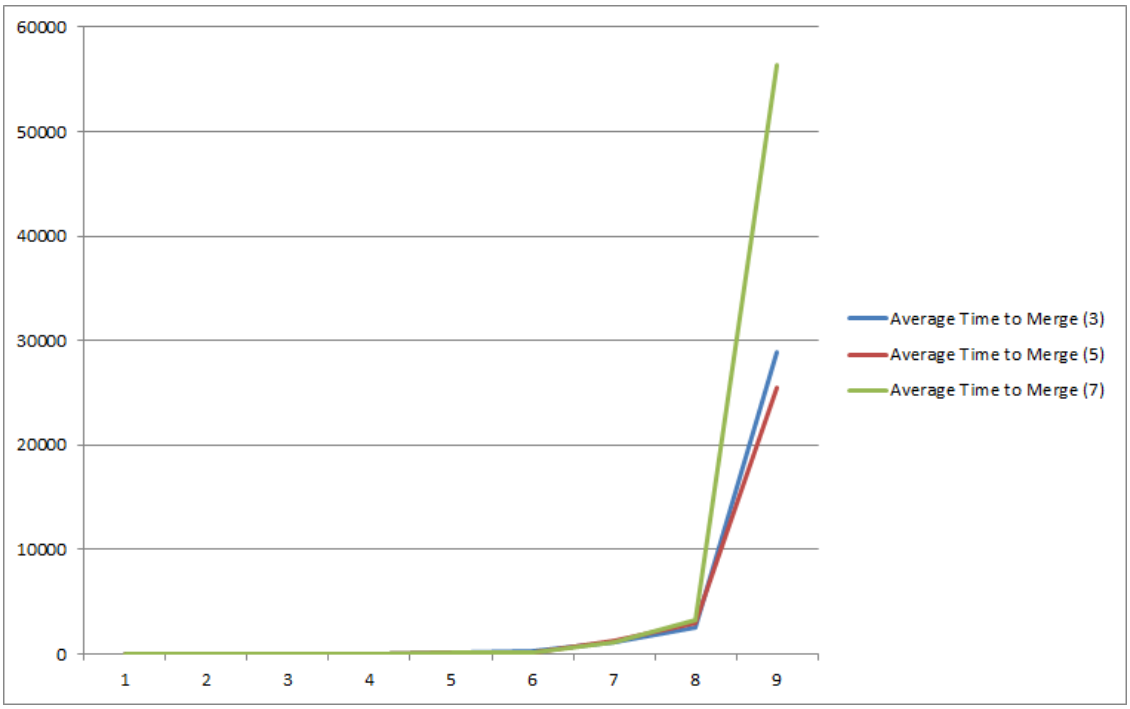

Figure 5.13: Time to merge networks with varying depth.

network. In our implementation, we create new objects in the primary network, as opposed to copying them from the secondary network.

Similarly, in figure 5.13, we test how merge time varies with increasing depth of the primary network. Here we fix the depth of the primary network to 3, 5 or 7 levels, and load a set of secondary networks from depth 1 to 9. The secondary networks are merged in sequence to the primary network. Each run of the algorithm corresponds to a single primary network (effectively, we reboot the JVM for each primary network). Here the exponential rise after the depth of the secondary network reaches 8 levels is more obvious. This again, shows the importance of constructing selective queries during the discover phase of the algorithm. The loading of all the secondary networks prior to the merge sequence allows us to study the cost of merges in an environment where memory is not available, and is shared by other citizen objects of the JVM heap.

Figures 5.14 and 5.15 show the time complexity of self merges. A self merge is merging a

network with an identical copy of itself. The perforance characteristics are very similar to the merge tests seen before.

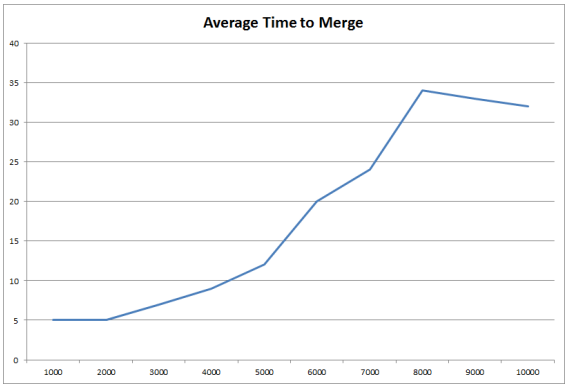

Figure 5.14: 1K-10K Nodes Self Merge

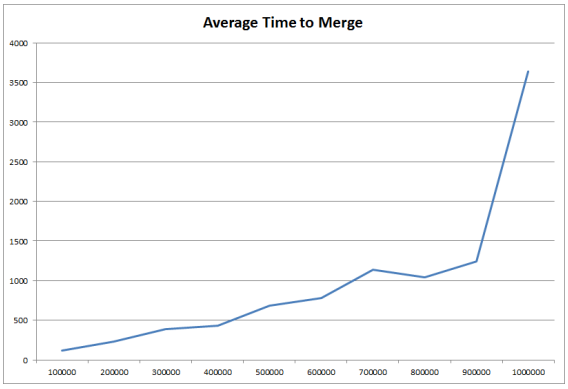

Figure 5.15: 100K-1M Nodes Self Merge

Now, we will look at how the merges behave with varying heap size (the `Xmx` runtime parameter provided to the JVM when it is initialized). In figure 5.16, we see 5 different curves, corresponding to `Xmx` values of 512m, 640m, 768m, 896m, and 1024m. We load a network twice, perform a merge (a *self-merge*), and record the time the operation takes. The experiments is repeated 5 times for each merge on different randomly created networks. We perform these operations on networks containing 1000, 2000, 3000 nodes upto 700K, 800K, 900K and 1M nodes. As we can see in the figure, we run out of heap space when the networks contain 600K and the Xmx value is 512m. Beyond a 896m, we see that the merge operations only grow linearly, and can be considered a safe limit when self-merges of 1M node networks are to be performed. For heap spaces of less than 896m, the performance time increases exponentially when the network size exceeds 600K nodes. This is, again, because of the creation of nodes in the primary networks.

Finally, we use the Classmexer library [34] to measure the memory footprint of the in-memory context network. We generate random networks of size 1000, 2000 ... 10,000 nodes, and measure the size in kilobytes. The results are shown in figure 5.17.

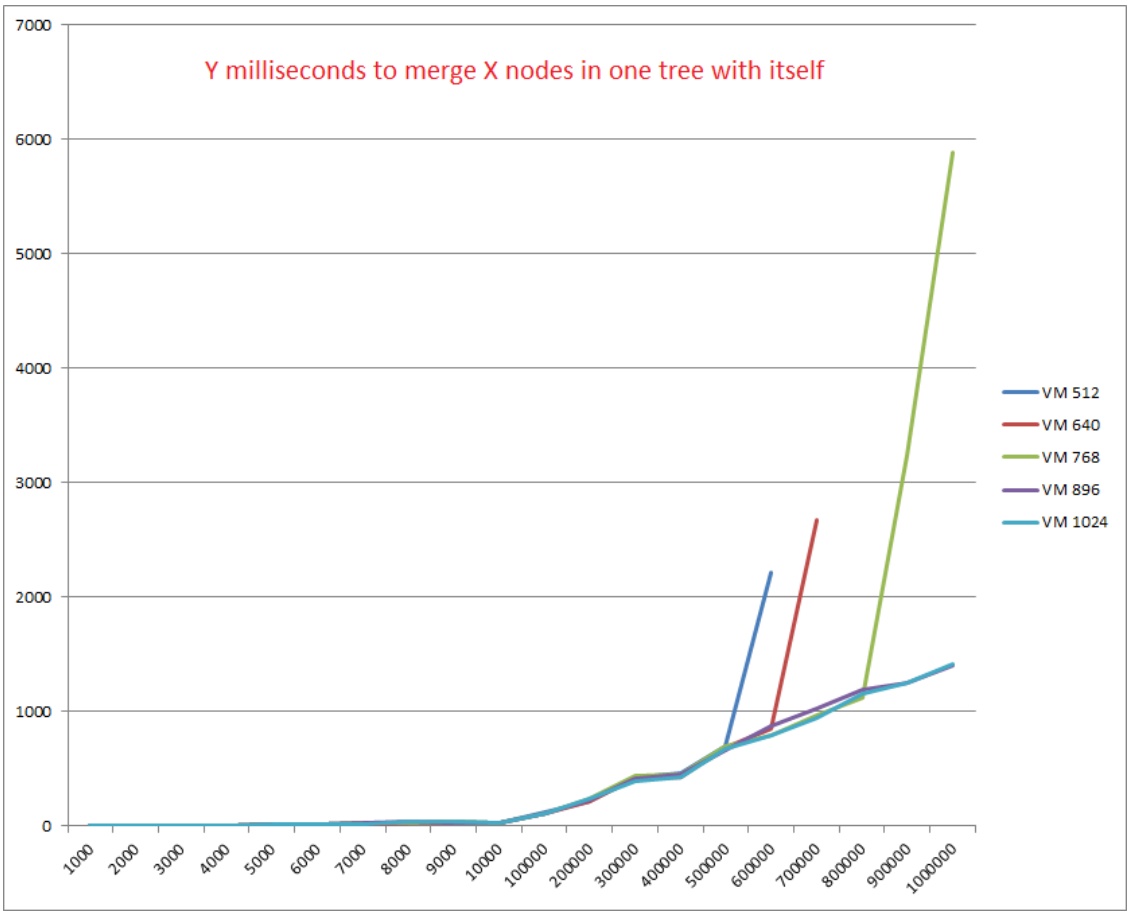

Figure 5.16: Time to merge networks with varying available heap space.

5.2.4 Conclusion

In analysing the above figures, we have seen that the in-memory merge algorithm performs very well for networks upto 500K nodes. Beyond this point, the performance exponentially decays, both as a combination of large graphs, and the JVM's garbage collector trying to clean up unwanted objects. And therefore it is best recommended to approach an disk based solution or partition the context networks across different machines.

Since performing many merges (especially with large secondary networks), it is advisable to not perform the merges when the rate of incoming networks is very high. Rather, the system can cache the networks, and extend techniques such as dynamic programming or *batch merges* to improve the overall performance. Such an algorithm, does not already exist,

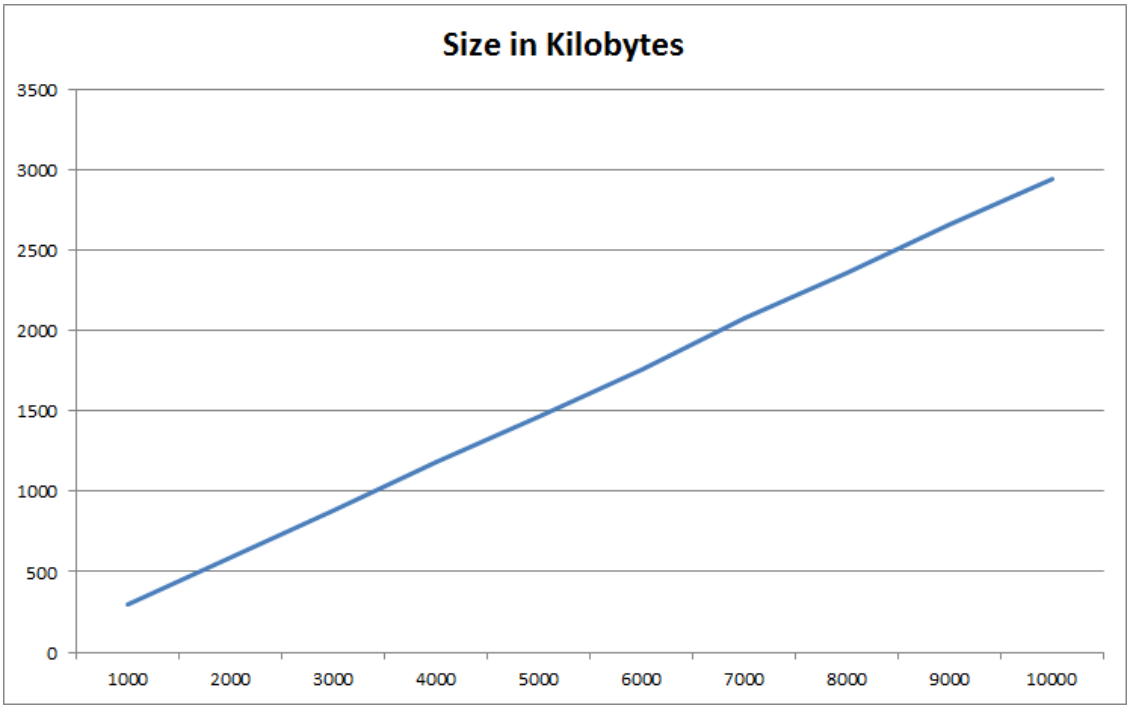

Figure 5.17: Sizes of in-memory context networks contain a few thousand nodes.

and can prove to be a very important future research problem to improve the utility of context networks in various domains with large workloads.

Chapter 6

Ranking Context Networks

So far, our treatment of context networks had made an assumption: context solely consists of events co-occurring with the photo-capture events, and these events are obtained from data sources. The second part of the assumption is not always true in practice. For example, a person's roommate might be a last minute addition to a road trip, who was neither on the email chain or the facebook event, will be ranked very low by the discovery algorithm. A professor whose students are receiving best paper award but was not part of the author list herself might be present at the award ceremony because the conference is being hosted 50 kilometers from her university. Similarly, people at a concert might run into acquaintances because they share the same musical interests. In these cases, the data sources will provide incomplete contextual information which leads the discovery algorithm to rank some candidates poorly.

In this chapter, we will attempt to address this problem of boosting the performance of the CueNet framework by introducing a technique to rank candidates based on their participation profile in previous context networks, personal interests or information. We will present our intuition through a series of examples, and introduce a technique similar to PageRank,

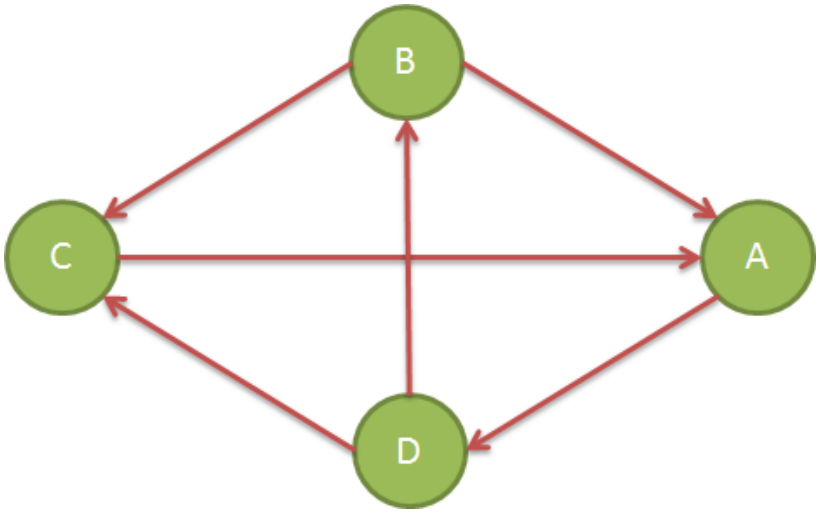

Figure 6.1: Example graph to demonstrate the original Pagerank algorithm.

which is used to rank pages on the world wide web. In our case, we will rank people in the `CandidateSet` given set of context networks. Our main differences will lie in the initialization of the score matrix and the propagation of scores *across* the different context networks.

6.1 Preliminaries

Rank propagation techniques have been used to rank nodes in directed graphs. Most commonly seen versions in today's literature are the HITS algorithm invented by Jon Kleinberg [66] and the PageRank algorithm developed by Larry Page and Sergey Brin [83]. The idea in the latter is to assign a set of initial values to a subset of nodes, and propagate a fraction of these values to their neighbors. Each iteration of the algorithm propagates their scores to the neighboring nodes until the overall ordering of the nodes, according to their scores, does not change. In practice, a few iterations (< 100) on web scale graphs is sufficient to achieve convergence.

The page rank [83] algorithm, designed to rank large nodes in web graphs is an extension of this idea. The intuition behind page rank is that of the random surfer model. The rank

Algorithm 3: Original Pagerank Algorithm

```
1 begin
2     R_0 ← S
3     loop:
4         R_{i+1} ← A R_i
5         d ← ||R_i||_1 - ||R_{i+1}||_1
6         R_{i+1} ← ||R_i||_1 + dE
7         δ ← ||R_{i+1} − R_i||_1
8     while δ > ϵ
9 end
```

propagation models the behavior of a random web surfer who follows a few links and then gets 'bored' and jumps to a random part of the web graph. The final scores of the ranking algorithm converge to the probabilities of such a surfer spending time at a particular page. Another analogy to understand PageRank would be the following: If many surfers (where number of surfers is greater than number of nodes) uniformly choose a page to start surfing from, and move to either one of the outgoing nodes or decide to jump to another page in the graph, then after some iterations, the page rank of a node is fraction of surfers left on it. Mathematically, the page rank of a given web page i at the k^{th} iteration of the rank computation algorithm, denoted as $PR(i,k)$ is defined recursively according to the equation [21]:

$$PR(i,k) = dD(i) + (1-d)\sum_{j \to i}[PR(j,k-1)/N(j)] \tag{6.1}$$

In equation 6.1, d is the probability with which the random surfer will jump to a sample from the distribution $D(.)$, and with probability (1-d) jumps uniformly at random to one of the pages linked from the current page.

Let us take a look at how rank propagation works in simple graphs. Consider the directed graph in figure. Let the initial scores of all nodes be 0.25. In the first iteration of the

algorithm, if we assume that the surfer will only follow links (i.e., d = 0), the score of node A would be computed as follows. Since A has two incoming edges from B and C, they will transfer a portion of their scores to A. Since B has two outgoing edges, it will transfer half of its score to A. C has only one outgoing edge, and therefore transfers all of its score to A. Thus, the value $PR(A)$ becomes $0.25/2 + .25 = 0.375$ at the end of the first iteration. Similarly, B obtains half the score of D; C obtains half of B and half of D; and D obtains the full score of A. For each node, we can rewrite 6.1 as follows:

$$PR(A,i) = \frac{PR(B,i-1)}{2} + PR(C,i-1) \tag{6.2}$$

$$PR(B,i) = \frac{PR(D,i-1)}{2} \tag{6.3}$$

$$PR(C,i) = \frac{PR(D,i-1)}{2} + \frac{PR(B,i-1)}{2} \tag{6.4}$$

$$PR(D,i) = PR(A,i-1) \tag{6.5}$$

If the probability of jumping to a web page from the distribution D is 0.15, and we assume that the probabilities in D are equally likely, the above equation to compute the new pagerank of node A becomes $PR(A,i) = \frac{d}{N} + (1-d)\frac{PR(B,i-1)}{2} + PR(C,i-1) = 0.35625$. Table 6.1 shows the different values of page rank for each node in our example graph per iteration until it converges. We assume that the initial ranks of each node is 0.25. The difference in ranks is computed using L1 norm of the difference of ranks in the current and previous iteration.

$$\delta = \sum_{\forall i} |PR(i) - PR(i-1)| \tag{6.6}$$

The page rank algorithm is shown in algorithm 3. R is a vector over all the nodes in the

k	A	B	C	D	δ
1	0.25000	0.25000	0.25000	0.25000	-
1	0.35625	0.14375	0.25000	0.25000	0.21250
2	0.31109	0.14375	0.20484	0.34031	0.18062
3	0.27271	0.18213	0.24323	0.30193	0.15353
4	0.32165	0.16582	0.24323	0.26930	0.09788
5	0.31472	0.15195	0.22243	0.31090	0.08319
6	0.29114	0.16963	0.23421	0.30501	0.05893
7	0.30868	0.16713	0.23922	0.28497	0.04508
8	0.31187	0.15861	0.22964	0.29987	0.03619
9	0.30011	0.16495	0.23236	0.30259	0.02352
10	0.30511	0.16610	0.23620	0.29259	0.02000
...	...	...	...	...	...
34	0.30554	0.16381	0.23343	0.29721	0.00002
35	0.30554	0.16382	0.23344	0.29721	0.00002
36	0.30554	0.16381	0.23344	0.29721	0.00001
37	0.30554	0.16381	0.23343	0.29721	0.00001
38	0.30554	0.16381	0.23344	0.29721	0.00001
39	0.30554	0.16381	0.23344	0.29721	0.00000
40	0.30554	0.16381	0.23343	0.29721	0.00000

Table 6.1: Page rank scores for nodes in the example graph per iteration.

graph. A is an adjacency matrix where each cell measures is the reciprocal of outgoing edge count. If u and v are two nodes, and there are N_u outgoing edges from u and one of them ends in v, then $A(u, v) = 1/N_u$. The expression $||R||_1$ is the L1 or Manhattan norm of a vector R, which is computed as $\sum_{\forall i} |R_i|$. The vector S is the initial score assigned to the nodes. On a large scale, pagerank scores are computed using the MapReduce framework [37]. Each iteration of the computation corresponds to one map reduce job. Multiple such iterations are performed to obtain the final results.

6.2 Intuition

To rank objects in context networks, we modify the random surfer model slightly to model the movement of objects between different networks. Consider a set of context networks,

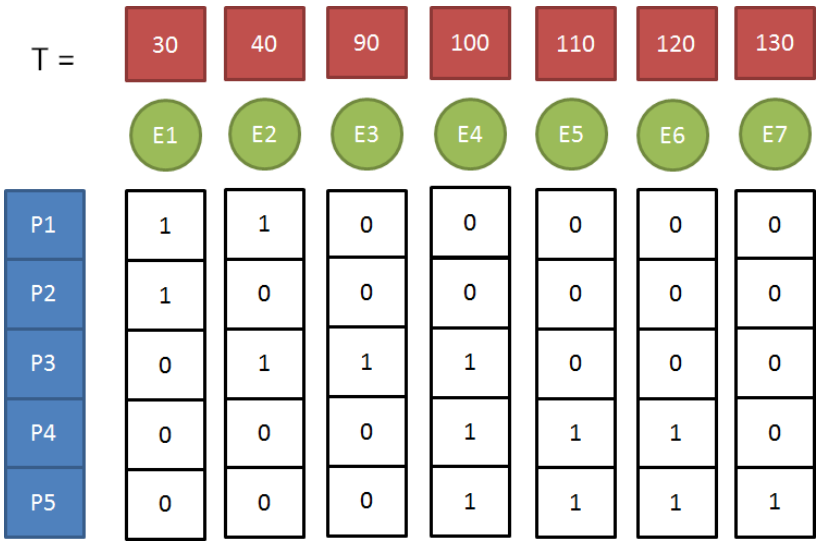

Figure 6.2: Propagating values through temporal relations.

each of which correspond to a `photo-capture-event`. Each network consists of a set of participating objects. An object in a particular network can move to a neighboring context network depending on the differences between properties of the two networks. If the two photos were captured right next to each other, and contain very similar types of events, and object instances, then there is a high probability that this person who was tagged in one of them, will be present in the other. Time is one such axis of propagation. Other properties such as event class, object co-occurrence, place (type of place or the distance) of event occurrence or subevent patterns in context network can provide different axes to propagate tags between photos.

Consider the context networks shown in figure 6.2. Events `E1 - E7` contain one or more persons from the candidate set `P1 - P5`. The binary vector below event describes whether an object was present in the event or not. `P1`, for example, participates in `E1` and `E2`. `P5` participates in events `E4` through `E7`. The top row shows the time of occurrence for all events. We shall assume that other properties of an event (class, location and subevent structure) are same, and therefore do not effect the propagation. Since `E6` contains `P4`, it can be said that there is a higer chance of `P4` participating in `E7` than in `E1` for two reasons. First, the

object distribution in `E1` is very different from `E6`. `P4` has never participated in any event with the `P1` and `P2`. Second, because the temporal difference is much higher, very little can be estimated about the presence of a person in an event. If we quantify the object similarity using a set similarity measure, such as the jaccard index (represented by $d_s(e_i, e_j)$, and the effect of time using a formula $1 - \frac{\Delta T}{T_{max}}$, where ΔT is the difference between occurrence times of two events in question, represented as $d_t(e_i, e_j)$, then we can say that the probability that a person will appear in some Ex at the i^{th} iteration, $R(p, Ex, i)$, who is known to have appeared in some other set of events $\mathcal{E}$, where $Ex \notin \mathcal{E}$ is:

$$R(p, Ex, i) \leftarrow \sum_{e \in \mathcal{E}} [d_t(Ex, e) d_s(Ex, e) R(p, e, i-1)] \tag{6.7}$$

Similarly, look at figure 6.3 where the context networks now differ in type and location of occurrence as well. We show the event class information (indicated by 'C') and location information with 'L' of the event above the nodes. Their timestamps are as before, but our original assumption about spatial and ontological consistency no longer holds. Here, can we need to exploit some ontological properties about event similarities to propagate scores. If we know that `L1` is a `conference`, `L2` is a `music-event` and `L3` is a party, then can we say that `L2` events are more likely to contain the untagged person than the `conference` events. Likewise, spatial relationships propagate higher scores to individuals between events which occur near each other, than those which occur further away.

In this chapter, we will modify the page rank technique (algorithm 3) to rank entities on a given set of context networks, to determine who might be participating in a new events. We will initialize scores of the nodes on the basis of what is known about the new event: the type of the instantiating class, the participating objects, and propagate the scores based on a set of propagation functions which we will introduce in the next section. Since the basic

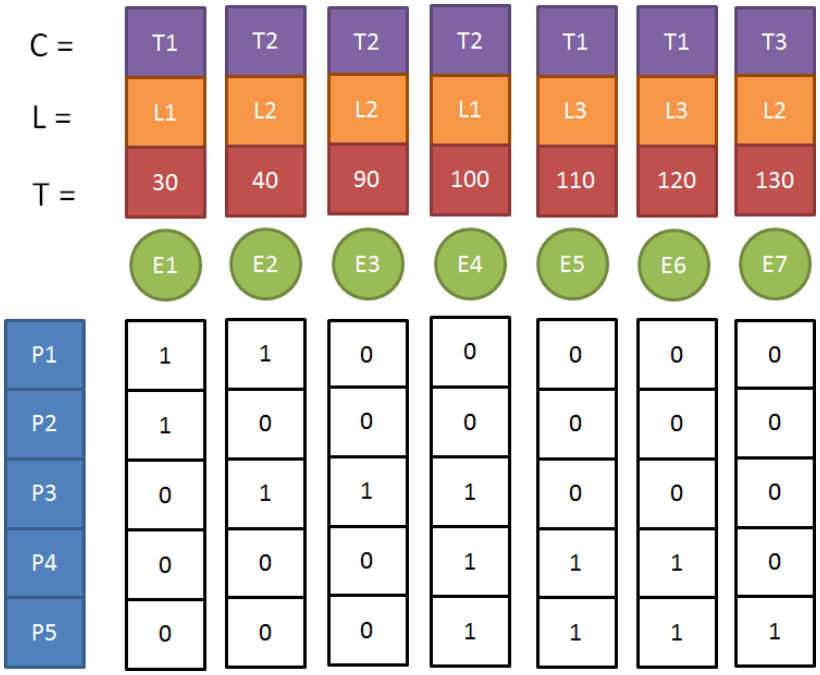

Figure 6.3: Other properties of events (space, type, subevent hierarchy patterns or object co-occurrence could be used to propagate scores.

propagation algorithm is an expensive one, we will look at some ways to reduce the size of the context networks, so the algorithm converges quickly, and still provides useful results.

6.3 Rank Propagation in Context Networks

In order to rank candidates for a new given photo, we go by the observation that people participate in similar kinds of types, and spend time with similar groups of people. In other words, the distribution of people in events is not random. By ranking events, we are ordering them by some metric of similarity with respect to the context network containing the new photo. Events encapsulate heterogeneous information. Thus, their similarities must be computed across the different kinds of entities they encapsulate. For example, two event instances of the same class which have no common participating object are less similar to each other than two others which are of the same type and contain same participating objects.

For our application of face tagging, computing similar similarities across time, space, type and object information of events. In order to formalize our rank propagation mechanism, we introduce the concept of a propagation function. A propagation function will transfer score from one object in a event to itself in another event through specific property of the event. For example, if we can design a propagation algorithm around two `temporal` and `spatial` properties, scores will be transferred based only on the temporal and spatial attribute values of the events respectively. We say that the propagation function contributes a *transfer ratio* across an axis. We follow the steps as shwon in algorithm 3 with the following differences: **First**, the vector S is setup based on the distribution of persons present and the attributes of the events in the new context networks. **Second**, the matrix A is computed as the product of all the transfer ratios between two events. For two events, u and w, and propagation functions P_t and P_s (temporal and spatial propagations), the propagation vector is $V = [P_t, P_s]$, and the score transfer vector from u to w, $V(u, w) = [P_t(u, w), P_s(u, w)]$ and the value $A(u, w) = \mathcal{C} \times P_t(u, w) \times P_s(u, w)$, where $\mathcal{C}$ is a constant scaling factor.

6.3.1 Propagation Functions

In this section we introduce the propagation functions used in our work to rank events for the face tagging application.

- **Temporal Propagation**: For events u and w, the temporal propagation function $P_t(u, w)$ propagates a fraction of the score of node u to w proportional to the difference between `u.duration` and `v.duration`. The transfer function can be modeled using uniform gaussian or zipf distributions. The advantage with such distributions is to model high transfer of scores who occur nearby in the temporal dimension, and ignoring those which happened too far away. The distribution parameters will be constant and set by the designer.

- **Spatial Propagation**: Similar to the temporal propagation function, this takes into consider the spatial attributes of an event while transferring scores. The distance between two coordinates can be computed through one of three ways. First, a table based scheme which allots scores based on the address information of the event. Here, high scores are given if the two events occur in the same city in comparison to those within the same country. Second, using euclidean distance between GPS coordinates of the two events. Third, compute the actual 'as the crow flies' distance between the two locations. We use the formula based on the spherical law of cosines: where the distance between two pair of coordinates d(lat1, lon1, lon2), is given by (where $\mathcal{R}$ is the radius of the Earth (6378 kms)):

$$\mathcal{R} \arccos(\sin(lat1)\sin(lat2) + \cos(lat1)\cos(lat2)\cos(lon2 - lon1)) \tag{6.8}$$

- **Type Propagation**: The class of the event instance is taken into consideration to compare two events. We use the concept of ontological distance to compute the type propagation score transfer. We extend the concept of `Is-A` distance given in [90]. For given a subsumption hierarchy, a is an ancestor of b iff there is a path between a and b. The set of concepts having a as ancestor is denoted by $desc(a)$, while its ancestors are denoted as $ansc(a)$. The set of exclusive ancestors (if a node is an ancestor of exactly one of the two nodes) of a and b is denoted by $exAnsc(a, b)$. The subsumption ontological distance is defined as:

  $d_{IS}(a, b) = |desc(exAnsc(a, b) \cup desc(a) \cup desc(b) - desc(a) \cap desc(b)|$.

  Events also exhibit subevent relationships. Two events could have very common subevents, but might have significant subsumption distance. In this case, we want to reduce this distance. We introduce the subevent distance as the number of common subevents as the number of common descendant in the subevent hierarchy of the

ontology. For a node a, $subdesc(a)$ is the set of all a's subevents and their transitive descendants.

$d_{SE}(a, b) = |subdesc(a) \cap subdesc(b)|$

Our ontology distance is the sum of subsumption distance and the subevent distance. This number can be normalized with the number of perdurant classes present in the ontology.

$d(a, b) = D_{IS}(a, b) + D_{SE}(a, b)$

- **Object Propagation**: If two events have the same set of participating objects, then the score transfered through object propagation is very high. If they have no common objects, then this transfer would be 0. We use Jacquard index to compute this ratio. For two sets of objects A and B, the jacquard index is computed as:

$$J(A, B) = \frac{A \cap B}{A \cup B}$$

- **Structural Propagation**: The last propagation function we use is the propagation from an event to its instance subevents. Here we use a function which is similar to the one used in pagerank while computing the transfer from a graph node through an outgoing edge. If event B is a subevent of A (we can use the web graph as an analogy where the subevent edge is an outgoing edge from A to B), and A has n subevents, then the propagation ratio from A to B is $1/n$.

6.4 Propagation Algorithm

The final transfer ratio is the product of all the individual transfer ratios. This allows us to transfer smaller scores if the two events differ largely in any one attribute. This is particularly useful when many events occur spatiotemporally close to each other, but are instances of

very different event classes. On the other hand, a sum based combination of the individual scores would propagate larger number of individuals because there is atleast one good axis.

We will compute the scores of objects using the following steps. Construct a map M, such that $M(V, o) = 1$, if the object o was participating in V. Let $\mathcal{P} = \{P_1, P_2 \ ... \ P_n\}$ be the set of all propagation functions. The score transfer, $\Delta(x, y, o)$ for the object o from a node x to y is given by the equation: $\prod_{1 \leq i \leq n} P_i(x, y) M(x, o, i-1)$. The score at the end of iteration i for o is given as the sum of all $\Delta(x, \bar{y}, o)$, where $\bar{y} \in \bar{Y}$. $\bar{Y}$ is the set of all nodes which are within satisfy the filter function F, such that $F(x, \bar{y}) = true$. This function is used to limit the number of propagations between events. Unlike the web graph where each nodes points to a very tiny fraction of web pages, propagation between events is $O(|V|^2)$ where $|V|$ is the number of events per iteration. If the score of an object is 1, then we do not update its score (the object is known to be present in the event). We call such entries in the map as immutable entries. For every mutable entry, the final score of o at the end of iteration i computed as:

$$M(V, o, i) = \frac{\sum_{y \in \bar{Y}} \prod_{1 \leq i \leq n} P_i(x, y) M(x, o, i-1)}{|\bar{Y}|} \tag{6.9}$$

We compute $\delta(V)$ as the difference between object scores of the event V between two iterations, using L1 norm (similar to the procedure in algorithm 3). The biggest difference in propagation is denoted by $\delta = max(\delta(V_1), \delta(V_2), ...)$. The iterative score computation is continued until $\delta < \epsilon$, a very small constant. At this point, the scores in each event are checked, and the objects with highest scores in each event are considered to be the next best guess. These objects can be verified using a face verification algorithm or manually. If they indeed exist, the score for this object must be updated to 1 in the M map and its entry made immutable.

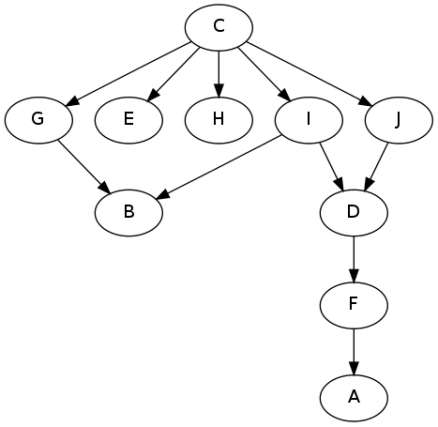

Figure 6.4: Example of a subsumption hierarchy created as part of the ontology generation process.

	A	B	C	D	E	F	G	H	I	J
A	0	6	9	2	7	1	7	7	5	5
B	6	0	9	6	7	6	5	7	5	7
C	9	9	0	7	9	8	8	9	5	6
D	2	6	7	0	7	1	7	7	3	3
E	7	7	9	7	0	7	3	2	6	5
F	1	6	8	1	7	0	7	7	4	4
G	7	5	8	7	3	7	0	3	5	6
H	7	7	9	7	2	7	3	0	6	5
I	5	5	5	3	6	4	5	6	0	3
J	5	7	6	3	5	4	6	5	3	0

Figure 6.5: The `is-A` distance matrix for classes in figure 6.4.

6.5 Experiments

We investigate the convergence behavior of our propagation algorithm. To simulate this experiment, we use the generative model described in section 5.2.1. We generate an ontology using the following steps: A random number of entities are chosen, and a subsumption hierarchy is created (similar to the one shown in 6.4). A fraction of these events are assigned a subevent hierarchy, which uses the step described in section 5.2.1. For this experiment we restrict the depth of the subevent hierarchy to be 2, and each event class to contain atmost 3 subevents. Events are instantiated randomly. If the ontology specifies a subevent structure, we instantiate subevents too. For each subevent class we use a uniform random generator to decide if it should be instantiated or not. We generate 1000 events for this experiment. It must be noted that this is a fairly large number as we are propagating scores for each event to all the others one due to temporal and spatial propagations. Thus, for 1000 events, we expect a million propagations per iteration. We use a linear distribution for spatial and temporal propagation functions. For the temporal function, this means that the propagation decreases linearly until 5% of the total timespan of the events, beyond which point no score is transferred. Similarly spatial propagation decreases 10% of the maximum distance between

two event instances in the dataset (these percentages were arbitrarily chosen).

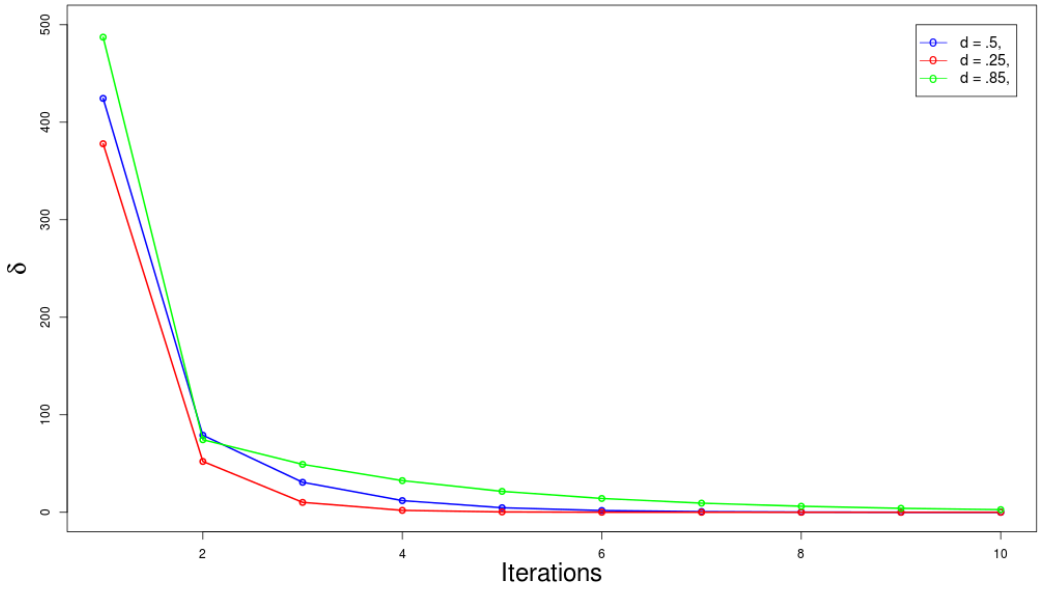

Figure 6.6: Rate of convergence of the propagation algorithm for varying values of d.

Figure 6.6 plots the L1 norm in the rank vector for different values of the decay factor per iteration. For a higher value of the decay factor, d, we see a faster convergence rate. It must be noted that for this convergence is achieved within a few iterations. We also plot the rate of convergence for datasets with different number of instances. In figure 6.7, we plot the decreasing value of δ on a log scale for two datasets, one containing 1000 nodes, and the other with 1500 nodes. The plot shows the increasing number of iterations for larger datasets to converge to the same value.

Now, we shall apply this technique to ranking real-world photos. We consider a dataset containing 120 photos taken during a `social-party` event. All photos were annotated by the owner, and contain required EXIF tags. In order to test the algorithm, we construct context networks for all photos (using the ground truth information supplied by the user). We select some photos randomly, and remove all object information from them. The propagation algorithm is used to order the collection of objects for these photos. We use precision and

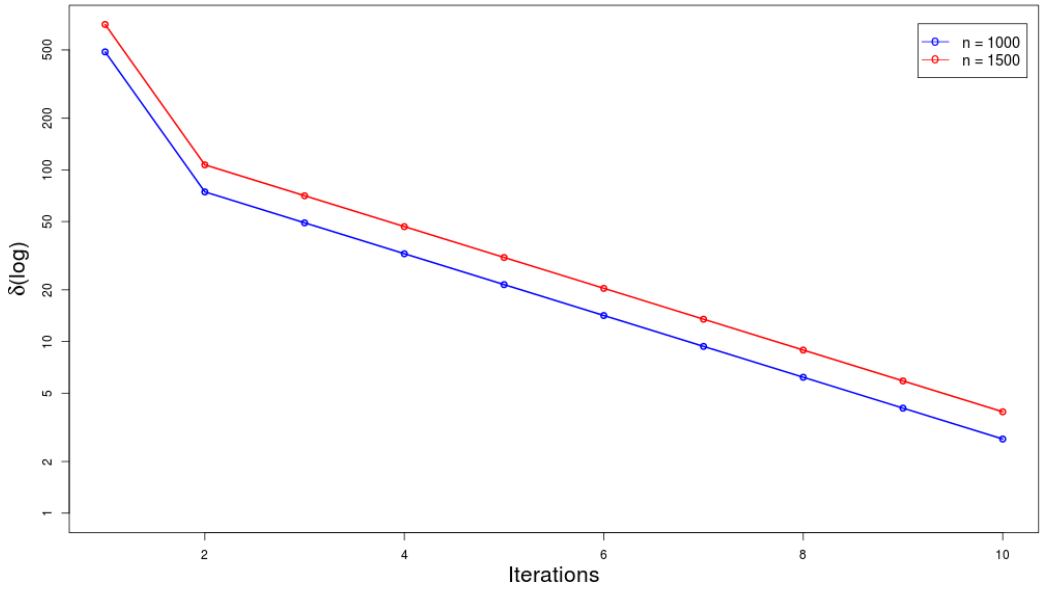

Figure 6.7: Rate of convergence with increasing number of instances, for $d = 0.85$.

recall to analyze the performance of the propagation algorithm. For this experiment, we only use temporal propagation with a linear transfer function which decreases linearly until 2% of the timespan. In figures 6.8 - 6.10, we see different precision recall graphs drawn for different photos. The performance for ranking people in a photo can vary because of many reasons. In our experiments, we see that many personal photos contain the same set of people in sequence. This is because the photographer was capturing a specific event where these people were playing a prominent role or the photographer was not satisfied with the aesthetic quality of photo, and tried multiple captures before settling on a photo. In such cases, we see all the objects in the ground truth rank very highly from the propagation algorithm. Figure 6.8 shows the precision recall graph for two such photos. On the other extreme, the algorithm fares poorly when new people are introduced in the photo. Here, they are ranked very low as the expected people are those who were captured in nearby photos. This can be seen in the figure 6.10. A set of photos also displayed intermediate results (figure 6.9), where some of the entities ranked highly but the rest were ranked very low. The distribution of people in these photos looked such that some people were present in nearby photos, but others were

not. This is due to the nature of the event. Since our linear propagation function does not model the behavior of photographers in such `party` events, the algorithm sees seemingly random behavior. What is actually happening is that the photographer moves his vantage between groups of people in the party. If the propagation model can be extended to capture this event specific behavior, we can possibly boost the performance of such an algorithm.

6.6 Conclusion

In this chapter, we outlined a technique similar to [83] to order the entities when only partial contextual information for a photo is available. There are two main advantages to such a technique. First, it can be used to tag photos which are captured after an event's duration has passed. Since the people are still participating in event, after its scheduled duration has passed, the photos taken during this interval would have *missed* the contextual information available from different sources. In such cases, the tagged people from the previous photos can be propagated to these photos. Second, during long term events like conferences people will participate in different types of events such as trips to landmarks, dinner with colleagues or visit friends in the city where the conference is occurring.

Our ontological distance measures take care of some of these cases, but there are many opportunities to explore and determine good ways of model people's behavior to propagate their scores across the different events. One way is to explicitly model such individual behavior using ontologies or similar tools. This will require many experts to work hands-on to tailor such models. On the other hand, appropriate data driven machine learning tools could be used to learn the behavior from a training set of photos, and apply them to propagate scores in a testing set. This technique requires a large amount of user data to train efficient models. But requires fewer experts to curate the models. Alternatively, a hybrid method can be devised, which utilizes the strength of both techniques, and achieves

reasonable results.

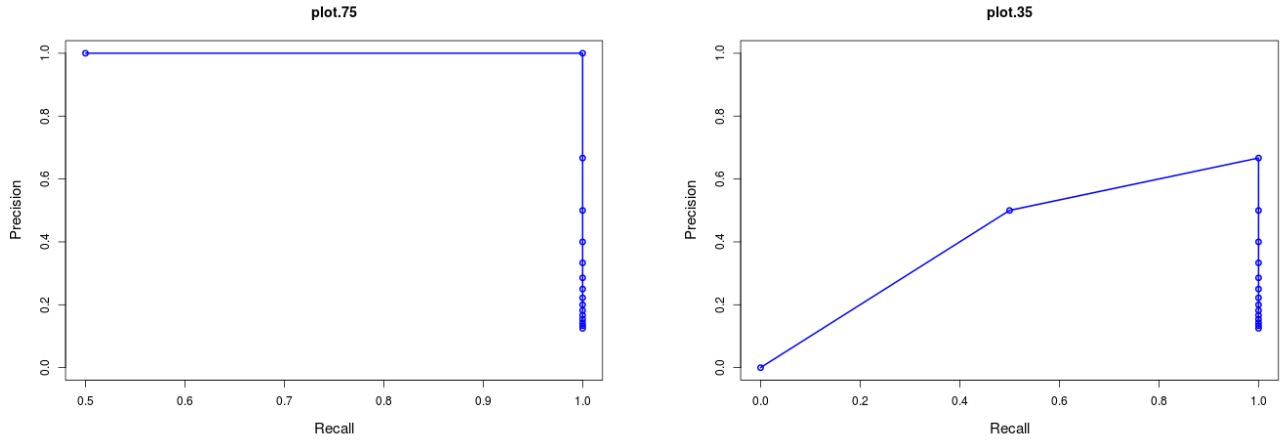

Figure 6.8: Photos where people were ranked very high.

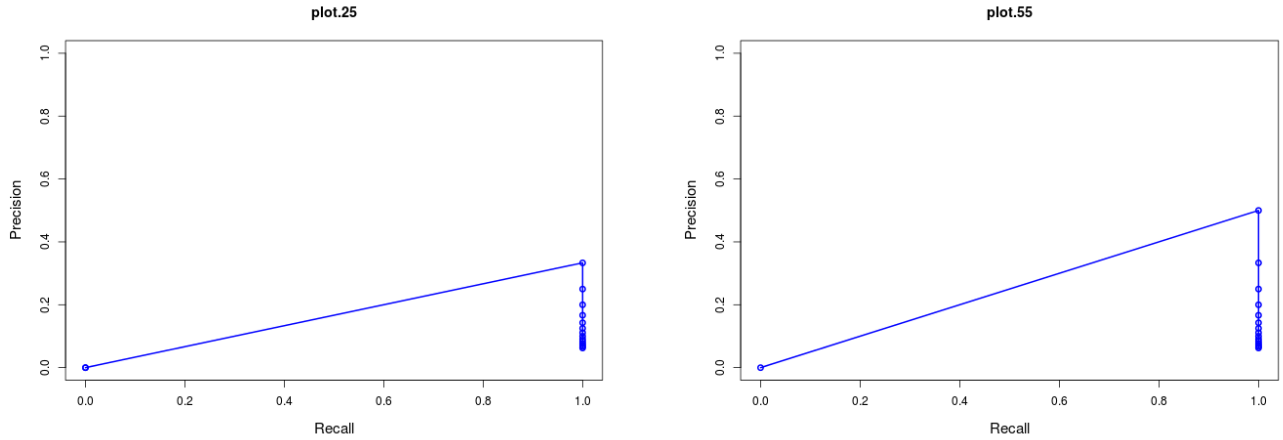

Figure 6.9: Photos where 'random' distribution of people were seen.

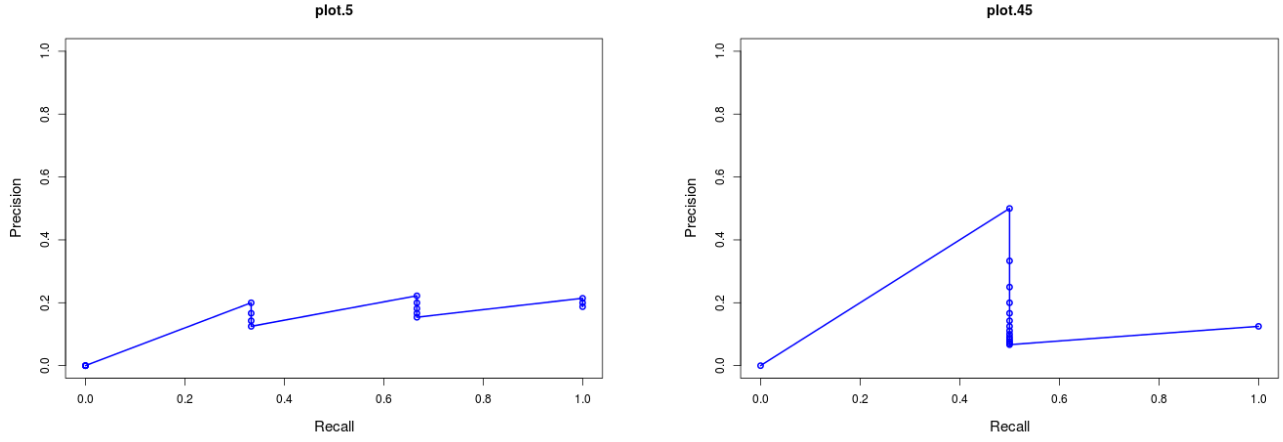

Figure 6.10: Photos where people were ranked poorly because they were not present in nearby photos.

Chapter 7

Conclusion

In this dissertation, we presented a innovative context based technique to prune the search space for the entity identification problem in multimedia objects, specifically personal photos. We introduce a relation centric perspective of contextual information to architect a novel context discovery framework CueNet, and designed its tail recursive algorithms to progressively query various heterogeneous data sources and merge context relevant to a given photo. We empirically analyzed the ability of context discovery to remove large numbers of irrelevant candidates and more importantly, verified the claims of dynamic discovery of the presented algorithms, by testing it over photos taken at different kinds of events, and comparing it with a baseline static linking algorithm.

In the cases where only partial context is available, we saw how propagation techniques can help us order persons by propagating context from nearby photos and other events. We modified the PageRank algorithm to propagate scores across events to rank and order participating objects by modeling event properties as propagation axes. The convergence of our algorithm was studied upon simulated data and the correctness was verified on photos taken during many events with different kinds of people.

7.1 Future Directions

The techniques described in this dissertation present a number of future opportunities. Here we list some of the open problems faced by context discovery techniques in general and some extensions to the algorithms described in the previous chapters.

1. The notion of relevance in context is currently specific to our application. Creating a generic notion, backed by mathematical infrastructure is an open problem. Since we are relating propagation techniques to context discovery, it might be possible to extend and relate some of the theoretical literature developed for random walks in graphs to events and time based context discovery. Any formulation of bounds in the sizes or growth rates of context network would result in very important implications of future algorithms.

2. All the data structures and algorithms described in the previous chapters cater to tagging few number of photos at a time. Performing the same work on a very large scale (billions of photos), is an open problem. It is also not clear if existing large scale computational infrastructures such as MapReduce or Pregel are the right computational foundations to construct large scale context discovery techniques.

3. In our work, we assume the correctness of context obtained from different sources. In the real world, this might not be so. Some context obtained from unmoderated data sources might be incorrect. For example, people checking in at the wrong place in a hurry or RSVP'ing to an event and never showing up. In order to deal with such cases, backtracking to a previously constructed accurate context network might become an important step in the discovery process. What data structures allow such an operation (in our case, we have no way of saying which nodes were merged in which previous stage) correctly, efficiently and in a scalable manner is not known.

4. The real world is very complicated. There are many entities which can influence each other, and in many different ways. Specifying real world knowledge has been one of the hardest problem faced by computer scientists for decades. OWL ontologies although meritorious in their own right, were designed to solve the data integration problem. It is cumbersome to describe many aspects of the real world knowledge using OWL, some of which is contradicting and contains exceptions. For example consider the piece of knowledge: bacteria *Bacillus Subtilis* is a catalyst to rock weathering. So large quantities of this bacteria can lead to severe rock weathering, which is usually accompanied by carbon dioxide consumption. By reducing carbon dioxide in the atmosphere can lead to reduction in overall temperature. The bacteria can only exist in parts of the world with favorable weather (an assertion on atmospheric temperature value). Notice the cycle of causation here. It is non trivial to model and reason with such knowledge which is very common in many domains and is crucial for this technology to be improved or engineered before advancing CueNet to such domains.

5. With the number of data sources and sensors networks increasing in the world, context discovery has a very exciting and promising future. More importantly, with the increase in sources, it will become extremely important to be selective about what information to consider context and what not. A technique to compute the importance of a data source (source selection techniques) for context discovery algorithms will be very useful.

7.2 Applications

Besides identifying entities in multimedia objects, CueNet can be utilized in a variety of other applications. We identify some of the brave new world applications, and describe them below:

1. **Contextual Narratives**: Multimodal stories are rich in context. Almost all social networks and public news feeds are overloaded with noise. There is so much content, and so little context, that the situation demands active attention from users to sift through the data to put information into context. With context discovery, models about real world events and heterogeneous information about entities from various sources can be combined and summarized to produce coarse or fine grained narratives about important public or personal events. Of course, there is sometimes no better substitute than adding a narrator in the loop to improve the final quality of such stories. In such a case, context discovery will have to be modified to work in synergy with humans for better results.

2. **Medical Diagnosis**: Studying the symptoms of a patient and coming up with a medical diagnosis is a very hard problem. Given the increasing number of medical sensors and personal health applications, it is becoming increasingly easy to construct a 'full body scan kit right at home'. With large amounts of computational power becoming available, and using the large number of tests available from their literature, a version of CueNet can be tailor made to diagnose people who are showing signs of weakening health. Instead of verifying faces, this system would be conducting specific medical tests (such as blood sugar sampling, WBC count or temperature check) in conjunction with analyzing with what the person ate, where he traveled or with whom he spent time (and if any of these people had shown such symptoms). Progressively, the system will discover which parts of the body are infected, and proceed to raise appropriate warning, along with presenting detailed reports on the infection's origin, current prognosis (expected symptoms over the next few days) and preventive measures for family members.

3. **Situation-Based Advertising**: Advertising on the web is largely driven by content on the web or textual results from search engines. Thus, many people see ads which are

irrelevant at that time. But with context discovery, context networks can be matched for specific patterns which can lead to more precise "needed-at-that-time" pattern. For example, suppose a person comes to know that a friend met with a skateboarding accident and is at the nearby hospital. The system can check this friend's wishlist on Amazon and find that it contains the new album by some of his favorite bands, which is available on a shop which is halfway between the friend and the person. The system can also check if any of these shops have signed copies of the records making the gift more meaningful.

7.3 Known Issues

In this section, we list some of the issues encountered in designing and building CueNet. Some of these are active areas of research, and whereas others are specific to our framework, and can be considered potential areas of research. Our experience with CueNet indicates that the following issues should be approached in a holistic manner, i.e., in conjunction with each other. Approaching these problems in the context of each other reduces the individual complexity of each sub-problem by possibly increasing the complexity of the entire framework, but making the problem more tractable.

Noise in Social Media: The problem of noise filtering in web data is a prominent one, and is being addressed by various communities in different ways. These range from entity matching and record linkage problems [45] to correcting missing data in information networks [93]. These problems get trickier because of the different variations in representing tiny details such as representation names of people, addresses of places, and time. In fact, there is a whole school of anthroponomastics [97] dedicated to studying variations in human names. Ideas in this field indicate that these differences arise due to cultural, historical and environmental issues[8]. Such issues cannot be trivially addressed. Consider, for example, the Hebrew name

below:

It can be translated to English as “Waisenberg” or “Vaisenberg”. So are Mr. X Waisenberg who authored a paper in a Security conference and Dr. X Vaisenberg who works at Google, the same person? They might be if X has semitic origins!

Face tags in social media sources like Facebook can also be very noisy. This strictly prohibits directly using this data to train verification/recognition models. Also, the quality of photos are poor, resulting in weaker features, which would have otherwise allowed better matching.

An exhaustive scientific characterization of noise in social media is beyond the scope of this dissertation, and is being investigated step by step in social media research communities.

CPU Efficiency: The query engine in CueNet is responsible for extracting data from different sources. If a very large number of photos are being tagged, our scheme of query generation and merging will prove inefficient. Processing many photos from different people provides a very rich opportunity to develop interesting heuristics using event semantics for the multi-query optimization problem. Also, partitioning the discovery algorithm such that the computations can occur in a distributed manner is a complex problem. Such steps will be required if the application workload is of the scales of Facebook or processing photos in real time at the scales of Instagram.

Face Verification: Even though face recognition has been studied in research for the last two decades, face verification, and its specific application to faces in the wild has been a relatively recent venture. Although the accuracy of these systems is commendable, the problems of occlusion, image quality, face alignment and differing lighting conditions exist.

These hard problems need to be solved before "perfect" or "near-perfect" verification can be established.

Execution Patterns: When is a good time to execute the algorithm? When a user takes a photo? Or before she uploads it to her favorite photo sharing site? For the current evaluation, contextual sources are assumed to be immutable. This is not true in the real world. Contextual sources are constantly being appended with new information, and old information is being updated. These updates may be vital in tagging a certain photo. So the question of when to execute the algorithm, or how and when to query the sources is an open question. If a large number of photos are to be tagged, and a busy source like Facebook is being used for context, the CueNet query engine must take into account various freshness metrics and crawling policies of the sources.

Open Datasets: The unavailability of a large public data set over which different techniques can be evaluated against each other is an open problem. As seen in our experiments, personal information is vital to contextual approaches, and this data is largely personal, and therefore cannot be shared openly. Optimal anonymization techniques need be to invented such that the privacy of the experiment participants are maintained, and at the same time the data is meaningful to be applied in contextual approaches to problems. This need to be solved so that new context discovery techniques can be evaluated independently and against each other, over a common platform.

In short, we have introduced the challenging, yet fascinating, problem of context discovery. Which becomes harder when considering scaling issues. Which becomes even harder when considering domains which are more detailed and complex than face tagging in photos. But a problem which is inevitable, and must be tackled to improve many aspects of human society and push it to the next frontiers. To put it in words which every science fiction geek has heard before:

To boldly go where no man has ever gone before

Star Trek (1966 -)

Bibliography

[1] http://downloads.cloudmade.com.

[2] http://snap.stanford.edu/data.

[3] http://wiki.openstreetmap.org/wiki/tags.

[4] http://www.cs.fsu.edu/ lifeifei/spatialdataset.htm.

[5] http://www.esri.com/data/download/census2000-tigerline.

[6] T. Ahonen, A. Hadid, and M. Pietikainen. Face description with local binary patterns: Application to face recognition. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 28(12), 2006.

[7] J. P. Ahrens, B. Hendrickson, G. Long, S. Miller, R. Ross, and D. Williams. Data-intensive science in the us doe: Case studies and future challenges. *Computing in Science & Engineering*, 13(6):14–24, 2011.

[8] A. W. Q. G. Al-Zumor. A socio-cultural and linguistic analysis of yemeni arabic personal names. *GEMA: Online Journal of Language Studies*, 9(2):15–27, 2009.

[9] J. F. Allen. Maintaining knowledge about temporal intervals. *Communications of the ACM*, 26(11):832–843, 1983.

[10] M. Ames and M. Naaman. Why we tag: motivations for annotation in mobile and online media. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM, 2007.

[11] V. Astakhov, A. Gupta, S. Santini, and J. S. Grethe. Data integration in the biomedical informatics research network (birn). In *Data Integration in the Life Sciences*, pages 317–320. Springer, 2005.

[12] L. Backstrom, D. Huttenlocher, J. Kleinberg, and X. Lan. Group formation in large social networks: membership, growth, and evolution. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 44–54. ACM, 2006.

[13] P. Barthelmess, E. Kaiser, and D. McGee. Toward content-aware multimodal tagging of personal photo collections. In *Proceedings of the 9th international conference on Multimodal interfaces*. ACM, 2007.

[14] M. Bazire and P. Brézillon. Understanding context before using it. In *Modeling and using context*, pages 29–40. Springer, 2005.

[15] P. Belhumeur, J. Hespanha, and D. Kriegman. Eigenfaces vs. fisherfaces: Recognition using class specific linear projection. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 19(7), 1997.

[16] C. G. Bell, J. Gemmell, and C. Rosson. *Total recall*. Dutton, 2010.

[17] T. Berg and P. N. Belhumeur. Tom-vs-pete classifiers and identity-preserving alignment for face verification. In *BMVC*, volume 1, page 5, 2012.

[18] K. Beyer, J. Goldstein, R. Ramakrishnan, and U. Shaft. When is nearest neighbor meaningful? In *Database TheoryICDT99*, pages 217–235. Springer, 1999.

[19] M. Bicego, A. Lagorio, E. Grosso, and M. Tistarelli. On the use of sift features for face authentication. In *Computer Vision and Pattern Recognition Workshop, 2006. CVPRW'06. Conference on*, pages 35–35. IEEE, 2006.

[20] S. Boll, P. Sandhaus, A. Scherp, and U. Westermann. Semantics, content, and structure of many for the creation of personal photo albums. In *Proceedings of the 15th international conference on Multimedia*, pages 641–650. ACM, 2007.

[21] A. Borodin, G. O. Roberts, J. S. Rosenthal, and P. Tsaparas. Finding authorities and hubs from link structures on the world wide web. In *Proceedings of the 10th international conference on World Wide Web*, pages 415–429. ACM, 2001.

[22] M. Boutell and J. Luo. Photo classification by integrating image content and camera metadata. In *Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on*, volume 4. IEEE, 2004.

[23] M. Boutell and J. Luo. Beyond pixels: Exploiting camera metadata for photo classification. *Pattern recognition*, 38(6), 2005.

[24] E. Brachmann, M. Spehr, and S. Gumhold. Feature propagation on image webs for enhanced image retrieval. In *Proceedings of the 3rd ACM conference on International conference on multimedia retrieval*, pages 25–32. ACM, 2013.

[25] P. Brézillon. Context-based modeling of operators practices by contextual graphs. In *Proceedings of the 14th Conference on Human Centered Processes*, pages 129–137, 2003.

[26] D. Brickley and L. Miller. Foaf vocabulary specification 0.98. *Namespace document*, 9, 2010.

[27] L. Cao, J. Luo, and T. Huang. Annotating photo collections by label propagation according to multiple similarity cues. In *Proceeding of the 16th ACM international conference on Multimedia*. ACM, 2008.

[28] L. Cao, J. Luo, H. Kautz, and T. Huang. Annotating collections of photos using hierarchical event and scene models. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*. IEEE, 2008.

[29] F. Capra. *The web of life: A new scientific understanding of living systems*. Anchor, 1997.

[30] J. J. Carroll, I. Dickinson, C. Dollin, D. Reynolds, A. Seaborne, and K. Wilkinson. Jena: implementing the semantic web recommendations. In *Proceedings of the 13th international World Wide Web conference on Alternate track papers & posters*, pages 74–83. ACM, 2004.

[31] G. Chen and D. Kotz. A survey of context-aware mobile computing research. 2000.

[32] Z. Chen, L. Wenyin, F. Zhang, M. Li, and H. Zhang. Web mining for web image retrieval. *Journal of the American Society for Information Science and Technology*, 52(10):831–839, 2001.

[33] M. Chlistalla, B. Speyer, S. Kaiser, and T. Mayer. High-frequency trading. *Deutsche Bank Research*, 2011.

[34] N. Coffey. http://www.javamex.com/classmexer/.

[35] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *Computer Vision and Pattern Recognition, CVPR 2005. IEEE Computer Society Conference on*, volume 1. IEEE.

[36] R. Datta, D. Joshi, J. Li, and J. Wang. Image retrieval: Ideas, influences, and trends of the new age. *ACM Computing Surveys (CSUR)*, 40(2), 2008.

[37] J. Dean and S. Ghemawat. Mapreduce: simplified data processing on large clusters. *Communications of the ACM*, 51(1):107–113, 2008.

[38] A. K. Dey. Understanding and using context. *Personal and ubiquitous computing*, 5(1):4–7, 2001.

[39] N. Diakopoulos and P. Chiu. Photoplay: A collocated collaborative photo tagging game on a horizontal display. *UIST Adjunct Proceedings*, 2007.

[40] J. M. Diamond and D. Ordunio. *Guns, germs, and steel*. Norton New York, 1997.

[41] A. Doan and A. Y. Halevy. Semantic integration research in the database community: A brief survey. *AI magazine*, 26(1):83, 2005.

[42] D. Dou, D. McDermott, and P. Qi. Ontology translation on the semantic web. In *Journal on data semantics II*, pages 35–57. Springer, 2005.

[43] S. Dumais, E. Cutrell, J. Cadiz, G. Jancke, R. Sarin, and D. Robbins. Stuff i've seen: a system for personal information retrieval and re-use. In *Proceedings of the 26th ACM SIGIR conference on Research and development in Informaion Retrieval*. ACM, 2003.

[44] A. Dumitrescu and S. Santini. Context based semantics for multimodal retrieval. In *IS&T/SPIE Electronic Imaging*, pages 72550C–72550C. International Society for Optics and Photonics, 2009.

[45] A. K. Elmagarmid, P. G. Ipeirotis, and V. S. Verykios. Duplicate record detection: A survey. *Knowledge and Data Engineering, IEEE Transactions on*, 19(1), 2007.

[46] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2):303–338, 2010.

[47] C. Frankel, M. J. Swain, and V. Athitsos. Webseer: An image search engine for the world wide web. 1996.

[48] D. Frohlich, A. Kuchinsky, C. Pering, A. Don, and S. Ariss. Requirements for photoware. In *Proceedings of the 2002 ACM conference on Computer supported cooperative work*. ACM, 2002.

[49] C. Galleguillos and S. Belongie. Context based object categorization: A critical survey. *Computer Vision and Image Understanding*, 114(6):712–722, 2010.

[50] J. Gemmell, G. Bell, R. Lueder, S. Drucker, and C. Wong. Mylifebits: fulfilling the memex vision. In *Proceedings of the tenth ACM international conference on Multimedia*. ACM, 2002.

[51] C. Geng and X. Jiang. Sift features for face recognition. In *Computer Science and Information Technology, 2009. ICCSIT 2009. 2nd IEEE International Conference on*, pages 598–602. IEEE, 2009.

[52] A. Graham, H. Garcia-Molina, A. Paepcke, and T. Winograd. Time as essence for photo browsing through personal digital libraries. In *Proceedings of the 2nd ACM/IEEE-CS joint conference on Digital libraries*. ACM, 2002.

[53] A. Gupta and R. Jain. Managing event information: Modeling, retrieval, and applications. *Synthesis Lectures on Data Management*, 3(4), 2011.

[54] J. Hailpern, N. Jitkoff, A. Warr, K. Karahalios, R. Sesek, and N. Shkrob. Youpivot: improving recall with contextual search. In *Proceedings of the 2011 annual conference on Human factors in computing systems*. ACM, 2011.

[55] A. Y. Halevy. Answering queries using views: A survey. *The VLDB JournalThe International Journal on Very Large Data Bases*, 10(4):270–294, 2001.

[56] J. Hays and A. Efros. Im2gps: estimating geographic information from a single image. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*. Ieee, 2008.

[57] K. Henricksen, J. Indulska, and A. Rakotonirainy. Modeling context information in pervasive computing systems. In *Pervasive Computing*, pages 167–180. Springer, 2002.

[58] D. Henter, D. Borth, and A. Ulges. Tag suggestion on youtube by personalizing content-based auto-annotation. In *Proceedings of the ACM multimedia 2012 workshop on Crowdsourcing for multimedia*, pages 41–46. ACM, 2012.

[59] J. R. Hobbs and F. Pan. Time ontology in owl. w3c working draft, 27 september 2006. *World Wide Web Consortium. http://www. w3. org/TR/owl-time*, 2006.

[60] J.-y. Hong, E.-h. Suh, and S.-J. Kim. Context-aware systems: A literature review and classification. *Expert Systems with Applications*, 36(4):8509–8522, 2009.

[61] G. B. Huang, M. Mattar, T. Berg, and E. Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. 2007.

[62] G. B. Huang, M. Mattar, T. Berg, E. Learned-Miller, et al. Labeled faces in the wild: A database forstudying face recognition in unconstrained environments. In *Workshop on Faces in'Real-Life'Images: Detection, Alignment, and Recognition*, 2008.

[63] R. Jain and P. Sinha. Content without context is meaningless. In *Proceedings of the international conference on Multimedia*. ACM, 2010.

[64] O. Jesorsky, K. J. Kirchberg, and R. W. Frischholz. Robust face detection using the hausdorff distance. In *Audio-and video-based biometric person authentication*, pages 90–95. Springer, 2001.

[65] T. Kindberg, M. Spasojevic, R. Fleck, and A. Sellen. The ubiquitous camera: An in-depth study of camera phone use. *Pervasive Computing, IEEE*, 4(2):42–50, 2005.

[66] J. M. Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM (JACM)*, 46(5):604–632, 1999.

[67] M. J. Krawczyk. Communities in social networks. In *Biometrics and Kansei Engineering, 2009. ICBAKE 2009. International Conference on*, pages 111–116. IEEE, 2009.

[68] N. Kumar, A. C. Berg, P. N. Belhumeur, and S. K. Nayar. Describable visual attributes for face verification and image search. In *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI), Special Issue on Real-World Face Recognition*, Oct 2011.

[69] N. Kumar, A. C. Berg, P. N. Belhumeur, and S. K. Nayar. Describable visual attributes for face verification and image search. In *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, October 2011.

[70] M. Lenzerini. Data integration: A theoretical perspective. In *Proceedings of the twenty-first ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pages 233–246. ACM, 2002.

[71] S. D. Levitt and S. J. Dubner. *Freakonomics Rev Ed LP: A Rogue Economist Explores the Hidden Side of Everything*. HarperCollins, 2006.

[72] X. Li, M. Larson, and A. Hanjalic. Geo-visual ranking for location prediction of social images. In *Proceedings of the 3rd ACM conference on International conference on multimedia retrieval*, pages 81–88. ACM, 2013.

[73] X. Li, C. G. Snoek, M. Worring, and A. W. Smeulders. Fusing concept detection and geo context for visual search. In *Proceedings of the 2nd ACM International Conference on Multimedia Retrieval*, page 4. ACM, 2012.

[74] X. Liu and B. Huet. Heterogeneous features and model selection for event-based media classification. In *Proceedings of the 3rd ACM conference on International conference on multimedia retrieval*, pages 151–158. ACM, 2013.

[75] J. Luo, Y. Ma, E. Takikawa, S. Lao, M. Kawade, and B.-L. Lu. Person-specific sift features for face recognition. In *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*, volume 2, pages II–593. IEEE, 2007.

[76] F. Monaghan and D. O'Sullivan. Automating photo annotation using services and ontologies. In *Mobile Data Management, 2006. MDM 2006. 7th International Conference on*. IEEE, 2006.

[77] G. K. Mostefaoui, J. Pasquier-Rocha, and P. Brezillon. Context-aware computing: a guide for the pervasive computing community. In *Pervasive Services, 2004. ICPS 2004. IEEE/ACS International Conference on*, pages 39–48. IEEE, 2004.

[78] M. Naaman. *Leveraging geo-referenced digital photographs*. PhD thesis, Citeseer, 2005.

[79] M. Naaman, R. Yeh, H. Garcia-Molina, and A. Paepcke. Leveraging context to resolve identity in photo albums. In *Digital Libraries, 2005. JCDL'05. Proceedings of the 5th ACM/IEEE-CS Joint Conference on*. IEEE, 2005.

[80] B. Nowack. Confoto: Browsing and annotating conference photos on the semantic web. *Web Semantics: Science, Services and Agents on the World Wide Web*, 4(4), 2006.

[81] N. F. Noy. Semantic integration: a survey of ontology-based approaches. *ACM Sigmod Record*, 33(4):65–70, 2004.

[82] N. O'Hare and A. Smeaton. Context-aware person identification in personal photo collections. *Multimedia, IEEE Transactions on*, 11(2), 2009.

[83] L. Page, S. Brin, R. Motwani, and T. Winograd. The pagerank citation ranking: bringing order to the web. 1999.

[84] Y. Pang, Y. Yuan, X. Li, and J. Pan. Efficient hog human detection. *Signal Processing*, 91(4):773–781, 2011.

[85] D. J. Patterson. *Assisted cognition: compensatory activity assistance technology*. PhD thesis, 2005.

[86] P. J. Phillips, P. J. Flynn, T. Scruggs, K. W. Bowyer, J. Chang, K. Hoffman, J. Marques, J. Min, and W. Worek. Overview of the face recognition grand challenge. In *Computer vision and pattern recognition, 2005. CVPR 2005. IEEE computer society conference on*, volume 1, pages 947–954. IEEE, 2005.

[87] P. J. Phillips, H. Wechsler, J. Huang, and P. J. Rauss. The feret database and evaluation procedure for face-recognition algorithms. *Image and vision computing*, 16(5):295–306, 1998.

[88] R. Poppe. A survey on vision-based human action recognition. *Image and vision computing*, 28(6):976–990, 2010.

[89] E. PrudHommeaux, A. Seaborne, et al. Sparql query language for rdf. *W3C recommendation*, 15, 2008.

[90] S. Ranwez, V. Ranwez, J. Villerd, and M. Crampes. Ontological distance measures for information visualisation on conceptual maps. In *On the Move to Meaningful Internet Systems 2006: OTM 2006 Workshops*, pages 1050–1061. Springer, 2006.

[91] T. Rattenbury and M. Naaman. Methods for extracting place semantics from flickr tags. *ACM Transactions on the Web (TWEB)*, 3(1), 2009.

[92] P. Reignier, O. Brdiczka, D. Vaufreydaz, J. L. Crowley, and J. Maisonnasse. Context-aware environments: from specification to implementation. *Expert systems*, 24(5):305–320, 2007.

[93] E. Sadikov, M. Medina, J. Leskovec, and H. Garcia-Molina. Correcting for missing data in information cascades. In *Proceedings of the fourth ACM international conference on Web search and data mining*, pages 55–64. ACM, 2011.

[94] N. Sambasivan, E. Cutrell, K. Toyama, and B. Nardi. Intermediated technology use in developing communities. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 2583–2592. ACM, 2010.

[95] N. Sawant, J. Li, and J. Wang. Automatic image semantic interpretation using social action and tagging data. *Multimedia Tools and Applications*, 2011.

[96] B. Schilit, N. Adams, and R. Want. Context-aware computing applications. In *Mobile Computing Systems and Applications, 1994. WMCSA 1994. First Workshop on*, pages 85–90. IEEE, 1994.

[97] M. Schneider. What's in my name? toward a generative anthroponomastics. *Anthropoetics*, 15(1):1–9, 2009.

[98] D. Siewiorek, A. Smailagic, J. Furukawa, A. Krause, N. Moraveji, K. Reiger, J. Shaffer, and F. L. Wong. Sensay: A context-aware mobile phone. In *Proceedings of the 7th IEEE International Symposium on Wearable Computers*, volume 248, 2003.

[99] P. Sinha and R. Jain. Concept annotation and search space decrement of digital photos using optical context information. In *Electronic Imaging 2008*, pages 68200H–68200H. International Society for Optics and Photonics, 2008.

[100] B. Smith, M. Ashburner, C. Rosse, J. Bard, W. Bug, W. Ceusters, L. J. Goldberg, K. Eilbeck, A. Ireland, C. J. Mungall, et al. The obo foundry: coordinated evolution of ontologies to support biomedical data integration. *Nature biotechnology*, 25(11):1251–1255, 2007.

[101] Z. Stone, T. Zickler, and T. Darrell. Autotagging facebook: Social network context improves photo annotation. In *Computer Vision and Pattern Recognition Workshops, CVPRW'08.* IEEE, 2008.

[102] M. Turk and A. Pentland. Eigenfaces for recognition. *Journal of cognitive neuroscience*, 3(1), 1991.

[103] V. Vieira, P. Tedesco, and A. C. Salgado. Designing context-sensitive systems: An integrated approach. *Expert Systems with Applications*, 38(2):1119–1138, 2011.

[104] A. Vinciarelli, M. Pantic, and H. Bourlard. Social signal processing: Survey of an emerging domain. *Image and Vision Computing*, 27(12):1743–1759, 2009.

[105] U. Westermann and R. Jain. Toward a common event model for multimedia applications. *Multimedia, IEEE*, 14(1):19–29, 2007.

[106] F. Wolter and M. Zakharyaschev. Spatio-temporal representation and reasoning based on rcc-8. *KR2000: Principles of Knowledge Representation and Reasoning*, 2000.

[107] T.-F. Wu, C.-J. Lin, and R. C. Weng. Probability estimates for multi-class classification by pairwise coupling. *The Journal of Machine Learning Research*, 5:975–1005, 2004.

[108] J. Xiao, J. Hays, K. Ehinger, A. Oliva, and A. Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *Computer vision and pattern recognition (CVPR), 2010 IEEE conference on*. IEEE, 2010.

[109] W. Yang, Y. Wang, and G. Mori. Recognizing human actions from still images with latent poses. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 2030–2037. IEEE, 2010.

[110] R. Yerneni, C. Li, H. Garcia-Molina, and J. Ullman. Computing capabilities of mediators. *ACM Sigmod Record*, 28(2):443–454, 1999.

[111] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang. A survey of affect recognition methods: Audio, visual, and spontaneous expressions. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 31(1):39–58, 2009.

[112] L. Zhang, D. V. Kalashnikov, and S. Mehrotra. A unified framework for context assisted face clustering. 2013.

[113] W. Zhao, R. Chellappa, P. J. Phillips, and A. Rosenfeld. Face recognition: A literature survey. *Acm Computing Surveys (CSUR)*, 35(4):399–458, 2003.

[114] A. Zimmermann, A. Lorenz, and R. Oppermann. An operational definition of context. In *Modeling and using context*, pages 558–571. Springer, 2007.

Appendix A

Appendix

Supplementary material goes here.

A.1 Source Mapping

```
(:source google-calendar
   (:attrs event email time location title description attendee)
   (:rel ev type-of event)
   (:rel owner type-of person)
   (:rel ti type-of time-interval)
   (:rel ev occurs-during ti)
   (:rel participant type-of person)
   (:rel owner participant-of ev)
   (:rel participant participant-of ev)
   (:io disk (:db mongodb))
   (:type personal)
```

```
   (:axioms
      (:map ev event)
      (:map ti time)
      (:map owner.email email B)
      (:map ev.occurs-during time)
      (:map ev.occurs-at location)
      (:map participant attendee)
      (:map ev.title title U)
      (:map ev.description description U)))

(:source fb-user
   (:attrs id name birthday location work email)
   (:rel person type-of person)
   (:rel named-place type-of named-place)
   (:rel address type-of address)
   (:rel person works-at named-place)
   (:rel person lives-at address)
   (:io disk (:db mongodb))
   (:type personal)
   (:axioms
      (:map person.name name)
      (:map person.dob birthday)
      (:map address.street-address location.name)
      (:map named-place.name work.name)))


(:source email
```

```
   (:attrs from to cc)
   (:rel pf type-of person)
   (:rel pt type-of person)
   (:rel pc type-of person)
   (:io disk (:db mongodb))
   (:type personal)
   (:axioms
      (:map pf.email from)
      (:map pt.email to)
      (:map pc.email cc)))


(:source fb-relation
   (:attrs name1 name2)
   (:rel p1 type-of person)
   (:rel p2 type-of person)
   (:rel p1 knows p2)
   (:io disk (:db mongodb))
   (:type personal)
   (:axioms
      (:map p1.name name1 F)
      (:map p2.name name2 U)))

(:source academix
   (:attrs name1 name2)
   (:rel p1 type-of person)
   (:rel p2 type-of person)
```

```
   (:rel p1 knows p2)
   (:io disk (:db mongodb))
   (:type public)
   (:axioms
      (:map p1.name name1 F)
      (:map p2.name name2 U)))

(:source conferences
   (:attrs time location ltitle stitle url)
   (:rel conf type-of event)
   (:rel time type-of time-interval)
   (:rel loc type-of location)
   (:rel conf occurs-at location)
   (:rel conf occurs-during time)
   (:axioms
      (:map time time)
      (:map loc location)
      (:map conf.title ltitle)
      (:map conf.name stitle)
      (:map conf.url url)))


(:source confattendees
   (:attrs url name time location ltitle stitle)
   (:rel conf type-of conference)
   (:rel time type-of time-interval)
   (:rel loc type-of location)
```

```
   (:rel attendee type-of person)
   (:rel attendee participant-in conf)
   (:rel conf occurs-at location)
   (:rel conf occurs-during time)
   (:axioms
      (:map time time)
      (:map loc location)
      (:map conf.title ltitle)
      (:map conf.name stitle)
      (:map conf.url url)
      (:map attendee.name name)))


(:source keynotes
   (:attrs url time location title name)
   (:rel conf type-of conference)
   (:rel k type-of keynote)
   (:rel k subevent-of conf)
   (:rel attendee participant-in k)
   (:axioms
      (:map conf.url url)
      (:map attendee.name name)
      (:map k.location location)
      (:map k.time time)
      (:map k.title title)))

(:source sessions
```

```
   (:attrs url time location title name)
   (:rel conf type-of conference)
   (:rel k type-of session)
   (:rel k subevent-of conf)
   (:rel attendee participant-in k)
   (:axioms
      (:map conf.url url)
      (:map attendee.name name)
      (:map k.location location)
      (:map k.time time)
      (:map k.title title)))


(:source talks
   (:attrs url time location title name)
   (:rel conf type-of conference)
   (:rel k type-of talk)
   (:rel k subevent-of conf)
   (:rel attendee participant-in k)
   (:axioms
      (:map conf.url url)
      (:map attendee.name name)
      (:map k.location location)
      (:map k.time time)
      (:map k.title title)))

(:source conflunches
```

```
   (:attrs url time location title name)
   (:rel conf type-of conference)
   (:rel k type-of lunch)
   (:rel k subevent-of conf)
   (:rel attendee participant-in k)
   (:axioms
      (:map conf.url url)
      (:map attendee.name name)
      (:map k.location location)
      (:map k.time time)
      (:map k.title title)))


(:source tweets
   (:attrs url name)
   (:rel conf type-of conference)
   (:rel attendee type-of person)
   (:rel attendee participant-in conf)
   (:axioms
      (:map conf.url url)
      (:map attendee.name name)))

(:source fb-events
   (:attrs event name time)
   (:rel ev type-of event)
   (:rel p1 type-of person)
   (:rel ti type-of time-interval)
```

```
(:rel ev occurs-during ti)
(:rel p1 participant-of ev)
(:axioms
   (:map p1.name name)
   (:map ev event)
   (:map ev.occurs-during time)
   (:map ti time)))
```