CharmMe: Applying Machine Learning to Facilitate Meaningful Interactions at the MIT Media Lab

by

Victor J Wang

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

Master of Engineering in Computer Science and Engineering

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

August 2012
[SEPTEMBER 2012]

Author ....
Department of Electrical Engineering and Computer Science
August 22, 2012

Certified by ....
Catherine Havasi
Research Scientist
Thesis Supervisor

Accepted by ....
Prof. Dennis M. Freeman
Masters of Engineering Thesis Committee

CharmMe: Applying Machine Learning to Facilitate Meaningful Interactions at the MIT Media Lab

by

Victor J Wang

Submitted to the Department of Electrical Engineering and Computer Science
on August 22, 2012, in partial fulfillment of the
requirements for the degree of
Master of Engineering in Computer Science and Engineering

Abstract

CharmMe is a social discovery application to help people connect with others of similar interests at a company, organization, or conference. Unlike traditional social networking or matching algorithms, CharmMe discovers connections automatically without the need for new profiles or tagging. By using natural language processing, we create a model of an organization by "reading" existing information related to the people being matched, such as their publications or social media accounts. Additionally, the application takes data provided by users Checking-in to conference talks or Liking projects, which are actions made popular by the social networking sites Facebook and Foursquare. To facilitate the actual introduction process, the application makes available the location of all recommended people using RFID technology. In addition, possible opening topics of conversation are suggested based on similar interests shared by users. In this paper, we investigate how effective CharmMe is at predicting new connections that are desirable and describe its deployment during a conference event at the MIT Media Lab. Additionally, we evaluate the effectiveness of the recommendations provided by the system and whether results improve with incorporating user feedback. Ultimately, we think this application will help people create better relationships by encouraging purposeful interactions, eliminating certain social inefficiencies, as well as decrease the opportunity for a missed but potentially meaningful connection.

Thesis Supervisor: Catherine Havasi
Title: Research Scientist

Acknowledgments

I am forever grateful to Catherine Havasi, my advisor, for accepting me into her group at the Media Lab, as well as for being perpetually patient and supportive of my work. My thanks also go out to Rob Speer for providing invaluable programming insights, Jon Ferguson for his technical expertise and help when things broke, and Lance Nathan for seamlessly upgrading the backend API calls.

Contents

List of Figures

Chapter 1

Introduction

Industry professionals often gather at large conferences to learn more about their field. According to the Convention Industry Council, it is estimated that nearly 205 million people attend the 1.8 million conventions and conferences held in the U.S. every year [5].

Implicit of attending these conferences, participants also have the intention of meeting other like-minded individuals who share similar interests. While the CIC reports that $106 billion is spent on hosting national conferences in the US [5], the total contribution is much greater, albeit harder to measure, in the form of business opportunities that arose as a result of people meeting at these conferences.

When conferences are attended by thousands of people (Apple's most recent WWDC sold out it's 5,000 tickets [4]), it can be very difficult to know who exactly are compatible participants to talk to. Because the chances of meeting the right person at the right time are already probabilistically improbable, attendees are left to rely on serendipitous encounters with strangers to forge meaningful connections. Rather than leaving this up to chance and foregoing on missed connections and opportunities, we can use technology to increase the likelihood of meaningful connections, thanks to the emergence of popular social networking websites.

Making meaningful connections is often the primary goal of many attendees at conferences, however, meaningful is defined differently from person to person. This is because each attendee may have a different purpose for participating in the conference.

These reasons may include recruiting candidates for a job opening, exploring business collaboration opportunities, or meeting interesting people who have similar passions. Finding another attendee who's intentions are compatible typically requires some serendipity because a persons intentions are not apparent by physical appearances alone. This often times leads to missed connections and other social inefficiencies.

CharmMe aims to correct these social inefficiencies by helping users facilitate connections that otherwise are left up to chance. Using a combination of each person's social network profile and their most recent activity around the conference, CharmMe algorithmically suggests eligible users to meet who have displayed similar interests and behavior patterns.

Chapter 2

Background

2.1 Online Social Networks

Social networking sites like Facebook and Twitter have discarded anonymity and made it possible for personal profiles and real identities to exist online. Facebook boasts an impressive 900 million users [8]. According to a recent report by Nielson, average user spends nearly seven hours per month on Facebook [14]. Much of their activity includes filling out their profiles, sharing their experiences, and browsing the content generated by their online connections. Back in October 2011, Twitter, another popular social networking site, saw users generate an average of 250 million "tweets" per day [19]. Tweets are short messages or status updates limited to 140 characters.

We can use tools from artificial intelligence to analyze the online data and predict compatibility of users. CharmMe can discover common interests between people and match users who have demonstrated similar online behavior patterns.

Here are some of the biggest existing social networks.

2.1.1 Facebook

Facebook is biggest social network, proudly boasting more than 900 million registered users. Much of someone's online presence is recorded on the site, as users divulge numerous pieces of personal information including their date of birth, hometown, school, etc. Facebook contains vital information about most users' real identities. Facebook allows its users to share a wide range of digital content from pictures, videos, and personal status updates. Users can also Like any material online to signal their preference or affiliation for that medium [8].

2.1.2 Twitter

Twitter is a social network that emphasizes more on helping users share content and distribute information. Unlike Facebook which requires friends to be mutually approved, Twitter allows for asymmetric relationships, under the pretense of following another user. By following another person, users receive text updates directly from that account in real-time [20].

2.1.3 LinkedIn

LinkedIn is a social network but only for business professionals. It solves the problem of trying to stay in touch with a user's professional relationships. Also, by making the site strictly for professionals, privacy issues are also mitigated compared to Facebook [13].

2.1.4 Foursquare

Foursquare is a location-based social network. It gives users the ability to Check-in to a location, thereby announcing users' physical presence to their online friends. Because this check-in information is given up voluntarily, privacy concerns are less of an issue on this service [9].

All of these sites have a public API that enables developers to build third-party applications that can leverage the social information contained in these social networks. Facebook has an API that gives developers access to the Social Graph, which is information about the connections or friendships that exist between users on the site.

2.2 Mobile Technology

With the rise of social media websites like Facebook and Twitter, more and more information about users' preferences and behaviors are being made publicly available online. However, with this explosion of user information comes the problem of deciding which pieces of information are relevant and which pieces are not. Recommendation engines can help users filter through this noise by providing personalized content based on the users' listed preferences, online and off-line actions, and as well as social interactions.

According to another Nielson study, smartphone proliferation has risen to approximately 50% of the U.S. population [15]. This has enabled users to maintain mobility while still being connected online. Users who are on-the-go experience environments that are constantly varying, marked by the changing peers they are immediately surrounded by. This concept of dynamic social networks is being increasingly investigated [18]. One such example is the formation of ad-hoc social networks which do not require Internet access, allowing for much heavier dependency on immediate location. Lee shows how to use mobile behavior in conjunction with information on social media profiles to effectively form these ad-hoc networks [12].

As the consumer products converge with social, mobile, and location technologies, speculation has begun to envision how such applications will impact our daily lives. In 2008, Michael Arrington predicted that mobile social networking was going to be the next big thing [3]. He.imagined a world where all the relevant online information about the people around him would be made available upon walking into a room full of strangers. Depending on the setting and privacy settings of others, someone's resume or listed interests online would be easily accessible just at the fingertips, all in hopes of making the process of meeting new people easier and more efficient. Fast forward to the conference SXSW 2012, people discovery mobile applications have erupted into the technology scene [21].

2.3 Related Work by Startups

2.3.1 South By Southwest

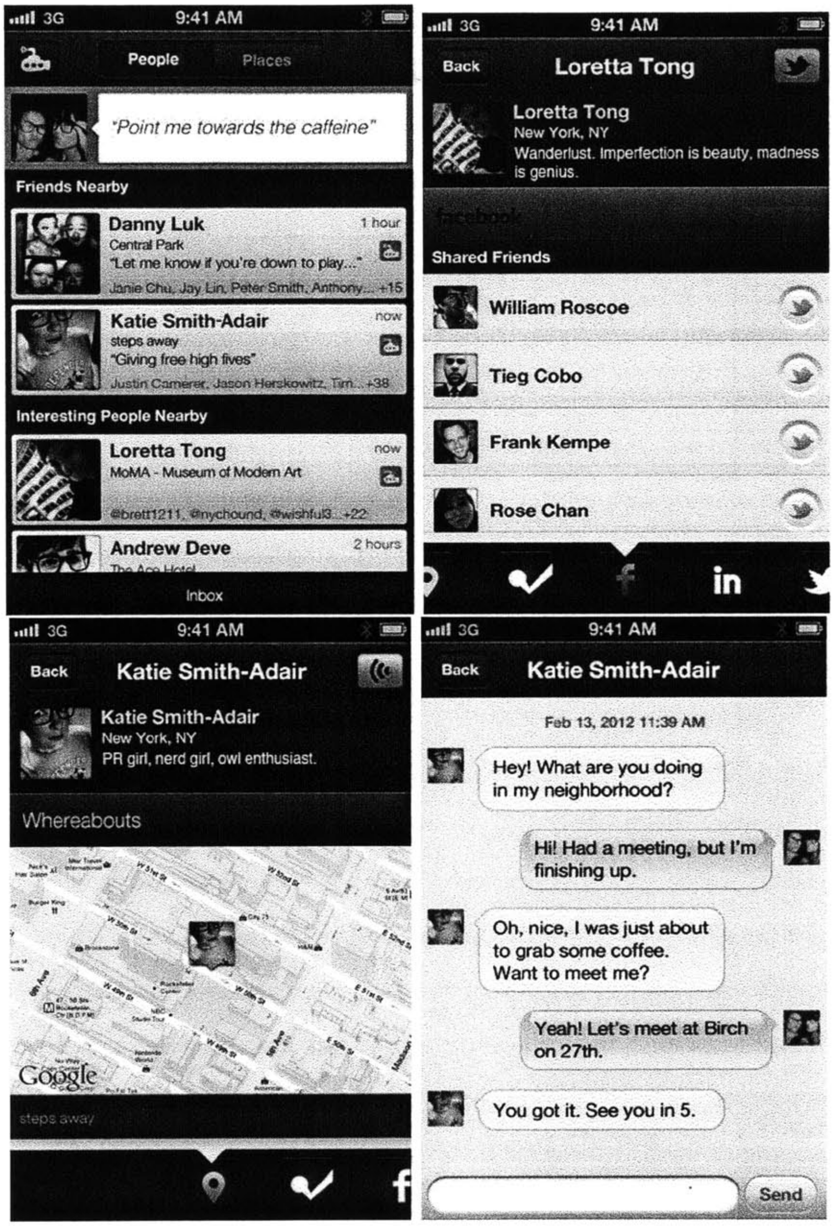
South by Southwest (SXSW) is a popular conference where technologists and startup companies attend to publicly introduce their product to the rest of the world. At SXSW 2012, there was an explosion of so-called people discovery applications, whose goal was to help attendees meet other attendees by recommending nearby and relevant participants to talk to based on common interest, affiliation, or friend. A number of startup companies currently are competing in the background location people discovery space. Such companies include Sonar, Gauss, INTRO, Highlight, and Glancee.

2.3.2 Sonar

Sonar is an iPhone application that uses Facebook, Twitter, and Foursquare to determine interesting people to talk to nearby. When a user comes within 500 meters of a relevant person, Sonar will push notify the user of interesting people. Sonar shows which Facebook and Twitter friends they have in common and also allows users to directly contact that person using a direct message via Twitter. Sonar believes that users are much more inclined to meet new people when they share a mutual friend [16].

2.3.3 Gauss

What separates Gauss from other applications is its feature to allow users to directly affect match results. Users can receive recommendations about surrounding people who also share their specific interests and hobbies based on which interest topic, or “magnet”, they have turned on or off. Gauss also leverages information from Facebook, Twitter, and Foursquare to focus on meeting new strangers in a new environment [10].



"Point me towards the caffeine"

"Let me know if you're down to play..."
Janie Chu, Jay Lin, Peter Smith, Anthony... +15

"Giving free high fives"
Justin Cameron, Jason Herskowitz, Tim... +38

@brett1211, @nychound, @wishful3... +22

Loretta Tong
New York, NY
Wanderlust. Imperfection is beauty, madness is genius.

Katie Smith-Adair
New York, NY
PR girl, nerd girl, owl enthusiast.

Feb 13, 2012 11:39 AM

Hey! What are you doing in my neighborhood?

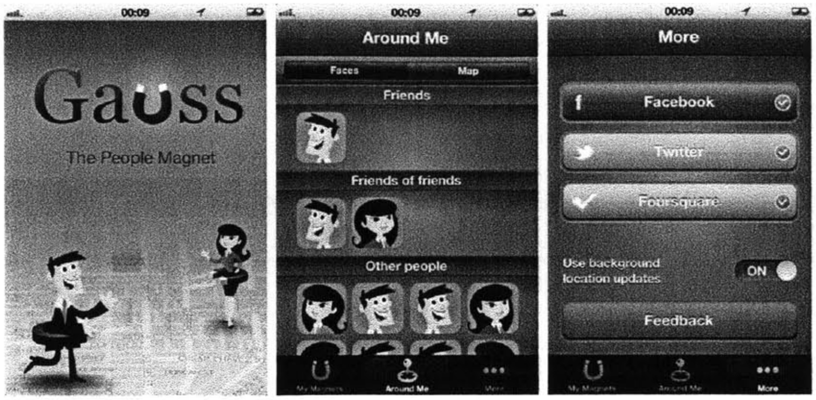
Hi! Had a meeting, but I'm finishing up.

Oh, nice, I was just about to grab some coffee. Want to meet me?

Yeah! Let's meet at Birch on 27th.

You got it. See you in 5.

Figure 2-1: Screenshots of Sonar

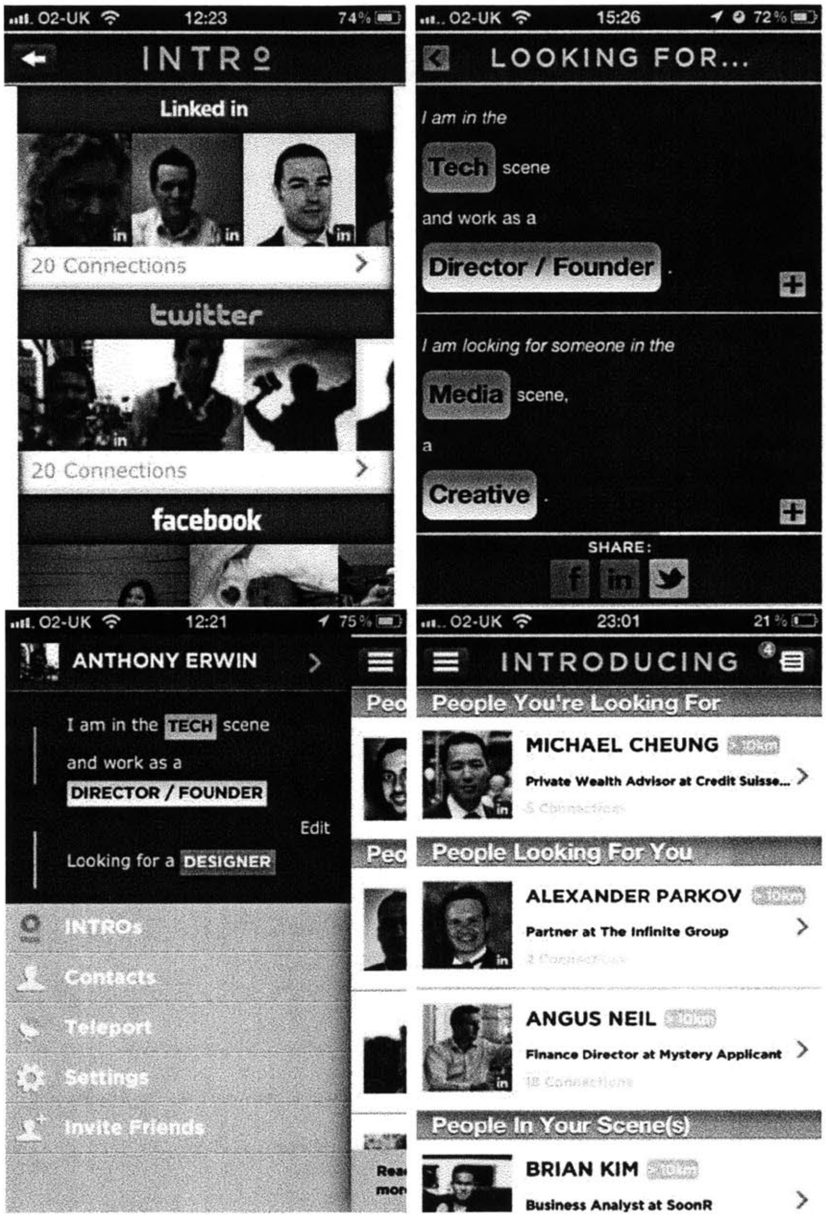


Use background location updates ON

Figure 2-2: Screenshots of Gauss

2.3.4 INTRO

INTRO aims to be the LinkedIn of this space by placing its emphasis only business contacts. Users must sign in using their LinkedIn account and then tell the service what type of professional he is currently seeking. INTRO then searches the user's LinkedIn connections to see if any such person in the user's indirect network is currently located within the user's immediate physical location. INTRO also has a nifty feature, called teleport, which allows users to perform searches in remote locations and reach out to relevant new contacts without being physically present [11].



I am in the Tech scene and work as a Director / Founder.

I am looking for someone in the Media scene, a Creative.

I am in the TECH scene and work as a DIRECTOR / FOUNDER

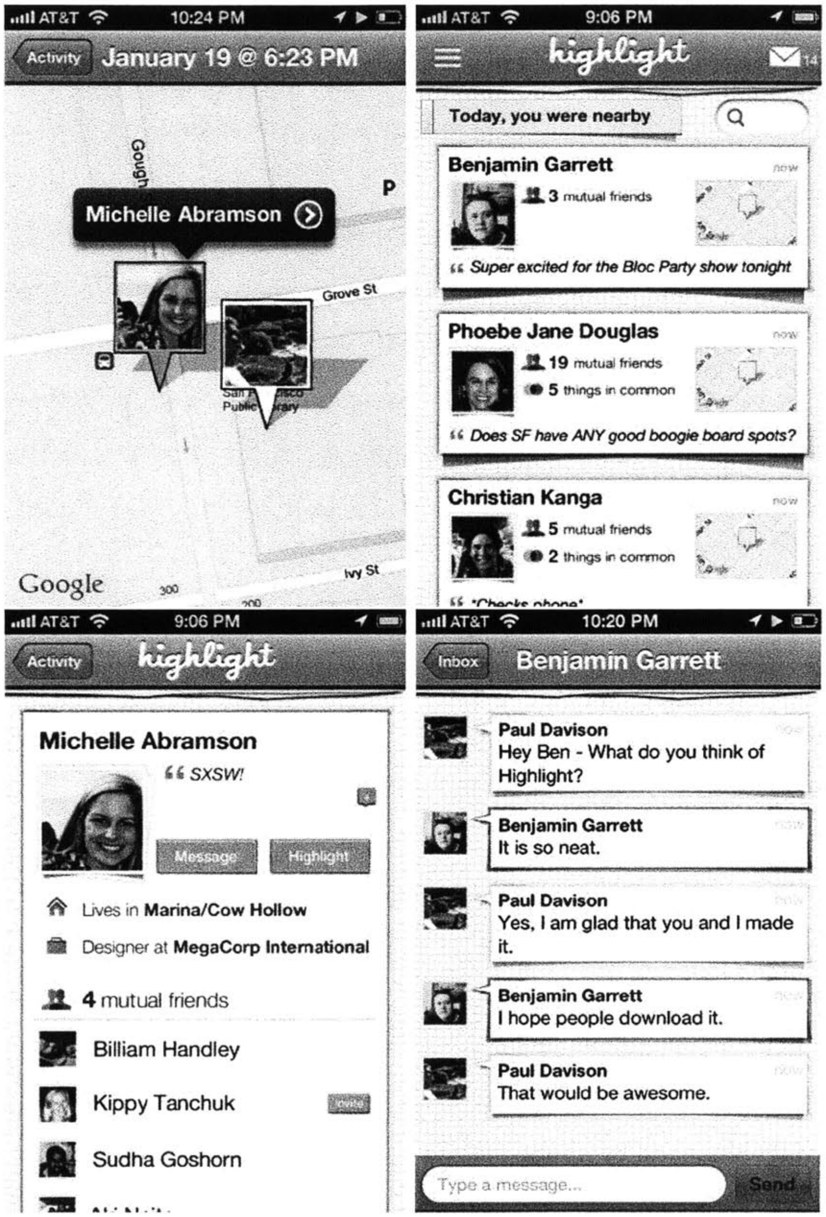
Looking for a DESIGNER

Figure 2-3: Screenshots of INTRO

2.3.5 Highlight and Glancee

Highlight and Glancee were two emerging applications that received a lot of press attention at SXSW 2012. Highlight specifically gave iPhone users the ability to bookmark specific people of interest. A common use case for this was that users could bookmark celebrities in order to receive alerts from Highlight when that celebrity was in proximity. Highlight would then bring up a map to show users how to navigate towards their target contact [2].

Glancee had similar features but was for the Android platform. But instead of showing a map, Glancee tells users the approximate distance of separation to another recommended person. It also makes its recommendations from analyzing the Facebook Likes of each user to find similar categories of interest using Wikipedia. By using encompassing categories in the matching algorithm, the intersecting space of the potential matches is expanded, which may lead to an increased number of actual matches [7].



Benjamin Garrett

3 mutual friends

Super excited for the Bloc Party show tonight

Phoebe Jane Douglas

19 mutual friends

5 things in common

Does SF have ANY good boogie board spots?

Christian Kanga

5 mutual friends

2 things in common

Michelle Abramson

SXSW!

Lives in **Marina/Cow Hollow**

Designer at **MegaCorp International**

4 mutual friends

Billiam Handley

Kippy Tanchuk

Sudha Goshorn

AT&T 10:20 PM

Benjamin Garrett

Paul Davison
Hey Ben - What do you think of Highlight?

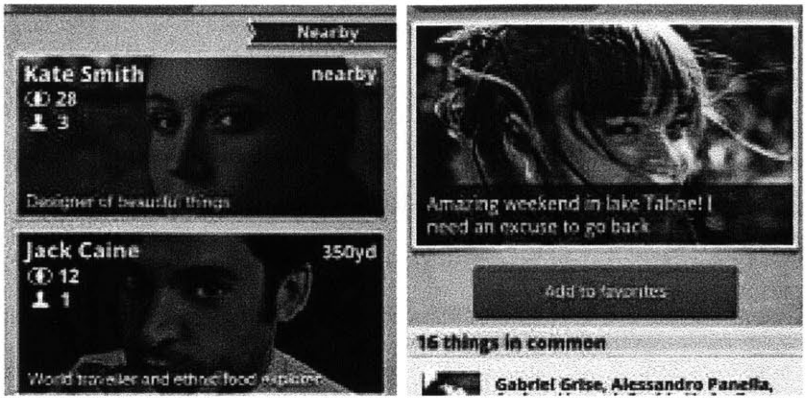
Benjamin Garrett
It is so neat.

Paul Davison
Yes, I am glad that you and I made it.

Benjamin Garrett
I hope people download it.

Paul Davison
That would be awesome.

Figure 2-4: Screenshots of Highlight



Designer of beautiful things

World traveller and ethnic food explorer

Amazing weekend in lake Tahoe! I need an excuse to go back

Figure 2-5: Screenshots of Glancee

CharmMe aims to apply artificial intelligence and machine learning techniques to make process of meeting new people at social entities or events easier and more relevant. In this this paper, we report the deployment and effectiveness of CharmMe at Sponsor Week at the MIT Media Lab. We also explore how accurate recommendations are using the support vector machine (SVM) methodology. Our approach is specifically for the Media Lab building, allowing for much more tailored experience.

Chapter 3

Theory

We briefly explain the machine learning techniques used, singular value decomposition (SVD) and support vector machine (SVM), as well as the underlying math behind these methods.

3.1 Unsupervised Learning and Singular Value Decomposition

Unsupervised learning techniques are used to find patterns or characteristics of unlabeled data. Using unlabeled data is the primary difference between supervised and unsupervised learning algorithms. Solutions find clusters that may exist in the data.

Singular value decomposition is used in linear algebra to factor matrices into component matrices with specific properties. It is given by this formula:

$$A = U\Sigma V^T \tag{3.1}$$

where:

- A is the matrix to be factored
- U is an orthonormal matrix
- Σ is a diagonal matrix containing the eigenvalues

- V^T is the transpose of another orthonormal matrix

To calculate U, multiply A by A^T, or A transpose, and solve for its eigenvectors and eigenvalues. The eigenvalues matrix contains the squared diagonal values of Σ and the eigenvector matrix is U. To calculate V, repeat using the product matrix of A and A^T.

3.2 Supervised Learning

The SVD method is one type of unsupervised learning in machine learning. Unsupervised learning is used to determine clusters in data that share common traits. When active feedback is present, supervised learning algorithms can be used improve predictions. Supervised learning requires training data that is labeled. Thus, the support vector machine, a supervised learning algorithm, was explored to test the accuracy of predictions if the user was able to provide the system with training labels. This provides users with a much more tailored and active approach to meeting people they explicitly say they want to meet. This transforms the space from people discovery to people search.

An example use case is when a recruiter is seeking professionals with a certain skill, for example "programming". By specifying these initial criteria into the application, those with a noted background in programming is deemed a positive example in the training set. As the recruiter meets more people and inputs into the system whether the previous person was actually a good match, the algorithm will modify the positive training set accordingly and retrain itself. Eventually, the algorithm will learn the recruiter's preferences over time and be able to more give more accurate predictions.

The support vector machine algorithm is implemented and its results are compared to how the original SVD algorithm works. We also investigate how the parameters of the Gaussian SVM can be tweaked to change results.

3.2.1 Support Vector Machines

Support vector machines is a type of supervised learning algorithm in machine learning. What this means is that the SVM requires an initial training set of data in order to make predictions. This data needs to contain positive and negative examples such that the SVM can learn which characteristics belong to which group. This is useful because it can answer questions that have binary answers. In this case, the question is "should this user meet this other person?", and appropriately, the answer is either "yes" or "no".

In mathematical terms, an SVM can be represented as a quadratic programming problem.

$$\begin{aligned} max_\alpha \quad & \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{m} y^i y^j \alpha_i \alpha_j x^i \cdot x^j \\ s.t. \quad & \alpha_i \geq 0, \quad i = 1, ..., m \\ & \sum_{i=1}^{m} \alpha_i y^i = 0 \end{aligned} \tag{3.2}$$

where:

- α_i is the calculated weight of data point i
- y^i is the label value of data point i
- x^i is the vector of data point i
- m is the total number of data points

The solution is the linear separator that best divides the positive from the negative data points. This is defined as finding the line that produces the largest margin between the boundary of positive and negative points. Ultimately, all of the α values will be equal to zero with the exception of the data points located on the boundary. Those α values will be positive.

What makes the SVM method particularly powerful is how this linear separator can be found even in cases where there does not appear to be a linear division using a kernel function. The kernel function, $K(x^i, x^j)$, is any function that takes parameters by pairwise replaces the dot product of x^i and x^j in Equation 3.2.

By changing the kernel function, the data points can be moved to a higher dimensional feature space where a linear separator does exist. This solution is then mapped back into the original feature space.

The Gaussian kernel function is used to introduce a nearly infinite feature space to the data set. It has the formula:

$$K(x^i, x^j) = e^{\frac{\|x^i - x^j\|^2}{2\sigma^2}} \tag{3.3}$$

The fit of the separator is dependent on the value of σ, or sigma. In general, the higher the sigma, the more likely the data will be overfit and not provide much predictive power. On the contrary, the lower the sigma the less defined the boundary is, which also gives poor predictions.

Later in the paper, we look at the effect of sigma on the predictive power of the SVM.

Chapter 4

Implementation

4.1 Sponsor Week at the MIT Media Lab

The MIT Media Lab Sponsor Week is a biannual conference held every semester where nearly 500 company representatives tour the Media Lab in a one week period to learn about all of the fascinating research being conducted. They are primarily interested in fostering relationships with Media Lab researchers to explore possible opportunities of collaboration. CharmMe aims to make establishing these relationships easier by suggesting corporate members to meet relevant researchers based on the member's activity around the Media Lab. Such activity include Checking-in to conference talks or Liking projects by taking pictures of a talk's or project's corresponding QR code.

A QR code, short for Quick Response code, is a visual pattern encoded with information not unlike the ubiquitous bar code. Taking a picture of the QR code with a smartphone QR code Reader typically leads the user to a website address. Each project in the Media Lab has a certain web address and is used to register as a charm.

Upon registration, each member of the conference is given a name tag, embedded with a small radio-frequency identification (RFID) metal strip. Each strip provides a unique signal, which directly maps to a four character personal ID called the webcode, also unique to each member. Members are identified by their individual webcodes in the Media Lab's internal database and system infrastructure. As members walk around the lab, their signals can be tracked using RFID readers strategically placed at various locations around the building. The reader itself is conveniently placed under a Glass Infrastructure (GI), which we will cover next.

4.2 The Glass Infrastructure

The GI is an interactive touch screen used to navigate and discover projects around the lab. From any GI screen, users can see a list of all the currently active groups in the lab. Drilling into any specific group brings up a description of the group, all affiliated researchers of the group, as well the projects being worked on by the group.

Because each RFID reader is uniquely matched to a specific GI at a known location, members can log in at the screen they are currently standing in front of. As members explore the content and research projects, they have the option of adding any project to their own personal list of favorite projects. The Media Lab also calls this "charming" a project, and each favorite project is now termed a "charm". This is analogous to Liking something in the physical real world.

Navigating on the GI screen is quite the pleasant experience as objects smoothly and elegantly move around to reposition themselves on the display in response to the user's touch. This rearrangement animation occurs when users tap on a button that drills down into a group's page or a project's page.

CharmMe can also be launched in the GI and viewed. The UI automatically adjusts its screen size to fit the dimensions of the display. Accordingly, the font, buttons and graphics are appropriately enlarged such that no manual zooming of the page is necessary for convenient viewing.

4.3 Backend of CharmMe

In total, there are 330 Media Lab researchers captured in this dataset. The original dataset was simply a list of text documents, each containing the description of one project. Each document also identified the associated list of collaborators by their user names and associated research group for that project. A separate list of users and their corresponding list of charmed projects were also included.

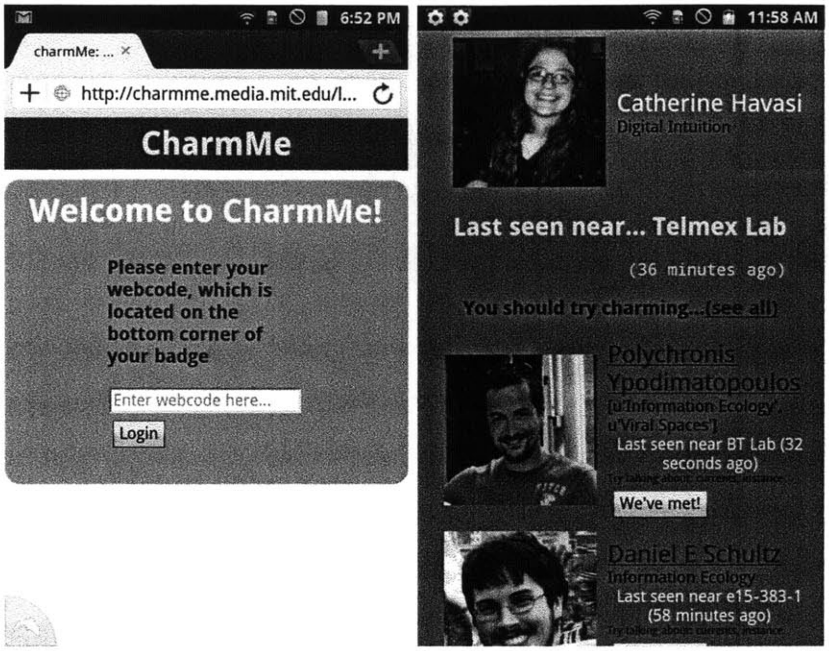
From the original dataset, each user is made into a vector who's component values were term frequency-inverse document frequency (TFIDF) weights. Each weight depicts how important a particular word associates with that user based on the dataset's listed text documents.

Similarity between vectors is determined by calculating the dot product between vector pairs. Because the dimensionality of the feature space is so high, this can be difficult to work with and computationally intensive [17]. When all the user vectors are combined to form a sparse matrix, singular value decomposition is used to reduce the dimension of the feature space. SVD calculates the principal components of the sparse matrix which can be used to represent each user vector in a reduced feature space. Calculating the dot product in this reduced space is simpler and serves as an appropriate approximation to the dot product in the original feature space. The users are then ranked in order of similarity and returned back to the frontend.

4.3.1 User Interface of CharmMe

CharmMe originally had a predecessor project called ConnectMe. CharmMe uses the same underlying SVD algorithm to recommend Media Lab researchers of other lab members to meet. With the ubiquity of smartphones usage for corporate members, CharmMe was born to be a mobile friendly version of ConnectMe that also included recommendations to sponsors, or corporate members in attendance. Having a mobile application was more appropriate for the on-the-go behavior displayed by attendees at the Media Lab Sponsor Week. A web application implementation was chosen because both Android and iPhone users could have access to the service.

Upon arriving to the site, users are prompted to input their unique webcode, found on their Media Lab member badge. If the webcode is valid, users will enter their personal profile page where their basic information is displayed along with the picture taken when they first arrived at the lab. Users are also given the option of logging into their LinkedIn accounts to facilitate following up with connections after the first



6:52 PM

charmMe: ...

http://charmme.media.mit.edu/l...

CharmMe

Welcome to CharmMe!

Please enter your webcode, which is located on the bottom corner of your badge

Enter webcode here...

11:58 AM

Catherine Havasi
Digital Intuition

Last seen near... Telmex Lab

(36 minutes ago)

You should try charming...(see all)

Polychronis Ypodimatopoulos
[u'Information Ecology', u'Viral Spaces']
Last seen near BT Lab (32 seconds ago)
Try talking about: currents, instance

Daniel E Schultz
Information Ecology
Last seen near e15-383-1 (58 minutes ago)
Try talking about: currents, instance

Figure 4-1: Screenshots of the CharmMe user interface

introduction. In yellow is the user's last seen location in the lab, or more precisely, the last GI and corresponding RFID reader that detected the user. Scrolling down on the page reveals the top three most relevant other members they should connect with, sorted by strength of match and location distance. These member recommendations are categorized into Sponsors and Researchers. Pressing the We've met button on a person will cause that person to disappear from the view. Scrolling to the end of the page displays the charms the user most recently input into the system.

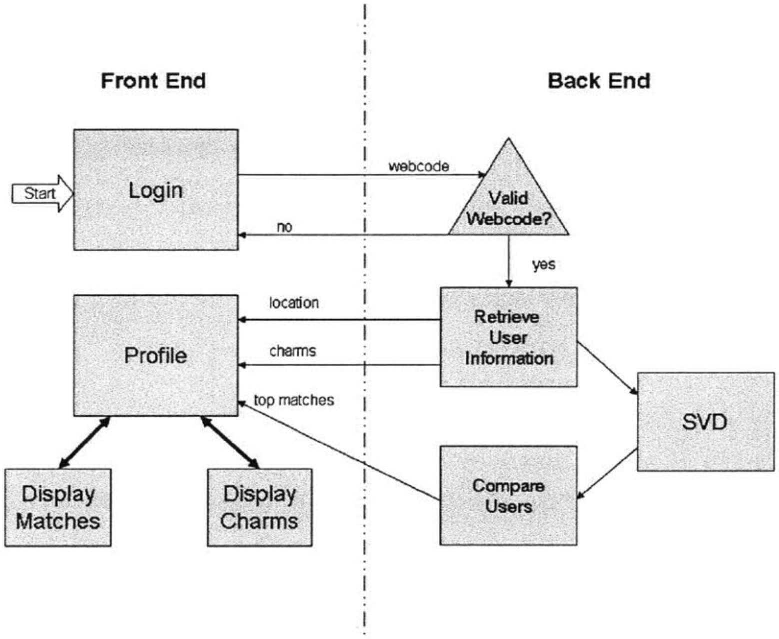


Figure 4-2: Software flow of CharmMe

Chapter 5

Deployment

In this section, we discuss the design considerations and unexpected problems that arose centered around deployment during the Fall 2011 and Spring 2012 MIT Media Lab Sponsor Weeks.

5.1 Design Considerations

5.1.1 Speed Optimizations

When running the SVD algorithm on a person, it takes approximately 10 seconds to return the top ranking recommended matches. From a user experience perspective, 10 seconds is much too long, especially since the application is being viewed on a handheld device. To fix this issue, a caching system was implemented. While first time users require the SVD algorithm to finish, it is redundant to run the SVD algorithm again for returning users. Therefore, upon first login, the user's results are stored into a SQL database. When returning users load their profile, the load time is reduced dramatically, typically to load times of under a second, because the information can just be retrieved from the database. To maintain freshness of results, should a user log in after a certain expiration time, the SVD algorithm will run again to update the recommendations for the user.

5.1.2 Mobile-Friendly View

With the prevalence of smartphones in today's corporate world, the UI was optimized for a mobile experience. CSS3 provides certain powerful commands that can allow automatic readjustment of front-end components, depending on the actual dimensions of the physical display. Buttons, images, and text font-sizes are all resized to fit nicely within the device display. The window itself is also readjusted for vertical scrolling only, while extra screen space in the horizontal direction is eliminated.

5.2 Unexpected Deployment Issues

During the initial deployment of CharmMe at the 2011 Fall Media Lab Sponsor Week, members signing in to the service for the first time were greeted rather unpleasantly with an error screen. This happened because many members did not have charms registered in the system. As a result, the SVD algorithm was unable to create vectors of the first time users and no recommendations could be made. Unfortunately, this case was neglected to be accounted for during testing. Instead of routing users to a custom application error page, the more generic Internal Server Error page abruptly appeared. Many first time users were discouraged from using the site and thought little of returning to using CharmMe.

When CharmMe was functional, many members had difficulties using the service. The main reason was because to get started required a rather lengthy setup process. As a web application, users needed to enter in a web URL. Because the string charmme.media.mit.edu is tedious to type on a phone, a QR code pointing to the CharmMe address was made available for faster access. However, this introduced another unforeseen step into the process because many corporate members surprisingly did not have a QR code reader application on their phones. Downloading a QR code reader added more time to the already time-consuming start up process.

When landing on the login page, members had to input their personal webcode. Many members did not know their webcode information, or even how to discover their own webcode by looking on their name badge.

The final step for the first time user was to charm projects and give the SVD algorithm data about the member. We had to demonstrate to members how to interact with the GI and charm projects for them in order to enter data into the system. This further added to the difficulty of adapting the product.

Another minor problem that was overlooked during testing was how sponsors can only see other sponsors who have already signed in to the CharmMe system. While sponsors can view recommended researchers, the CharmMe algorithm only uses sponsors who have previously logged in before in matches. Therefore, the first sponsors who use the application see a limited to nonexistent number of sponsor matches. This is a problem at the start of the conference when nobody knows of CharmMe, but disappears as more people hear about and use the application.

Chapter 6

Follow-Up Experiment

The support vector machine algorithm is implemented and its results are compared to how the original SVD algorithm works. We also investigate how the parameters of the Gaussian SVM can be tweaked to change results.

Two different ways of labeling the training data were explored, conveniently named Experiment 1 and Experiment 2. Experiment 1 marked positive examples as all names that shared a document with the current user. This means that all researchers that collaborated with the current user on a project were marked as positive examples. Experiment 2 marked positive examples as all names that shared a group with the current user as described in the data documents.

6.1 Error Validation and Sigma

Predictions can be made once each SVM is properly trained. These predictions were then validated by the actual researcher by having each of the four researchers confirm whether they know or would like to know the predicted users. Validation simply consisted of presenting 13 researchers with a list of 66 randomly generated user names sampled from the entire set. The number of errors (or wrong predictions) is counted from the predicted matches generated from the Gaussian SVM. An error is a positive prediction that the actual researcher claims she does not know or a negative prediction that the researcher claims to know or want to meet.

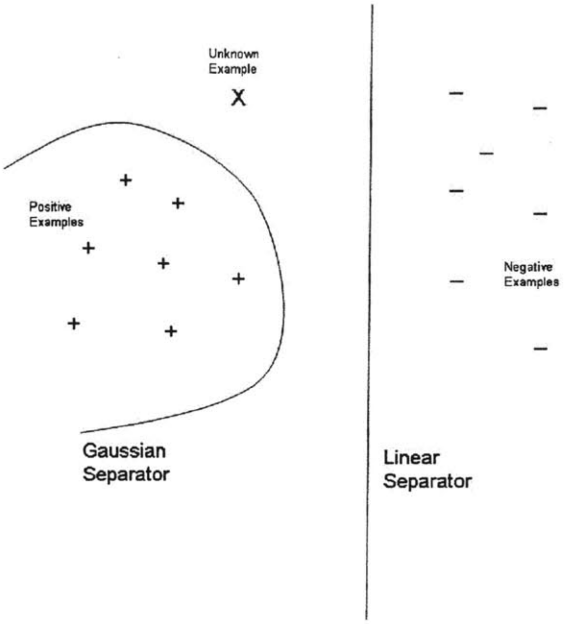


Figure 6-1: Gaussian SVM overfitting

The best sigma is determined by minimizing the following formula:
minimize with variables e, d with objective function $e + d$ where $e =$ error, $d =$ (total recommended by SVM less total extrapolated positives from validation results)

This formula takes into account both the accuracy and predictive power of the SVM. The error, e, is the number of wrong predictions made while d is how close the is the number of SVM predictions made comes to the extrapolated number of positive results from each researcher's validation results.

It surprised me when we first discovered how it was possible that a Gaussian kernel could produce less recommendations than a linear kernel. However, this could be a product of overfitting. If a Gaussian kernel's sigma is too small, the SVM loses its predictive power simply because only the positive training examples will ever be positively labeled. Figure 6-1 illustrates this point. When sigma becomes too large, the Gaussian boundary becomes more flat in shape. This also causes the Gaussian kernel to lose its predictive power since it will just end up including the entire data set as positive examples.

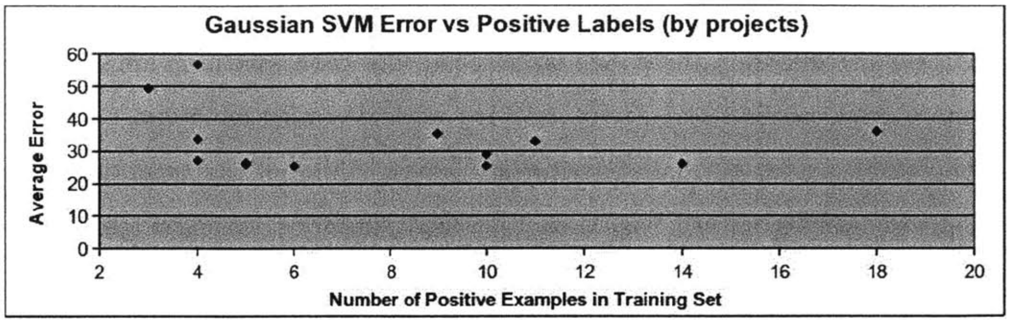


Figure 6-2: Error versus positive labels, by projects.

When a Gaussian kernel has more positive examples to train, the optimal sigma may take on a lower value. This is because the Gaussian can form a more complex separator to envelope other positive examples. Of course, one must worry about overfitting. On the other hand, if there are not enough positive examples in the training set, then the sigma must be big enough so that the separator can extrapolate more on other data points.

6.2 Training Data and Average Gaussian SVM Error

Next, we investigate if the average error rate from the Gaussian SVM is dependent on how many positive examples in the training data. Both methods are analyzed, by groups and by projects.

While the method by projects had a lower number of positive examples in the training set, the observations in error rate was very similar. Error rate was around 30 with no correlation to the number of positive examples in the training set.

6.3 Training Data and Total Predicted Matches

In both cases, there seems to be a slight positive correlation between number of positive labels and number of positive predictions when using the linear SVM. Both cases

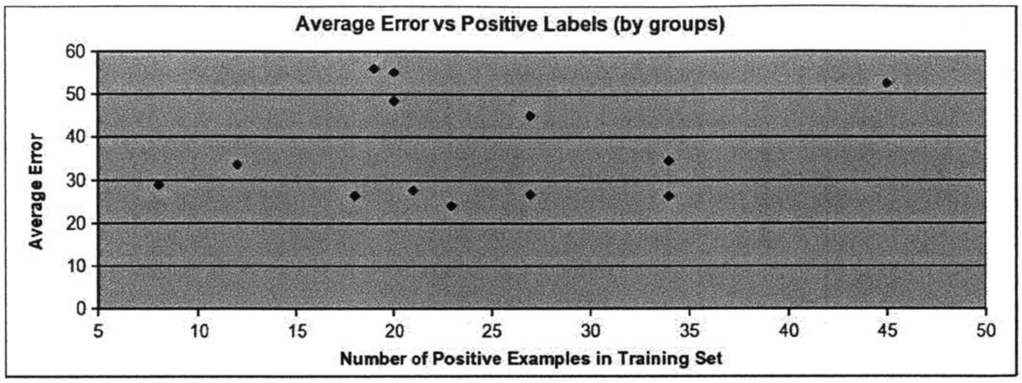


Figure 6-3: Error versus positive labels, by groups.

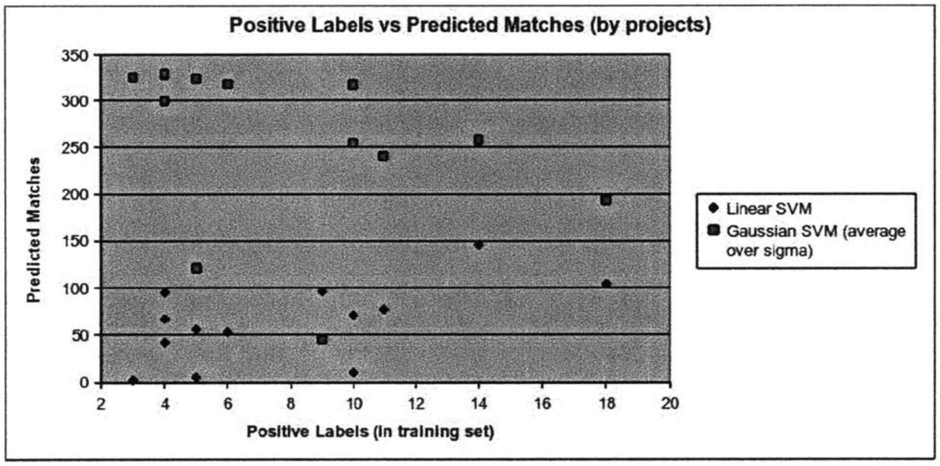


Figure 6-4: Positive labels versus number of predicted matches, by projects.

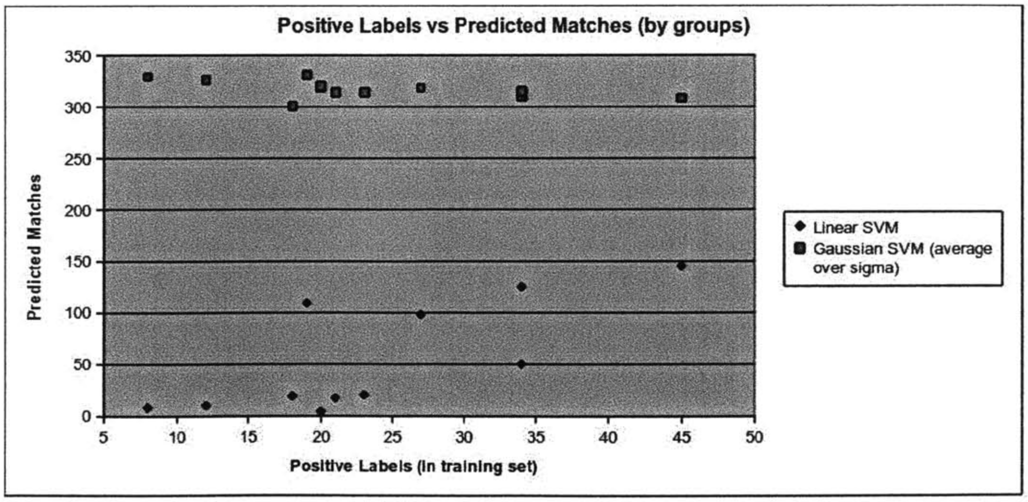


Figure 6-5: Positive labels versus number of predicted matches, by groups.

also show a significant difference in average number of predicted matches between the linear SVM and Gaussian SVM techniques. By projects, however, shows a slight negative relationship when using the Gaussian SVM, when compared to its linear counterpart.

6.4 Discussion

The biggest challenge from a computation standpoint was solving the optimization problem presented by an SVM. Using the python package CVXOPT, training each SVM on 330 points took on average of about 10 seconds. To boost results, Leave-One-Out Cross Validation (LOOCV) was implemented but never tested with because running LOOCV would take about an hour to complete. When testing on my data, running the entire script would take approximately 8 minutes.

Labeling the training data was also an interesting challenge to ponder. Without the help of social networks, it is very hard to determine who knows who. An initial starting point is having positive labels for those who work in the same research group. However, this was problematic as certain research groups are very small, only consisting of 4 total people. Upon reiteration, a good starting point is indeed deeming those who worked on the same project as connected. This is because there are more projects per research group. This was Experiment 1.

Validating the results of the predictions also poses a challenge. The goal of this project is to determine new people to meet based on the people already known. However, the line between relevant people that one should meet versus people that one wants to meet are different. For simplicity, we treated all positive predictions as people whom should be met. But one could imagine how different positive training examples would be needed to determine those that users would like to meet. One possible future work could be to use the validation results as positive training examples. This method might be a more suitable approach to recommend people by interest.

The last challenge is computing the error. Due to the validation process, it is appropriate to extrapolate because the list of 66 users were generated at random.

However 66 is only a fifth of the total sample size and it is possible that this is too small to capture enough information. False negative errors could therefore be unrepresentative because we don't know if the researcher actually doesn't know the person in question or that person's name simply did not come up in the random sampling.

Chapter 7

Results

Predictions can be made once each SVM for each person is properly trained. These predictions were then validated by the actual researcher by having each of the researchers confirm whether they know or would like to know the predicted users. Validation simply consisted of presenting the researchers with a list of 66 randomly generated user names sampled from the entire set.

7.1 Comparing Error Rates by Algorithm

First, we compare the error rate across the three algorithms, SVM with Gaussian kernel, SVM with linear kernel, and the SVD. An error is a positive prediction that the actual researcher claims she does not know (false positive) or a negative prediction that the researcher claims to know or want to meet (false negative).

We also vary the initial positive examples given in the training data using two separate methods. Experiment 1 marked positive examples as all names that shared a document with the current user. This means that all researchers that collaborated with the current user on a project were marked as positive examples. Experiment 2 marked positive examples as all names that shared a group with the current user as described in the data documents.

As you can see in Figure 7-1 and Figure 7-2, there is not much discrepancy between the two experiments in terms of error rates. However, it is interesting to note that the

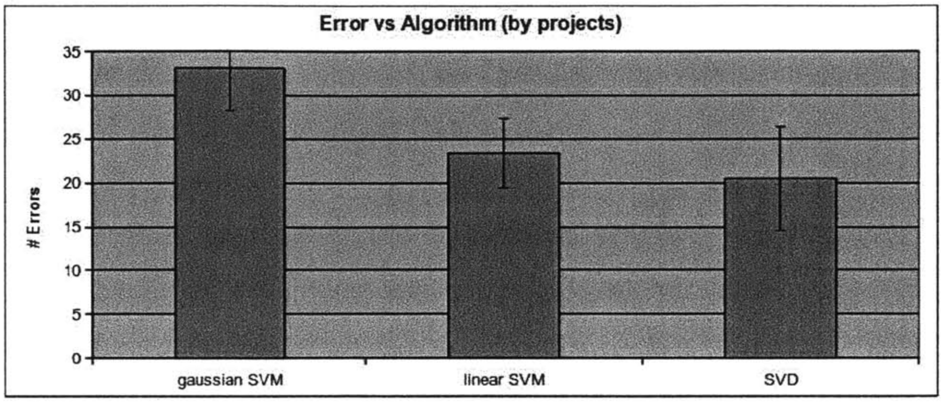


Figure 7-1: Error across the three algorithms, by projects.

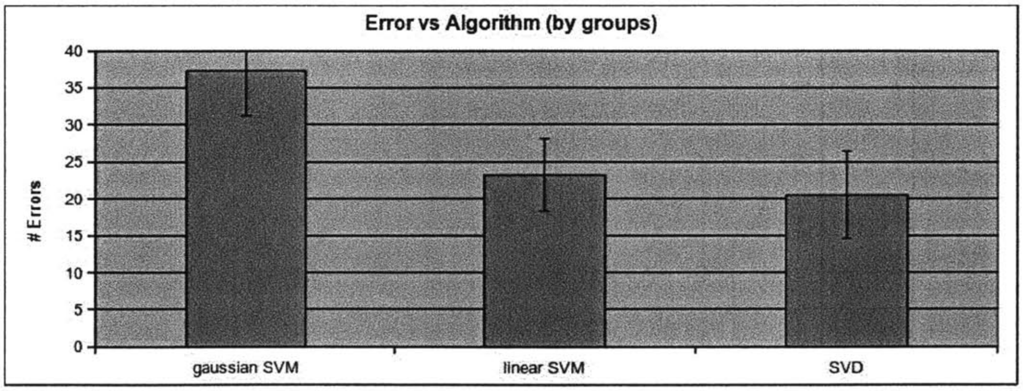


Figure 7-2: Error across the three algorithms, by groups.

Gaussian SVM consistently had the highest error rate while the SVD consistently had a lowest error rate. The SVD also in both methods had the highest standard deviation which implies that it had the potential to be the most accurate of the algorithms.

An important note about the SVD algorithm is requirement for a threshold value to distinguish matches from non-matches. Because SVD is a clustering algorithm, it produces a range of values that determine the strength of the match for each pair of user. The threshold value for evaluating the strength of a match is set arbitrarily, in this case to 0.75. Moving this number higher or lower would result in less or more number of matches, respectively. As a future work, this value could be left up to the user to input, depending on how many results the user wanted.

Chapter 8

Discussion and Future Work

8.1 Interview with Reid Hoffman

Besides the technical difficulties of getting started to use CharmMe, there are various social reasons why these background location applications for meeting new people have not reached their pinnacle in popularity. Reid Hoffman, Founder of LinkedIn, discusses in an informal conversation why this could be.

First, meeting new people simply isn't the most pressing problem on the top of people's minds. In a world where so many different things compete for your attention, it is very easy to become distracted by answering emails or responding to the connections users already have.

Even in settings where it is an explicit goal of the event to make new connections, people naturally still gravitate towards those that they already know. This is because it is inherently awkward to make a cold introduction to another person. It takes a certain level of individual confidence and comfort to act with extroversion. Those who have more introverted personalities often find meeting new people even more daunting due to their inherent personality.

Furthermore, the avid "networker" is the exact person we try to avoid when at any social gathering. These types of people appear "slimy" and "sleazy" - words that carry rather negative connotations. Appearing disingenuous or insincere comes off as self-serving and makes others feel manipulated and used.

Perhaps a better route to take is to make meeting new people an implicit goal, rather than an explicit goal. Many startups, including Meetup and UrbanOrca, do this by helping users come together in person through organization of an activity, event, or cause. While the explicit goal is to participate in the event itself, in this process, members get to meet other like-minded people similar to themselves, which allows them to form new connections. Meetup boasts over 11 million users in 105,000 different local groups from 45,000 cities worldwide [1].

8.2 Future Work

There are many directions with potential that builds on this project. Partnering companies have particularly expressed interest in an application that helps the formation of teams by suggesting compatible personnel with complementary skill sets and interests. The application would take in as input various specifications revolving around the purpose of the team creation, involved tasks, and necessary skills required. The application uses such criteria to search through people in the system to find the ideal group of people suited for the specification description. Other elements into the algorithm can also be introduced, such as previous work history and personality types, to output a team that is optimized to collaborate the most efficiently and effectively.

This idea of discovering relevant people can be applied to other use cases as well. A large company with thousands of employees has to cope with the natural friction that arises when processes for collaboration scale. Often times, a rigid and hierarchical social structure is put into place, leading to a consequently slow deliberation of decisions. Getting things done efficiently becomes dependent on who you know.

CharmMe can be used at large corporations to discover opportunities of synergy between employees who may not have many chance encounters. This may be because large companies tend to have big office buildings which inherently limit how much interaction any single employee may have with other coworkers due to natural barriers of proximity.

By suggesting other colleagues based on profile information, employees are empowered to their job much more efficiently. This enables employees to leverage their relationships within the company to increase productivity. Such mixing of business backgrounds may also encourage creativity to perpetuate innovation [6]. This also promotes a friendly and close-knit culture which may make a more enjoyable work environment.

Another use case for a location-aware intra-network may be to improve promptness when attending company meetings and events. While this may initially seem like infringement on personal privacy, employers could claim to have the rights to know where their employees are during work hours. Furthermore, this service may be useful for times when a meeting adjudicator needs to know if it is worth delaying the start of a meeting when attendees are running late. Usually, during business settings such as at work or conferences, participants do not mind having their location publicly known since they are in public location. They especially do not mind if having their location broadcasted can result in more connections with meaningful people.

Chapter 9

Conclusion

People have long attempted to use technology to help meet new friends. This goal of making new meaningful connections is now more possible than ever thanks to the emergence of popular social, mobile, and location-based Internet products. We have explored the effectiveness of such an application at the MIT Media Lab, applying Machine Learning techniques to the data at hand to recommend other relevant people to users. While social convention is largely responsible for the current impediment against the widespread adoption of this technology, this trend may be short-lived as we rely more and more on our mobile devices to make connections with those around us.

Bibliography

[1] About meetup. http://www.meetup.com/about/, July 2012.

[2] Highlight about. http://highlig.ht/about.html, June 2012.

[3] Michael Arrington. I saw the future of social networking the other day. *TechCrunch*, April 2008.

[4] Scott Austin. Live: Apples keynote at wwdc 2012. *The Wall Street Journal*, June 2012.

[5] Economic significance study - key findings. *Convention Industry Council*, June 2012.

[6] Beth Comstock. Want a team to be creative? make it diverse. *Harvard Business Review Blog Network*, May 2012.

[7] Eric Eldon. Glancee: A nice-guy ambient social location app for normal people. *TechCrunch*, February 2012.

[8] Facebook newsroom - key facts. http://newsroom.fb.com/content/default.aspx?NewsAreaId=22, June 2012.

[9] About foursquare. https://foursquare.com/about/, June 2012.

[10] Gauss. http://www.getgauss.com/, June 2012.

[11] Intro. https://getintro.net/, June 2012.

[12] Jun Lee and Choong Seon Hong. A mechanism for building ad-hoc social network based on user's interest. *Asia-Pacific Network Operations and Management Symposium*, September 2011.

[13] Linkedin about us. `http://press.linkedin.com/about`, June 2012.

[14] May 2012 top u.s. web brands and news websites. `http://blog.nielsen.com/nielsenwire/?p=32201`, June 2012.

[15] Americas new mobile majority: a look at smartphone owners in the u.s. `http://blog.nielsen.com/nielsenwire/?p=31688`, May 2012.

[16] Sonar. `http://www.sonar.me`, June 2012.

[17] Robert Speer, Catherine Havasi, and Henry Lieberman. Analogyspace: reducing the dimensionality of common sense knowledge. *Association for the Advancement of Artificial Intelligence*, 2008.

[18] Chayant Tantipathananandh and Tanya Y. Berger-Wolf. Finding communities in dynamic social networks. *IEEE International Conference on Data Mining*, 2011.

[19] Alexia Tsotsis. Twitter is at 250 million tweets per day, ios 5 integration made signups increase 3x. *TechCrunch*, October 2011.

[20] Twitter about. `https://twitter.com/about`, June 2012.

[21] Jenna Wortham. New apps connect to friends nearby. *The New York Times*, March 2012.