**Cros-Domain User Modeling: Applying Grph Techniques for Reasoning on Social Network Personal Data**

**Research proposal for M.sc degree in M.I.S department**

**Oz Shapira**

Contents

[1. Introduction 3](#_Toc341959324)

[2. Background and Related works 5](#_Toc341959325)

[2.1. Background 5](#_Toc341959326)

[2.1.1. Recommender systems 5](#_Toc341959327)

[2.1.2. Graphs as data structures 6](#_Toc341959328)

[2.1.3. Social networks 6](#_Toc341959329)

[2.2. Related work 7](#_Toc341959330)

[2.2.1. Generic Semantic-based Framework 7](#_Toc341959331)

[2.2.2. Link Data cloud 7](#_Toc341959332)

[2.2.3. Collecting data form social networks (SN) 8](#_Toc341959333)

[2.2.4. Social network analysis (SNA) 8](#_Toc341959334)

[2.3. Summery 8](#_Toc341959335)

[3. Research Goals and Questions 9](#_Toc341959336)

[4. System and Methods 9](#_Toc341959337)

[4.1. Methods 10](#_Toc341959338)

[4.1.1. Data Source 10](#_Toc341959339)

[4.1.2. Selecting pinterest Social network s 11](#_Toc341959341)

[4.1.3. Crawling Pinterest 12](#_Toc341959342)

[4.1.4. Crawling process 12](#_Toc341959343)

[4.1.5. Data collecting: 13](#_Toc341959346)

[4.1.6. Graph creation 13](#_Toc341959347)

[4.2. Tools 14](#_Toc341959350)

[4.3 Evaluation 15](#_Toc341959351)

[5. Timetable 16](#_Toc341959352)

[6 Research Contributions 16](#_Toc341959353)

[7 References 16](#_Toc341959354)

# Introduction

What if we know that someone loves pizza, James bond movies and he also loves jogging on the beach and will ask for a recommended recipe of a cake he may love? How can we make use of the information we already know about the personal preferences regarding fast food, entertainment and sport in order to suggest a preferred recipe for a cake? This research will try to address this challenge and suggest a recommender system able to answer on this kind of questions.

“Recommender systems represent user preferences for the purpose of suggesting items to purchase or examine. They have become fundamental applications in electronic commerce and information access, providing suggestions that effectively prune large information spaces so that users are directed toward those items that best meet their needs and preferences “[Burke 2002]. Recommender systems became an important research area since the appearance of the first papers on collaborative filtering since the mid-1990s [18, 19, 20]. There has been much work done both in the industry and academia on developing new approaches to recommender systems over the last decade. Examples of such applications include recommending books, CDs and other products at Amazon.com, movies by IMDB , and news at VERSIFI Technologies (formerly AdaptiveInfo.com). [Adomavicius and Tuzhilin 2005].

Most recommender systems are focused on providing a personalized service in a specific domain, as does Pandora – music recommendation system or IMDB movies recommender system (see figure 1 as an example).



Figure 1: from IMDB website as movie recommender system

In order to provide a personalized service to their users, recommender systems need to have relevant personal information about their users, or a “model” of their users - a “User Model”. When this information is available, then the task of recommendation may be straight forward – provide a service based on the relevant information – in our case, if the system knows the user’s preferences about cakes, then finding a similar recipe becomes and easy task. However, in many cases, like in our example, the system does not have this information. This may be the case of a first time visitor to a cultural heritage site/a new city/ restaurant etc. The lack of sufficient user modeling data at the onset of a service is among the classical and well known problems of user modeling and recommender systems – the “cold start” problem [Guo 1997].

Nowadays, on the World Wide Web many sites contain user’s models which reflect user characters. As we surf and visit websites we leave identifiable digital “fingerprints” not to mention explicit definition of interests and preferences, that are used or may be used for personalization purposes. In recent years, social networks became a major source for personal information [boyd 2007].

This freely available personal information, scattered over various online sources (including social networks) may be a valuable source of information for building an initial user model for recommendation. However, even though these social networks may be rich in personal information, they may lack specific personal information that is required for a specific personal service requested.

In order to address this issue, “cross domain” recommendation/personalization was defined – how can we use personal information available about the user in one domain for providing service in another domain [berkovsky 2007] [provide a reference – Shlomo berkovsky’s paper?].

The proposed research is intended to address this issue by integrating personal information that is freely available in social networks with a simple yet powerful graph based representation. User characteristics will be represented by nodes and relations between them will be represented by edges. Traversing the graph will enable to find out relations and links between characteristics that were not explicitly defined in the original information sources. We plan to explore, demonstrate and evaluate the ability to use graph based representation of user modeling for representing and reasoning on data elicited from social network in order to help solving the cross-domain user modeling challenge.

# Background and Related works

## Background

### Recommender systems

Recommender systems are now an integral part of some e-commerce sites such as Amazon.com and CDNow [Schafer, Konstan & Riedl, 1999]., Recommender system applies knowledge discovery techniques to the problem of making product recommendations during a live customer interaction. These systems are achieving widespread success in E-commerce nowadays, especially with the advent of the Internet [Sarwar et al.2000] .there is many type of recommendation systems and in each one of them have different approach for recommitting (vs. Pandora[[1]](#footnote-1) , Google[[2]](#footnote-2) search ,YouTube[[3]](#footnote-3) ,amazon[[4]](#footnote-4) etc) There are several types of recommendation techniques:

#### Content-Based,

Recommendations are based on matching semantic properties (preferences) of items similar to those that user liked in the past. A content-based recommender learns a profile of the user’s interests based on the features present in objects the user has rated. User profile is depends on the learning method employed. There are many methods for establishing recommendation Decision trees, neural nets, and vector-based etc.

Adomavicius [et el 2005] conclude this technique in algorithm if utility u(c, s) (when u is the method used for recommendation) of item s for user c is estimated based on the utilities u(c , si) assigned by user c to items si ∈S that are “similar” to item s. for example let’s assume we have music recommender system M in order to recommend song s to user c, the system will try understand (using utility u) the commonalities among music user c has rated highly in the past – them only songs that have the highest similarity degree to users preference will recommended.

**Content-based disadvantage**-recommendation is based on past user commonalities and not updated to user specific preferences.

#### Collaborative-Based

the most common and mature technique, implement in verities systems where the recommendations are based on previous ratings of users to items, with the assumption that users who agreed in the past on item ratings are likely to agree again in the future, in some case the ratings is binary (like/dislike) like Pandora or real-valued indicating degree of preference like movie rating in IMDB. Some of the most important systems using this technique are GroupLens/NetPerceptions [Resnick et al. 1994], Ringo/Firefly [Shardanand & Maes,1995], Tapestry [Goldberg et al. 1992] and Recommender [Hill et al. 1995]. These systems can be memory-based, some use users to compare against each other use direct approach or other measure. Other system use model-based which a model is derived from the historical rating data to make predictions [Breese et al. 1998]. Model-based recommenders have used a variety of learning techniques including neural networks [Jennings & Higuchi, 1993], latent semantic indexing [Foltz, 1990], and Bayesian networks [Condliff, et al. 1999].

collaborative techniques is are completely independent of any machine-readable representation of the objects being recommended, and work well for complex objects such as music and movies where variations in taste are responsible for much of the variation in preferences. Schafer, Konstan & Riedl (1999) call this “people-to-people correlation.

**Collaborative-based disadvantages** – collaborative based system have suffer from the cold start problem- when recommendation is needed to new user when we have very few data on him.

#### Hybrid systems

”Hybrid recommender systems combine two or more recommendation techniques to gain better performance with fewer of the drawbacks of any individual one” [burke 2002]. Since hybrid systems are combination of several techniques they have the abilities to overcome on each techniques weakness.

All recommender systems, regardless of their underlying technology need to obtain user profile for their recommendation process. Some of those system are an application of a particular type of Knowledge Discovery in Databases (KDD) [Fayyad et al. 1996] technique how investigate the useable database for discovery the “best option” for the recommendation.

### Graphs as data structures

Graph is a representation of a set of objects where some pairs of the objects are connected by links The interconnected objects are represented by mathematical abstractions called vertices, and the links that connect some pairs of vertices are called edges (Wikipedia 2012 ).

The Implementation graph as data structures is common and has many aspects, the main advantage of deploying graph in computerized systems is the abilities to automatically traverse and reason on it. Used graph is to reduce analysis of source-target implantation based on web linked hierarchy (MapReduce problem ) [Dean & Ghemawat 2004] .

### Social networks

Social networks are… They have been with us since 1997 (the first one was sixDegrees.com), social networks site (SNS) have successfully changed worldwide communication. They gave personal users the ability to reach any user in the world. They attracted millions of users, many of whom have integrated these sites into their daily practices. As of this writing, there are hundreds of SNSs, with various technological affordances, supporting a wide range of interests and practices (for example Facebook) those abilities allow SNS the to connect between separate type of population using SNS users, [boyd 2007] rise the fact the SNS can provide rich sources of naturalistic behavioral data. Profile and linkage data from SNSs can be gathered either through the use of automated collection techniques or through datasets provided directly by the company, enabling network analysis researchers to explore large-scale patterns of friending, usage, and other visible indicators [*Hogan, B* 2007], and continuing an analysis trend that started with examinations of blogs and other websites .

SNS basically contain social circles when each one of those circles can relate to different aspect. For example in LinkedIn (LinkedIn is professional SNS that specialize work relation between work colleague) most likely that a regular user will have a work circle (he will have connection to people in his/her work place/domain) but s/he also can be at a different circle for example with friend from school or military service. The same can happen in Facebook: A user may have friends from different circles: school, university, work place, neighborhood, preferred music, food etc… in Google+ they even coded this feature as you can create or join a social circle – when each circle have a common topic. With the value in the social networks we can establish large collection user’s preference data, the effort of collecting data have been mention before by S D Rhodes1, D A Bowie2, K C Hergenrather[2003] in their research they concluded that using the web as empiric tools for behavioral science research will increase the tested population from local to global distribution. Using the web as resource for data mining is not new and used in many researches, in our effort we will try to found normally distribution population for secure or data misleading, for this effort we can search many type of populations at the socials networks and retaliate their information to characteristic relation graph.

## Related work

### Generic Semantic-based Framework

Fernández-Tobías et al. [2011] try to create an automated system that will recommend to user preference by to different domains in their approach the used graph for mapping the connection between the domain and analyzing the nodes relation in graph, this approach adopt the *Content-based recommendations* mention on paper by Adomavicius1 and Tuzhilin [2005] used past user data for create the recommendation, in their system they used DBpedia as the database source. DBpedia is a graph based database that contains values from Wikipedia. The main problem with that experience is that DBpedia is not updated Daily, in fact it has versions that come out once every ??? Since they can be depend on that database they created description framework built upon semantic networks, the problem with attitude it’s for obtaining data you to maintain and upgrade your frameworks.

### Link Data cloud

Figure 2: Fragment of the Linked Data cloud

Within the Semantic Web initiative, the Linked Data5 project aims at publishing structured datasets – usually described by standard metadata models such as RDF6 – on the Web, and setting (RDF) links between data items – usually called semantic entities – from different data sources. The adoption of Linked Data has thus led to the extension of the Web with a global data space connecting data from diverse domains such as people, companies, books, films, television, music, statistical and scientific data, and reviews [5]. This enables new types of applications. For instance, there are search engines that crawl Linked Data by following the links between data sources, and provide expressive query capabilities (see e.g. SPARQL7 RDF query language) over aggregated data, similar to how a relational database is queried today.

### Collecting data form social networks (SN)

Abdesslem et al. [2011] used social networks and mobile devices for collecting user characteristics and understand in behavior, in their research they used Facebook API for obtaining users data form Facebook while the activate questionnaire at user mobile devices, they also crossbreed their data with user location (they use GPS in the mobile device) for “understanding” where the user is and collect effective data and avoiding bad filtering. The main achievement they concluded guideline how collect data form social networks, they also concluded that the collecting process is depend on the purpose of the data.

### Social network analysis (SNA)

At the same way Fehmi ,lain and Tristan [2011] we also have to make sure that our subjects will be normally distributed and **not** belong to some type or circle in the SN but unlike fehmi we don’t want to collect random data from user , at the end we want to understand the nature of the connection , how we can identify interests ? , how we can seek for user friend with command interest? , how large the users sampling we need for understanding common phenomenon in SN ?

Boyd and Elision [2007] understand the advantage SN since the involved from variety of millions of users we can extract from them concluded information check for interests connection .Liu [2006] have mentions the user interests are basically can gather together and can be called “tastes” (examples: favorite music, books, film, etc), they argued that Friend connections are not the only network structure worth investigating. They examined the ways in which the performance of tastes constitutes an alternate network structure, which they call a ‘‘taste fabric.’’ The used system of classification derived from the practice and method of collaboratively creating and managing tags to annotate and categorize content called Folksonomies were used in the experiment to weave the taste fabric. Their idea was based on philosophical and sociological theories of taste and identity to weave a semantic fabric of taste. They mined 100,000 social network profiles, segmented them into interest categories and then normalized the folksonomies in the segments and mapped them into a formal ontology of identity and interest descriptors.

Shardanand and Maes [1995] represent users as vectors of (item, rating) pairs, and compute taste-similarity as statistical correlation between the two vectors, or alternatively as cosine similarity of the two vectors in n-dimensional item space.

## Summery

Recommendation systems use common people knowledge about items in an information domain to help people choose other items. Most recommendation systems determine the “decision” by internal data which gain by collaborative way, contend base or the hybrid of both attitudes. Recent studies shift their focus on cross domain recommendation, in most cases the cold start problem is increase on cross domain system cause by verity data of lack of system information.

Resent attempt [Liu, Maes & Davenport 2006 ] ,[Shardanand and Maes 1995] try to obtain user interest throw social networks and used their huge variety , collect data throw social networks have become native approach and many studies are base their result on social networks.

The used of graph in content-based recommendation system have been successfully try on Ricci Tobías and Kaminskas [2011] recommendation system the successfully the use of this approach even their recommendation based on two domains.

In this research we research recommender system with large scale graph base data which obtain throw social network analyze by graph database for offering solution to cold stat problem.

# Research Goals and Questions

The “Cold Start” problem is a well-known problem in user modeling and recommender system – how to bootstrap a user model in order to provide the user with a specific personalized service. Given the fact that a lot of personal information may be available in various sources, including social networks, but that this information may not exactly represent the user interests/needs/preferences in the target domain, a question is how can we use information available about a user in one domain for modeling the user in another domain, or “cross domain” modeling or recommendation.

The goal of the proposed research is to explore the possibility to use freely available information on a social network for cross-domain recommendation using a graph representation of a user model.

It is assumed that the wealth and variety of information that is available in social networks can be used for cross-domain user modeling when represented in a simple graph data structure and by applying generic graph search techniques.

The research will answer the following question:

*How we can use the social network of curators for mining the links between topics of interest for cross domain recommendation?*

# System and Methods

We will construct a research tool called **TraitsFinder** that will allow us to collect user’s information and extract the data [[5]](#footnote-5)from a specific social network – Pinterest and build a graph that will represent the connection between traits, the research will work by two steps:

* **TraitsFinder**  will crawl pinterest social networks and collect user’s information in our servers, the outcome of this step is users folder with user information save in XML files.
* **TraitsFinder** will create graph base users crawled data – this step can run in offline mode or online (user information are update immediately after saved) the outcome of this step update neo4j graph.

## Methods

The research is a design research (Hevner[2004]). As such, an experimental tool will be built. It will be used for representing user models over a graph and graph based techniques will be used for cross domain recommendation generation. Personal information will be collected from web Social network called Pinterset By Our research tool Called **TraitsFinder** The information will then be uploaded to the graph using also **TraitsFinder.**

### Data Source

Since we are interested in creating affective database with reflect all type of population the collecting must be breath spreading and must have direct links between some of individual , it’s also must contain many domains with some kind of link between them ,the source must also be dynamic, scalability and can be change .

#### Social Network as a source

We found that the social networks (SN) can be effective sources for establish database, the main key in social networks is to shard the individual to the common population. That value can attribute to our goal, each social network are depend user data, in general at SN user are upload their data to the SN, they are deiced how to shard to the common population, to your link friend only or save as private to your used also can benefit the free cataloging – when users are update their data they choose how to call and catalog it form this act we enumerate the SN. as follow : when each user is update some data we can refer it as new object in the database the name of that field is the name that the user is catalog it , for example : Let us suppose that some random user at Facebook with cooking hobby , some day he decided to bake special birthday cake for his person in his family, when he finish making cake he decided to photograph the cake and upload the picture and the recipe to Facebook , he also label “ homemade Birthday cake for <person >” from that act we can examine the this picture as object , we can understand that some person have birthday , we can also understand that this user love to bake ,and it’s person love cakes, when different users are press like for that cake we can also assume different conclusion for example :we can assume that users how press like or send message to the cake are probably people that love cakes or at least love the shape of that cake , people that will press like on the recipe most likely people that love to bake to. This case is one many cases that happen hundred or maybe even thousand time at each day, the conclusion that we mention before are easy for human to understand, but we need to save those conclusion and save the result in database.

Additional aspect that tribute for our effort it’s SN’s are daily even hourly updated, danah m boyd and nicole elision [2007] already mention that SNS combine form millions of users, we can construct automated systems that can give a “big picture”- large scale analysis that can daily update and synchronize with SNS , also since SNS are basically web system we can access to their data from anywhere ,since we base on data that exist in the computerize system we can create system that can used the SNS interface for extracting relevant data.

Abdesslem, Parris, and Henderson [2011] research and concluded the use of SN for collecting data, they speared the collecting to two sections – collect user social behavior in SN and collect user characteristics. from their aspect when we used SN we collect not only is preference and characteristics we can actually build social profile from the user data, in their research they collect data from SN , but in their twitch they not collect the data only from SN the also use mobile devices and for crossbreed their collected data with user location , by this way they success to “understand ” user act’s ,behavior and desire . they successfully collect data , but the main benefit that we can concluded is the importance of random sampling , Fehmi ,lain and tristan make excrements with that interacts with user and used Facebook as the SN , for creating random sampling they select users for his friend , their application assist with “all friend” user and extract new user – with this approach they achieve random sampling and keep their subject normally distributed.

### Selecting pinterest Social network s

Since we want to rely on social networks (SN) as source for our graph database the logical conclusion is to use Facebook, at least for now Facebook is the biggest SN exist, it’s have more them one billion members, it’s update very frequently – today most of their users user updates data in the mobile devices, since it’s popularity it’s cover almost any type of population at any age.

massive investigation with Facebook API we discover that Facebook not allow to normal small developers to collect data from the entire Facebook users (if this ability was exist and publish most likely Facebook have been busy with defending herself agents legal claim) instead normal user need to access to some kind of application (it’s can be game, puzzle, quiz interview or any application we want) once the user have been access to the application we can access user profile using Facebook query language (FQL) .

we decide to abandoned Facebook for several problems we encounter:

* **Sampling problem** - in Facebook we can’t samples random users , instant we need to create some kind of “bait” for calling our users ,with this action our effort will be leaning for particular population (population that was interest in our application ).
* **Circle problem -** since we not sampling random users we can’t access to users friends , we can only publish in our participant users wall – this action will not help to expand our sampling , intend it’s will create circle of users that will use the application.
* **Semantic problem –**in Facebook users upload picture , updated status ,join groups , check in places etc… form all this action it’s hard to understand user characteristic needed advance sematic parser for analysis user preference and Traits , for example if user upload birth cake with no explanation what we can understand form this picture , that he have birthday party, he love to bake cake or he just love cake.
* **Legal issues** – if we use Facebook we will need to ask or mention to the user this is academic experiment - this can also harm our user sampling.

Since we can random sampling in Facebook we will use Pinterset SN for collecting users information, Pinterest is curator photograph sharing social network.

Pinterest not only simple and have specific attribution we need – in Pinteres user interested **catalogue** to subjects, we also get the connection between users – when user upload photo and catalogue it ,any other user that will pin this picture we can understand and analyze is connection to that picture , we have also very big advantage in Pinteres the subjects are basically our characteristic that we seek . in additional we don’t have to become entangled with random sampling issue – when can just sample all the users .

### Crawling Pinterest

Unfortunately pinterest doesn’t have API, for crawl the website we contacts web application that parsing HTML web page and extract the data from web site: The data in Pinterest exists in site in chronologic way at the main page there pictures with comment when you press at one picture you access to user that upload the picture under is relevant subject. The hierarchy in pinterest is simple and constant is work as follow way:

Figure 3 : Pinterest hierarchy

Since hierarchy is simple and fixed we can used the web crawler for establish our data.

### Crawling process

The crawling process is recursive process when the crawler algorithm works as follow:

Go to pictures group G

Crawl(**G**)

{

If **G** is empty exit

Else

{

Find pictures **P** from **G**

Save Comment **C** from **P** under **P**

Foreach user **X** in **C**

{

Add user **X** to group **U**

Foreach subject **S** in **X**

**{**

Save subject **S**i under user **X**

Save all items (it) under **S**i

**}**

}

Crawl(X)

Crawl(U)

}

}

### Data collecting:

The data we are going to extract will be saved in files for each user, subject, picture and comments. The crawler process will convert the HTML pages standard XML files under folder hierarchy (see figure 5) at same way we save the subjects and pictures page files.

Second stage we parse each xml file for creating the graph.

### Graph creation

The crawling process will create us hierarchy of folders and xml files. From those files we can parse the content and retrieve information a graph .with the ability to represent the native ontology of user curator subjects.

#### Main general graph

The main graph will represent all type of connections inside Pinterest website, each object will be document in our main graph.

The main graph is bend to Pinterest hierarchy will marked as graph G when each node (V) will represent object in Pinterest hierarchy (users, subjects , items ) therefor V={users ,subjects ,items} , the edges (E) are the connection is represent belonging at the hierarchy when EE ⊆ {u,v∈V } for example : let’s presume user\_x ∈ V when he owned subjects like pizza ,animals (animals ∈ V, pizza ∈ V) and we also have user\_y that also love animals and pizza we will create edge between user and is subjects (user\_x,pizza)∈E and (user\_x,animals)∈E and relation between user\_y ((user\_x,pizza)∈E & (user\_x,animals)∈E)

User\_z love cars(cars ∈ V) and pizza also ,therefore (user\_z,pizza) ∈E & (user\_z,cars) ∈E.

pizza subject owned items likes tuna\_pizza ,olives\_pizza and mushrooms\_pizza (tuna\_pizza ∈ V ,olives\_pizza ∈ V and mushrooms\_pizza ∈ V) will create edges between subject to item (tuna\_pizza ,pizza) ∈ E , (olives\_pizza, pizza) ∈ E and (mushrooms\_pizza ,pizza ) ∈ E .

With this graph we can analysis the connection between each object in pinterest site



Figure 4: graphical view main graph

#### Interests connection graph

our interests graph is undirected graph G that will extract from the main graph ,each node character will mark as (V,E) when V is our node group (interest) and E represent the edge when EE ⊆{u,v∈V} the edge E is represent connection between the interests, the connection is establish when user X have both interests , the weight of the edge is the number of users obtain the same two interests. In our example the interest graph have only the interest nodes: {animals, pizza, cars} = V the edges are represent the native of the interests connection when (animals, pizza)∈E & (pizza ,cars)∈E when ι(animals,pizza)=2 and ι (Cars,Pizza)=1 the weight animals-pizza edge is 2 since user\_x and user\_y are both like animals and pizza.



Figure 5: graphical view Interests connection graph

Eventually we will have weighted undirected graph that will represent the our graph database when the number of nodes (vertex) is equal to the number of characters |V|=numof(characters) , the |E| represent the number of connections between characters

## Tools

For the purpose of the planned research to use the following tools:

* **TraitsFinder** - we will construct research tool called TraitsFinder that will collect data from Pinterest website , TraitsFinder will Crawled Pinterest website and will save data as local xml files , TraitsFinder have also the ability to upload the information to grap database .

**TaritsFinder** is multithread application cross OS implemented in java.

* **Neo4J[[6]](#footnote-6)** – is a high-performance, NOSQL graph database with all the features of a mature and robust database. **TraitFinder** will upload user’s interests to Neo4j graph.
* **Gephi[[7]](#footnote-7)** – Gephi is an interactive visualization and exploration platform for all kinds of networks and complex systems, dynamic and hierarchical graphs. Gephi will be used to survey and analysis neo4j graph .

## 4.3 Evaluation

The proposed research will be evaluated by collecting publicly available data from social networks regarding users’ preferable items and using it to train a graph based recommendation engine. Approximately 100,000 profiles are to be crawled, initial test crawls show an average of X subjects of interests per user, with Y items in average listed under each interest album.

The cross domain recommendations will be evaluated using a 10 fold validation as described in Kohavi [1995]. Of the collected data, 80% will be used to train the recommendation engine and 10% for its evaluations. The 10% selected for evaluation will be changed during the evaluation process to include different features of the data. Once the evaluation iterations will result with a minimal error (matching recommendations considered as ‘hits’) single test iteration will provide a final result.

As part of the evaluation it is also intended to analyze how the size of the dataset/size affects recommendations results. This will be done by taking different subsets of the available data and measuring the changes in recommendations quality.

Since our system is cross-domain recommendation system the variety of interest is basically infinite R.Kohavi [1995] have been investigating using cross-validation and bootstrap for analyzing bias learning, he concluded the more the K-fold is bigger it’s reduce the variance while increase the bias, since our graph will be high interests variance we will need to equalize graph to the K value in the K-fold cross validation – for better evaluation we find the K value by measuring the all graph interest nodes.

For evaluate this graph we be cross validation technique, we will runs cross validation runs as described in the following table (in this example k-fold test when k=1000):

|  |  |  |
| --- | --- | --- |
| Train size (creating graph based on X users) | Number of folds tested user check | Accuracy needed |
| 1000 | 1 |  |
| 10000 | 10 |  |
| 100000 | 100 |  |

The tested fold user will checked by checking recommitting hits, for each tested user we scan is interests and valid our recommitting algorithm on is only 30-40 % of is interests , the recommitting that **TraitFiner** will return will cross examined with is actually traits. Using cross validation is common technique especially for learning system, since our recommendation system is learning with graph base learning algorithm , we will applied this technique on our system.

# 5. Timetable

Each of the phases discusses the primary focus of each time period:

Phase I - Literature Survey and Focus.

Phase II - establish first stage of TraitsFinder – the web crawler

Phase III - collecting users traits (exit criteria: at least 100,000 users)

Phase IV - broadening TraitsFinder - add graph builder.

Phase V - improve TraitFinder: add automatic graph analysis GUI and RI

Phase VI - Graph analysis, evaluation and algorithm establish

Phase VII - Writing Thesis.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Phase | Spring 2012 | Summer  2012 | Fall  2013 | Spring  2013 | Summer  2013 |
| I |  |  |  |  |  |
| II |  |  |  |  |  |
| III |  |  |  |  |  |
| IV |  |  |  |  |  |
| V |  |  |  |  |  |
| VI |  |  |  |  |  |
| VII |  |  |  |  |  |

# 6 Research Contributions

The proposed research main contribution to the field of user modeling will be a cross domain recommender algorithm. The algorithm will apply graph analysis methods for the purpose of recommendation generation and will be based on data (interests) extracted publicly available data.

A secondary contribution will be estimation for dataset scales from which valid recommendations become effective.

# 7 References

1. Fehmi Ben Abdesslem, Iain Parris, and Tristan Henderson (2011)- “Reliable Online Social Network Data Collection”
2. Ignacio Fernández-Tobías , Marius Kaminskas , Iván Cantador and Francesco Ricci (2011) , “ A Generic Semantic-based Framework for Cross-domain recommendation”
3. Rubi Boim (2011)(Under the supervision of Prof. Tova Milo) Tel-Aviv University “Methods for Boosting Recommender Systems”
4. danah m. boyd1, Nicole B. Ellison (2007) - “Social Network Sites: Definition, History, and Scholarship “
5. Shlomo Berkovsky , Tsvi Kuflik , Francesco Ricci (2007) “Mediation of user models for enhanced personalization in recommender systems”
6. peter brusilovsky and eva millan (2007) – “The Adaptive web : user models for adaptive hypermedia and adaptive Educational systems”
7. B.Hogan (2007). “Analyzing social networks via the Internet. In N.Fielding, R.Lee, & G.Blank (Eds.), Sage Handbook of Online
8. Hend S. Al-Khalifa and Hugh C. Davis (2006) “FolksAnnotation: A Semantic Metadata Tool for Annotating Learning Resources Using Folksonomies and Domain Ontologies”
9. Liu, H., Maes, P., & Davenport, G. (2006). “Unraveling the taste fabric of social networks.” International Journal on Semantic Web and Information Systems, 2(1), 42–71
10. Adomavicius and Tuzhilin, A. (2005) - “Towards the Next Generation of Recommender Systems:A Survey of the State-of-the-Art and Possible Extensions”
11. Alan R. Hevner, Salvatore T. March, Jinsoo Park and Sudha Ram (2004) “ Design Science in Information Systems Research Design Science in Information Systems Research”
12. Jeffrey Dean and Sanjay Ghemawat “MapReduce: Supplied Data Processing on Large Clusters” *Google, Inc (2004)*.
13. S D Rhodes1, D A Bowie2, K C Hergenrather3 (2003) –“ Collecting behavioural data using the World Wide Web: considerations for researchers”
14. Robin Burke (2002) “Hybrid Recommender Systems:Survey and Experiments”
15. Badrul Sarwar, George Karypis, Joseph Konstan, and John Riedl (2000) “Application of Dimensionality Reduction in Recommender System – A Case Study”
16. Schafer, J. B., Konstan, J. and Riedl, J. 1999, “Recommender Systems in E-Commerce”.
17. H.Guo. “Soap:live recommendation through social agent”. In fifth DELOS workshop on filtering and Collaborative Filtering , Budapest (1997)
18. Fayyad, U. M., Piatetsky-Shapiro, G., Smyth, P.,and Uthurusamy, R., Eds. (1996) - “Advances in Knowledge Discovery and Data Mining”
19. Hill, W., L. Stead, M. Rosenstein, and G. Furnas(1995). “Recommending and evaluating choices in a virtual community of use”.
20. Breese, J. S., Heckerman, D. and Kadie, C (1998), “Empirical analysis of predictive algorithms for collaborative filtering”
21. Resnick, P., N. Iakovou, M. Sushak, P. Bergstrom, and J. Riedl. GroupLens 1994: “An open architecture for collaborative filtering of netnews”. In Proceedings of the 1994 Computer Supported Cooperative Work Conference.
22. U. Shardanand and P. Maes. (1995).” Social information filtering: Algorithms for automating `word of mouth'”.
23. Jennings, A. and Higuchi, H.(1993) ”A User Model Neural Network for a Personal News Service”.
24. Ron Kohavi (1995)“A study of Cross-validation and bootstrap” for accuracy Estimation and model selection” ,Stanford university.
25. Condliff, M. K., Lewis, D. D., Madigan, D. and Posse, C.(1999), “Bayesian Mixed-Effects Models for Recommender Systems”
26. Foltz, P. W.(1990) “Using Latent Semantic Indexing for Information Filtering”
27. Thomas H. Cormen, Charles E. Leiserson, Ronald L. Rives “Introduction to Algorithms” chapter VI Graph Algorithms (1990)

1. <http://www.pandora.com> [↑](#footnote-ref-1)
2. <http://www.google.com> [↑](#footnote-ref-2)
3. <http://www.youtube.com> [↑](#footnote-ref-3)
4. <http://www.amazon.com> [↑](#footnote-ref-4)
5. www.Pinterest.com [↑](#footnote-ref-5)
6. http://neo4j.org/ [↑](#footnote-ref-6)
7. https://gephi.org/ [↑](#footnote-ref-7)