**Traits Finder:**

**For better characteristic links understanding**

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

**Oz Shapira**

Contents

[1. Introduction 3](#_Toc341699161)

[2. Background and Related works 5](#_Toc341699163)

[2.1. Background 5](#_Toc341699164)

[2.1.1. Recommender systems 5](#_Toc341699165)

[2.1.2. Graphs as data structures 5](#_Toc341699166)

[2.1.3. Social networks 6](#_Toc341699167)

[2.2. Related work 7](#_Toc341699168)

[2.2.1. Generic Semantic-based Framework 7](#_Toc341699169)

[2.2.2. Link Data cloud 7](#_Toc341699170)

[2.2.3. Collecting data by SNS 7](#_Toc341699171)

[2.3. Seeking for normally distribution source 8](#_Toc341699172)

[2.3.1. Social Network as a source 8](#_Toc341699173)

[2.4. Connection Analysis 9](#_Toc341699174)

[2.5. Summery 9](#_Toc341699175)

[3. Research Goals and Questions 9](#_Toc341699176)

[4. System and Methods 10](#_Toc341699177)

[4.1. Methods 10](#_Toc341699180)

[4.1.1. Selecting pinterest Social network s 10](#_Toc341699181)

[4.1.2. Crawling Pinterest 11](#_Toc341699182)

[4.1.3. Crawling process 12](#_Toc341699183)

[4.1.4. Data collecting: 12](#_Toc341699186)

[4.1.5. Graph creation 13](#_Toc341699187)

[4.2. Tools 14](#_Toc341699189)

[4.2 Evaluation 14](#_Toc341699190)

[6 Research Contributions 15](#_Toc341699191)

[7 References 15](#_Toc341699192)

# 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 became an important research area since the appearance of the first papers on collaborative filtering since the mid-1990s [45, 86, 97]. There has been much work done both in the industry and academia on developing new approaches to recommender systems over the last decade. The interest in this area still remains high because it constitutes a problem rich research area and because of the abundance of practical applications that help users to deal with information overloads and provide personalized recommendations, content and services to them. Examples of such applications include recommending books, CDs and other products at Amazon.com, movies by MovieLens , and news at VERSIFI Technologies (formerlyAdaptiveInfo.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 (see figure 1).

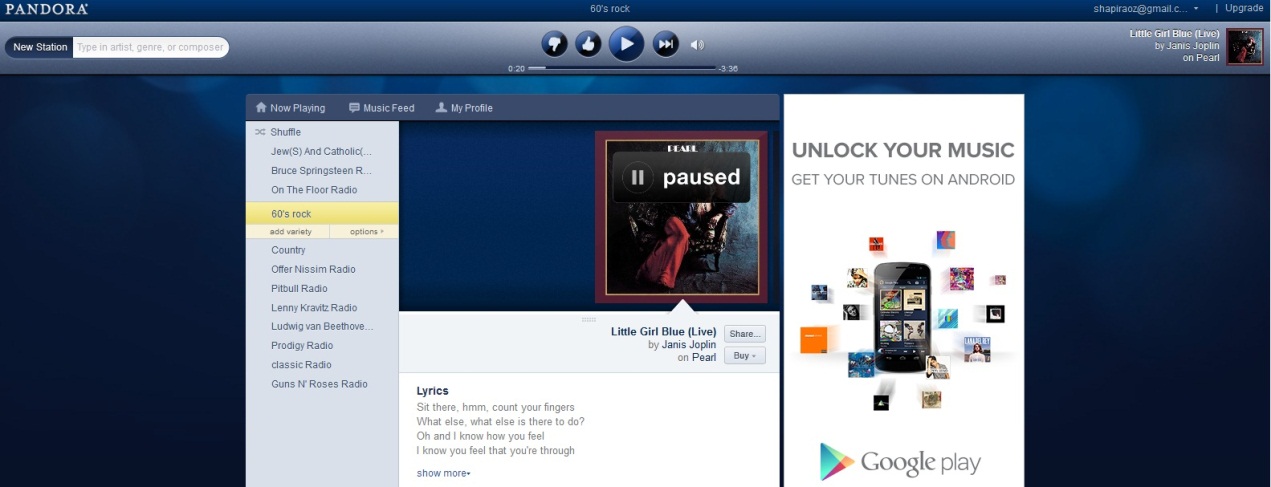


Figure 1: Pandora as recommender system

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, than 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/ anew city/ restaurant etc. The lack for 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 (H.Guo [1997]).

Nowadays, At the World Wide Web many sites contain user’s profile which reflects user characters. As we surf and visit websites we leave identifiable digital “fingerprints” not to mention explicit definition of interests and preferences. In recent years, social networks became a major source for personal information (ref…). For example in Facebook the user insert is interests while in Pandora (Pandora its music recommended radio website system with broadcast music according to user flavor) try to find user flavor according his user music analysis.

This freely available personal information, scattered over various online sources such as 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, this may not be relevant to the 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.

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 information sources.

We plan to explore, demonstrate and evaluate the ability to use graph based representation of user modeling to help solving the cross-domain user modeling challenge.

In most cases user model systems are trying to obtain user knowledge, learn their styles and individual traits, research on this mattes have been concluded by p.brusilovsky & millan [2007]. the more the system is open, generic and shareable its increase the possibility for more accurate user profile, the main problem in user modeling is attempting to modeling new user when we have very little data on him, this research outcome can assist on this problem by identify individual traits using collection of many user traits (at the first stage ) and find the connection between those traits and applied ontology for new users .

.

# Background and Related works

## Background

### Recommender systems

Recommender systems have the interactive recommendation abilities, Recommender system apply 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 (M. Sarwar, Karypis,A. Konstan,T. Riedl)[2000] .there is many type of recommendation systems and in each one of them have different approach for recommitting (vs. Pandora , Google search ,YouTube ,amazon… ) There is tree type of recommendation system

* **Content-Based**, - where recommendations are based on semantic properties (preferences) of the items (users) .
* **Collaborative-Based** - where the recommendations are based on previous ratings of (similar) users to (similar) items, with the assumption that users who agreed in the past on item ratings are likely to agree again in the future.
* **Hybrid systems –** thecombination of Content-Based and Collaborative-Base based the recommendation

but most of those system are need to obtain user profile for their recommendation ,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. in this paper we will not analysis or extend at this matter , instead we will focused on collecting stage and discuss on ways to obtaining their internal database also we offer different solution for obtaining the database and analysis in different approach by using graph’s .

### 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 ).

The Implementation graph as data structures is common and has many aspects, the main advantage of deploying graph in computerize system is the abilities to analyze in automatic format. used graph is to reduce analysis of source-target implantation based on web linked hierarchy (MapReduce problem ) (Dean & Ghemawat [2004] ) .

Few years ago new wave of NOSQL have breach the concept of none using traditional database with standard SQL transaction .

### Social networks

Social networks have been with us since 1997 (the first one was sixDegrees.com) , social networks site (SNS) have successfully change worldwide communication they gave personal user the ability to reach any user in the world , SNS site 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 cause to SNS the basic ability to connect between separate type of population using SNS users ,(danah m boyd and nicole elision [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 from the company, enabling network analysis researchers to explore large-scale patterns of friending, usage, and other visible indicators ([Hogan, in press](http://onlinelibrary.wiley.com/doi/10.1111/j.1083-6101.2007.00393.x/full#b45)), and continuing an analysis trend that started with examinations of blogs and other websites .

SNS are basically contain social circle when each one of those circle can related to different aspect for example in Linkedin (Linkedin is professional SNS that specialize work relation between work colleague) most likely to normal user will have work circle (he will have connection to people in is work) but he also can be at different circle for example is friend from school and is army service. At same way it’s can happen in Facebook user have friend from different circle : school, university, work place, neighborhood, music he love , food etc… in Google+ they even coded this feature as you can create or join to circle – when each circle have the 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 Hergenrather3 (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

Ignacio Fernández-Tobías , Marius Kaminskas , Iván Cantador and Francesco Ricci public article (A Gene6ric Semantic-based Framework for Cross-domain Recommendation 2011) try to create automated system that will recommend to user preference by to different domain 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 Gediminas Adomavicius1 and Alexander Tuzhilin (Towards the Next Generation of Recommender Systems:A Survey of the State-of-the-Art and Possible Extensions -2005) used past user data for create the recommendation ,in their system they used DBpedia as the database source. BBpedia is graph base database that obtain is values from Wikipedia , the main problem with that experience is DBpedia is not updated Daily , in fact it have version that come out once in some time ,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 by SNS

Fehmi Ben Abdesslem, Iain Parris, and Tristan Henderson (2011) used the social networks and mobile devices for collecting user characteristic 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” here user and collect effective data and avoiding bad filtering.

In this paper we will try to obtain database by using the SNS system the biggest question is since SNS are combine for many type of the population when each item is basically related to some circle how we sample subject in random way without taking our samples from the same circle in this paper will try to answer this question.

## Seeking for normally distribution source

Since we interested of 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/upload some data/status/new item , we can refer it as new object (new field) 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 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 (it depend on site access – we will enlargement at section 2.4) 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. Using automated system like we describe will allow collecting anytime anywhere data.

Fehmi Ben Abdesslem, Iain Parris, and Tristan 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.

## Connection Analysis

At the same way Fehmi ,lain and Tristan 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 between the user and their preference ,therefore we will have to collect that data in some way that allow us to understand the native of the connection between each preference and characteristics for building characteristics connection graph.

In the birthday cake example that we mention in section <> user can connect each other – this in one of the strongest abilities exist in SN , the fact that users can add link to each other (via friend in Facebook) is create a grid of users with connection to each other – we can look at this grid as graph and use it for creating characteristics connection graph example , in this research our main purples is to collect the data , but also to arrange in in graph , the characteristics graph have to build dynamically and update form each user at the SN , therefore the process of collecting the data is automatically we be depended on the relation with SN users .

## Summery

<Need to complete>

# 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 data and extract the data to mathematical graph that will represent the connection between traits, this tool (Traits Finder) will crawled SN (pinterest) and collect user’s traits in our servers.

At the second stage is graph building, in this stage we will add to our tool graph building abilities with basic analysis.

## Methods

The research is a design research (Hevner….). 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.**

### 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 used Facebook we will need to ask or mention to the user this is academic experiment - this act 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 is exist 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 #: 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 save 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 and cars (animals ∈ V, pizza ∈ V, cars ∈ V) we will create edge between user and is subjects (user\_x,pizza) ∈ E ,(user\_x,animals) ∈ E and (user\_x,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 pinterset site

Graphical view main general view

#### Characteristics connection graph

our Traits graph is undirected graph G that will extract form the main graph ,each node character will mark as (V,E) when V is our node group (characters) and E represent the edge when EE ⊆{u,v∈V} the edge E is represent connection between the characters , the connection is establish when user X have both characteristics for example if user love cakes and movie in our general graph will be two node v\_movie ∈ V & v\_cakes ∈ V therefore will create edge between them (v\_movies, u\_cakes ) ∈ E .

Graphical view of connection Nodes 1

If we will find another user that love also movies and cakes as well we just simply add weight to the edge when each new user that we will find with this connection we just simply increases the weight by one.

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** – TraitFinder will use Neo4j graph database for upload the collected data.
* **Gephi** – will be used to survey and analysis graph .

## 4.2 Evaluation

The performance of the system will be assessed as we will create detailed graph that will cover more the 100000 users the graph have to represent the accurate relation between traits and will used for us as map for understanding strength of each relation . by this graph we will predict create valid recommendation

For evaluate this graph we will run simulation that will used known user and tested system hitting recommendation rate.

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

# 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”
   * 1. danah m. boyd1,•, Nicole B. Ellison (2007) - “Social Network Sites: Definition, History, and Scholarship “
3. peter brusilovsky and eva millan (2007) – “The Adaptive web : user models for adaptive hypermedia and adaptive Educational systems”
4. Adomavicius and Tuzhilin, A. (2005) - “Towards the Next Generation of Recommender Systems:A Survey of the State-of-the-Art and Possible Extensions”
5. Jeffrey Dean and Sanjay Ghemawat “MapReduce: Simplied Data Processing on Large Clusters” *Google, Inc (2004)*.
6. S D Rhodes1, D A Bowie2, K C Hergenrather3 (2003) –“ Collecting behavioural data using the World Wide Web: considerations for researchers”
7. Hogan, B. (in press). “Analyzing social networks via the Internet. In N.Fielding, R.Lee, & G.Blank (Eds.), Sage Handbook of Online
8. Robin Burke (2002) “Hybrid Recommender Systems:Survey and Experiments”
9. Badrul Sarwar, George Karypis, Joseph Konstan, and John Riedl (2000) “Application of Dimensionality Reduction in Recommender System – A Case Study”
10. H.Guo. “Soap:live recommendation through social agent”. In fifth DELOS workshop on filtering and Collaborative Filtering , Budapest (1997)
11. Fayyad, U. M., Piatetsky-Shapiro, G., Smyth, P.,and Uthurusamy, R., Eds. (1996) - “Advances in Knowledge Discovery and Data Mining”
12. Thomas H. Cormen, Charles E. Leiserson, Ronald L. Rives “Introduction to Algorithms” chapter VI Graph Algorithms (1990)
13. Rubi Boim (Under the supervision of Prof. Tova Milo) Tel-Aviv University “Methods for Boosting Recommender Systems”