**By Oz Shapira**

**Abstract**

.

**1 Introduction**

What if someone will tell us that he love pizza ,James bond movies and he also love to beach jogging and we can replay to him that he probably love to bake, hike and flower interweaving. This research will try to constant a valid recommender system with the ability to answer on those kinds of questions in high absoluteness.

Recommender systems are part of user modeling systems with Recommendation abilities, those system “knows” according to past users knowledge how to make recommendation according to his flavor

At the World Wide Web many sites contain user’s profile which reflects user characters.

In some of the site this data have been add by the user itself while in different sites are analyze user profile and added automatically, 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.

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 (the adaptive web 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 .

A type of systems that analyzing the user data reference and recommend to user on new user preference called recommendation systems, 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 .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 two 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 ratingsare likely to agree again in the future.

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.

In most knowing recommender systems today are specializing on one domain (for example Pandora as music recommender, Netflix offers predictions of movies) and simply called domain recommendations system , the ability to focused on single domain give to those systems the focused they need for correct proposals. Usually those system are using single dimensional database or many databases with common fields therefore they most of those systems have been successfully develop algorithm for collecting the data.

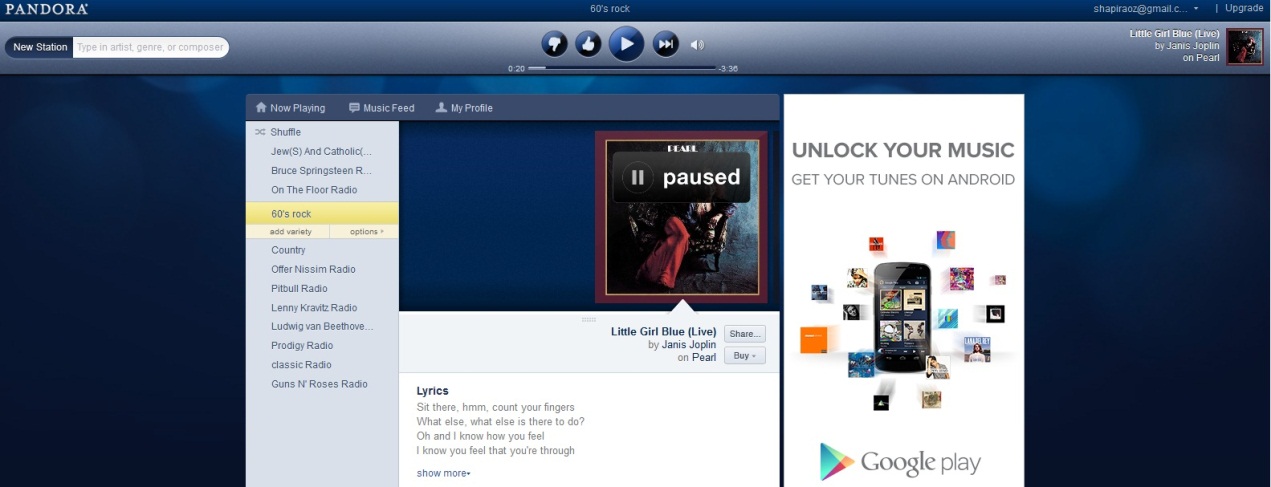


Figure 1: Pandora as recommender system

the effort of collecting data have been mention before by [S D Rhodes](http://jech.bmj.com/search?author1=S+D+Rhodes&sortspec=date&submit=Submit)[1](http://jech.bmj.com/content/57/1/68.short#aff-1), [D A Bowie](http://jech.bmj.com/search?author1=D+A+Bowie&sortspec=date&submit=Submit)[2](http://jech.bmj.com/content/57/1/68.short#aff-2), [K C Hergenrather](http://jech.bmj.com/search?author1=K+C+Hergenrather&sortspec=date&submit=Submit)[3](http://jech.bmj.com/content/57/1/68.short#aff-3) (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 find many type of populations at the socials networks.

Social networks have been with us since 1997 (the first one was sixDegreee.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 Linkin (Linkin 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 preference data

In section 1.1 we will show related works, section 2.1 will show our proposal to used social network s, in sections 2.2, 2.3 and 2.4 we will explain and show effort and conclusion.

**1.1 Related works**

**1.1.1 Generic Semantic-based Framework**

Ignacio Fernández-Tobías , Marius Kaminskas , Iván Cantador and Francesco Ricci public article (A Generic 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.

**1.1.2 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.

**1.1.3 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.

**2 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 .

**2.1 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.

**2.2 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 .

**2.3 Establish characteristics connection graph**

The collecting and the building database graph is process that will need be automated, cross domain oriented.

our characteristics connection graph is undirected graph G When each node character will mark as (V,E) when V is our node group (characters) and E represent the arc when E the Arc 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 therefore will create arc between them .

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 arc 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 ,but is this graph the is no way to know how many users are combine it - this data will save separately.

**2.4 Building our database**

**2.4.1 First attempt: Facebook as SN source**

Since we want to rely on SN as source for our graph data base the logical conclusion is to use Facebook, at least for now Facebook is the biggest SN 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.

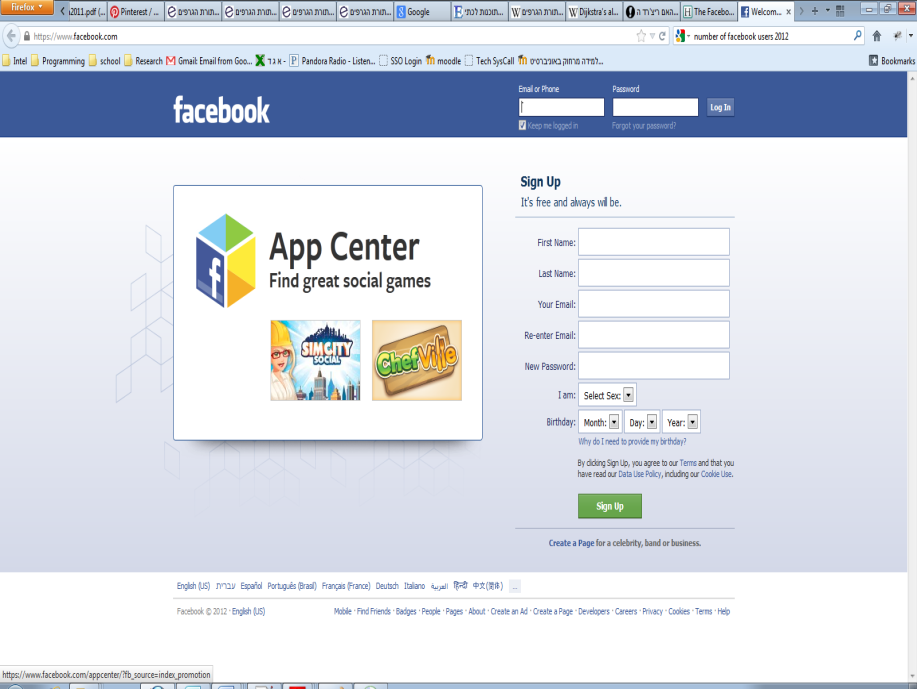


Figure 3 Facebook -still the biggest SN

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 – Facebook have query language (FQL) with this ability you can access to any user object (just by using FQL) .

As far we gather to Facebook solution we encounter to major problem:

* **the Random problems -** despite fehmi , ben and lain success we found their attitude can be problematic , unlike them we can’t collect from few user for achieving results ,we need those’d of user sampling. Experiment at small scale can tribute to our effort .
* **The Circle problem –** SN composite of social circles, at Facebook you can consider yourself to some circles (like Google+ ) , so how build quiz online questionnaire that will cover all type of users ? - when we build some kind of application we risk that our application will used only by some of the users type – for example lets assuming we will build Facebook game – how can we build a game that all type of user for all common population will play it ? , the same problem we found with any Facebook app we can build
* **Semantic problem -** since we seek for user characteristic sometime it’s hard to “guess” in Facebook some of the data object are fixed in Facebook for example status (single , married, works , school etc…) but some of the characteristic will need to analyze
* **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.

**2.4.2 Second and final attempt: Pinterest**

Since application at Facebook can’t work for our collection goal, we concluded that we need to “push” collector that will collect from SNS (not to wait and “pull” user data) therefor we need to used crawling technique that will search at SNS for user characteristic .

A **Web crawler** is a computer program that browses the World Wide Web in a methodical, automated manner or in an orderly fashion.

Crawling Facebook was our first attempt, since we have Facebook API the task look easy, but Facebook not allow to get user data without user action, the effort was to construct Facebook App when at the lower level at this app will be the web crawler.

Since Facebook was not a good option we continue to seek for new solution when we our search was focus in open SNS with user characteristic.

The answer was at SNS called Pinterest - pinterest is photograph sharing SN (very similar to histogram) user upload there photos catalogue them to subject, write comments to photos , also other user can comments on photos , add specific photo to their subjects , when user add specific photo he pin the photo to his profile.

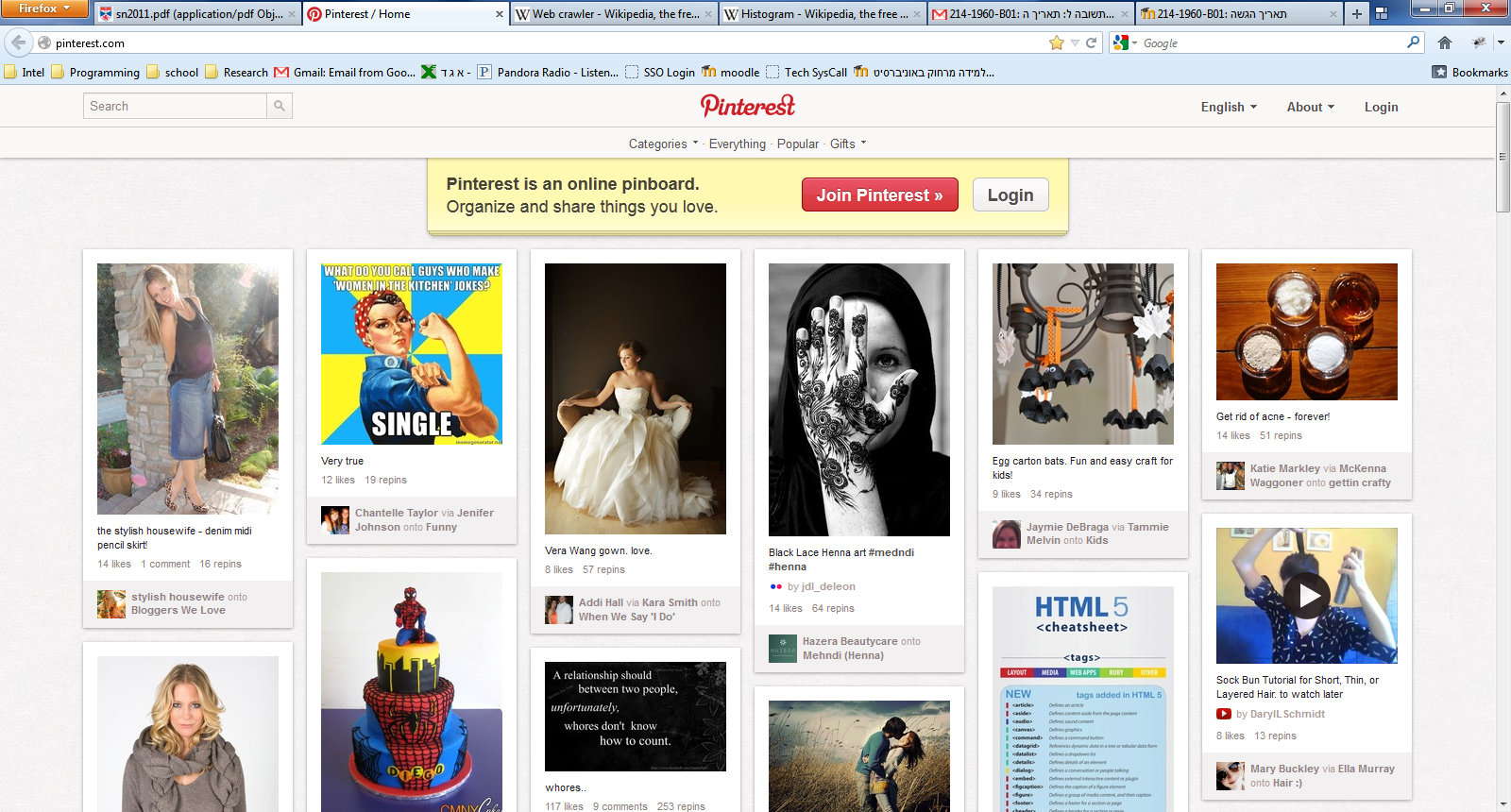


Figure 4: Pinterest as our source

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 .

**2.4.2.1 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 their 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 5: Pinterest hierarchy

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

**2.4.2.2 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
* Add user X to group U
* Save subject x to under user X
* Crawl(X)
* Crawl(U)

}

}

**2.4.2.3 Saving the data:**

The data we are going to extract is save in files, database and graph database for each user, subject ,picture and comments . The crawler convert the HTML pages to that encounter in XML file under folder hierarchy (see figure 5) at same way we save the subjects and pictures page files. In the database we save them by using SQL server when the tables are users ,subjects , pictures , comments –at each table there is foreign key (FK)to his parent table (according to the site hierarchy : for example in subjects table there is FK of users)

To save the data in graph database the saving technique will accord section 2.3

**2.4.2.4 Implantation:**

The system been write in Java (SE 1.6 version), the chosen SQL server is MySql , and the graph database is newo4j (open source graph database base on java).

**2.5 Result**

So far our crawler is having the ability to parse and collect until the picture (item) depth, we successfully download user’s data and extract there subjects – the process is extremely efficient:

The rating of adding new user is 50 users per sec, when at each user it’s take less them 1 sec to parse all is subjects (it’s depend in the number of subject the rate is also ~50 subject for sec) so if we can estimate that we can at 50 sec collect 2500 characteristic from 50 user, for now the data is saved only at XML format but future work the system will support tabled and graph database.

**3 Conclusions:**

In this document we show summarize and describe our effort to collecting user characteristic and establish database to Collaborative-Based recommender , we describe the searching way , explain the used in social networks (SN) as live source ,we explain our way to extract the characteristic and save data with the ability to easily analyzing the connection between them.

**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 ecommendation”
3.  danah m. boyd1,, Nicole B. Ellison (2007) - “Social Network Sites: Definition, History, and Scholarship “
4. peter brusilovsky and eva millan (2007) – “The Adaptive web : user models for adaptive hypermedia and adaptive Educational systems”
5. Adomavicius1 and Alexander Tuzhilin (2005) - “Towards the Next Generation of Recommender Systems:A Survey of the State-of-the-Art and Possible Extensions”
6. [S D Rhodes](http://jech.bmj.com/search?author1=S+D+Rhodes&sortspec=date&submit=Submit)[1](http://jech.bmj.com/content/57/1/68.short#aff-1), [D A Bowie](http://jech.bmj.com/search?author1=D+A+Bowie&sortspec=date&submit=Submit)[2](http://jech.bmj.com/content/57/1/68.short#aff-2), [K C Hergenrather](http://jech.bmj.com/search?author1=K+C+Hergenrather&sortspec=date&submit=Submit)[3](http://jech.bmj.com/content/57/1/68.short#aff-3) (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. Badrul Sarwar, George Karypis, Joseph Konstan, and John Riedl (2000) “Application of Dimensionality Reduction in Recommender System – A Case Study”
9. Fayyad, U. M., Piatetsky-Shapiro, G., Smyth, P.,and Uthurusamy, R., Eds. (1996) - “Advances in Knowledge Discovery and Data Mining”
10. Rubi Boim (Under the supervision of Prof. Tova Milo) Tel-Aviv University “Methods for Boosting Recommender Systems”
11. Facebook API –FQL - <http://developers.facebook.com/docs/reference/fql/>
12. Wikipedia terms :
    * Web crawler - <http://en.wikipedia.org/wiki/Web_crawler>