PROPOSAL

CONTEXT BASED ALGORITHM FOR PROMOTION RECOMMENDATION SYSTEM

RESEARCH PAPER

Shan Markus Susanto

1701498326

Program Pascasarjana Ilmu Komputer

PROGRAM STUDI TEKNIK INFORMATIKA JENJANG S2

UNIVERSITAS BINA NUSANTARA

JAKARTA

2015

PROPOSAL

CONTEXT BASED ALGORITHM FOR PROMOTION RECOMMENDATION SYSTEM

RESEARCH PAPER

Shan Markus Susanto

1701498326

Supervisor:

Raymondus Raymond Kosala, Ph.D.

12 – 02 – 2016

TABLE OF CONTENTS

LIST OF TABLES

LIST OF EQUATIONS

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

1.1 Background

According to Merriam Webster, recommendation is the act of saying that someone or something is good and deserves to be chosen. We tend to give recommendation in our daily life such as recommending the best coffee in town to our best friends, giving a clue that the last Star Wars movie are worth watching. The essence of recommendation is reliable because we know that the item which is recommended by other people will more likely to satisfy us. As for in our Information Technology society, we define a system that give a recommendation to people is called a Recommender system (RS). This system has become more and more popular in our society and implemented in various application, because it could filter the related data that user interested rather than give all of information to the user. The most popular ones are probably to recommend movies, music, news, books, research articles, search queries, social tags, and products for people. In addition, in October 2006, a big company Netflix open a challenge for public in order to create a recommendation towards their movie recommendations ("The BellKor Solution to the Netflix Grand Prize," n.d.). They offer 1 million US Dollar to improve their recommendation system called *Cinematch* by 10% improvements.

The reason we implement RS in our society are, to create a better content for the user. Firstly, by using RS, be could identify our user behavior. By monitoring their preferences and likes for a certain object the RS will create an output just as like a wizard who know what the customer wants.

For example, the Last.fm create a station which contains recommended songs by calculating the user behavior towards the system. The system will also play the song which are not in the user library but played by other people who has similar taste to the music. This behavior also applies to the e-commerce recommendation system such as Amazon. By knowing the user preference, it will create an interactive perspective which give the user a good feeling seeing the item that RS has recommend and persuade them to buy or use the service.

Finally, conversion perspective which helps the company to grow. Such as, increase hit rate, optimize sales and profits margin, and many more. (Jannach, D). Recommendation also helps the company to plan what is the best item or move to be offered to the customer when selling their product or service. As for in IT terms, this tools can be seen in many applications for helping the customer making their next decision. For example, Amazon website, as the user can see several option after they bought or see several items in the website. It said the recommendation tool helps the market gain almost 35% of the profit rates. The other best example taken from "Now Touching The Void and Into Thin Air" example case. In 1988, a British mountain climber name Joe Simpson wrote a book which told his story to be the first to reach the summit of the Siula Grande in the Peruvian Andes with his friend. The book was having a good review which resulting a good success but soon disappearing in years later. Then a decade later, Jon Krakauer wrote a book Into Thin Air which also has the same plot with Joe Simpsons book. Instantly, Joe Simpsons book's become hits again, the reason behind this is Amazon.com recommendation acknowledge the pattern of their customer buying behavior and told their user that if they liked Now Touching the Void then they will likely to like Into The Thin Air. Their user took the suggestion then it create an impact of a rising demand for unpopular book to become popular.

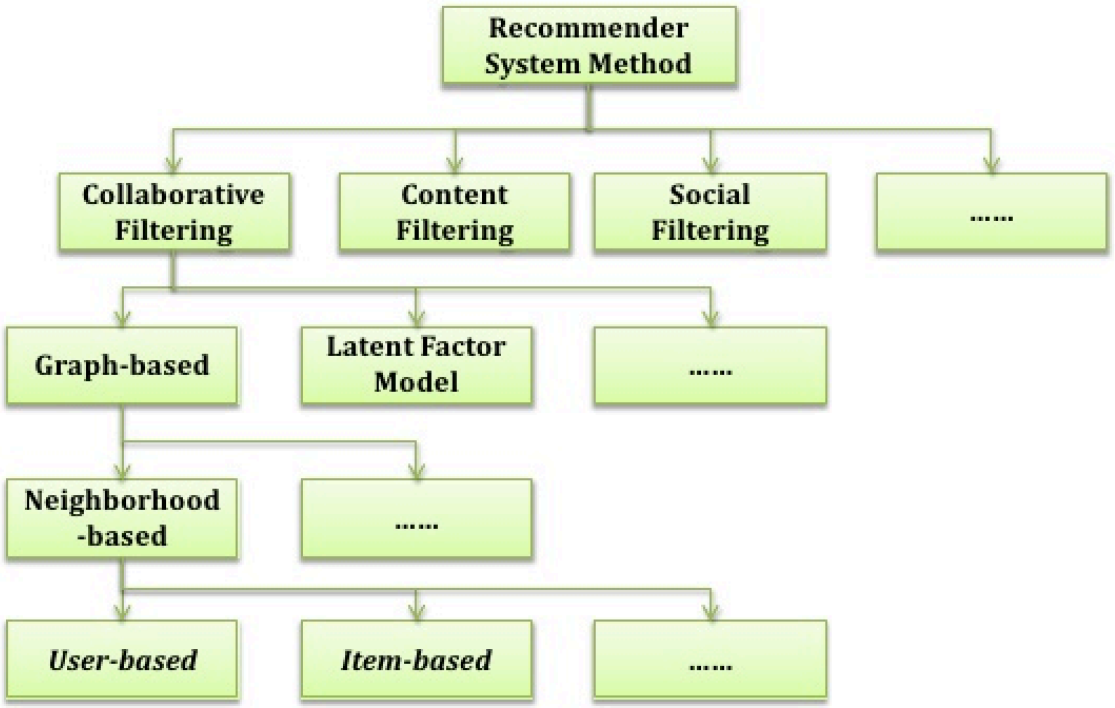

Figure 1 Recommendation Algorithm

There are several other methodology regarding RS, but the most common algorithm was Contenbt Based Filtering (CBF). According to research conducted by Joeran Beel (2015), from 62 reviewed approaches, 34 used CBF (55%). From these CBF approaches, the majority utilized plain terms contained in the documents. However, there are some fascinating algorithm which can be used to the RS. The Natural Algorithm is an algorithm that mimic an animal or plant behavior to the system. Janusz Sobecki (2014) found that there are several natural algorithms that can be applied towards the RS. One of them is PSO (Particle Swarm Organization), this algorithm was invented in 1995 by Eberhart and Kennedy. This algorithm was based on several things which are, position,

velocity and acceleration, and the movement of the particle around space. PSO can be referred as the school of fish, which each fish has their acceleration, coordinate, speed.

1.2 Problem Definition

Recommendation System has proven to bring great benefits towards human society. For example, it has become the tool for a company to attract more customer by giving them list of item that might interested. Not only it benefits the customer by giving them a less price or targeted item, but also benefits the company by giving them a loyal customer and profits. In these recent day, we can see many variables which can be put to the recommendation system which can be related to the RS. A context, which defined by Merriam Webster as the interrelated conditions in which something exists or occurs. A use of mobile device has a capability for monitoring user location and other service related that can play a part as the context for the RS. For instance, a location is a context which can be putted to a RS to create a recommendation to people near that location. In addition, time is another context that can be assign to give a specific time related recommendation to the user. The author believes, that by implementing the context to the RS will bring a great impact or improvement towards the accuracy to the RS.

1.3 Aims and Benefits

The Goal of this project can be defined in below

1. To compare context based algorithm between Collaborative Filtering (CF) and Content Based Filtering (CBF) and to investigate if the process can improve the recommender system.

2. To be able to create a recommendation system that using previous algorithm (context CF and CBF) and providing promotion recommendation data for PT. XYZ in daily basis.
3. To explore graph database technology to implement Recommendation System.

The benefits of this thesis are

1. Help the current company to gain more customer and profits.
2. The result of this thesis can become the foundation for other people when developing the next recommendation system.
3. As the base foundation for the author to create a recommendation platform with the same or similar data and structure in the future.

1.4 Scope

The scope of this thesis are:

1. To identify which algorithm that produce the best result for the PT XYZ. The author will use Content Based Filtering and Collaborative Filtering as the basic recommendation algorithm to use in this thesis
2. The data sample will be collected from the PT XYZ and other source
3. The recommendation data will be shown in the company application which is in IOS Applications

CHAPTER 2

LITERATURE REVIEW

In this chapter, the author will describe and explain regarding the theory that've been used for creating this thesis. In this chapter the author will explain several literatures which connected with the fundamental block for this thesis. Firstly, the author will discuss regarding the promotion which will be taken as the main data or item for the recommendation system. The aim for this research is not only to compare two different algorithms but also to produce the data which are usable for a certain party (PT. XYZ). Secondly, the explanation of RS algorithm itself will be explained in this chapter thoroughly and deeply. Finally, several database management systems will be explained in order to compare which system will be the most fitted towards this research analysis.

2.1 Promotional Marketing

According to McCarthy, Jerome E. (1964) Promotion can be define as to raising the customer awareness regarding services, item, improving sales, and getting customer loyalty to our product. There are several basic promotion type which available in our society. Firstly, there is a traditional promotion where the company gives the user a printable media such as column in a newspaper, electronic advertising such as in television and radio broadcast. Secondly, a new type of media which called digital promotion where the promotion can be spread using internets. Since almost 40% world population are online therefore the digital media prove to be the most effective way for company to expand their brand to their customer on their daily basis

2.1.1 Promotional Model

In the past 10 years, more than 200 studies regarding promotion has been published compare with year 1965 – 1983 which only 40 studies shown. Many definitions of promotions have been pass around the internet but it shares the common idea that promotions are a momentary and tangible modification of supply which has a purpose to create an impact towards customer, retailers, and sales behavior (Vemette 1990; Desmet 1991; Guilbert 1991; Ward and Hill 1991; Jones 1992). In addition, according to Diamond and Johnson (1990), the most widely accepted promotion topology contains three core foundations which are retail promotions, trade promotions, and consumer promotions. The topology is based on the nature of profits, effort, and relationship to the other product.

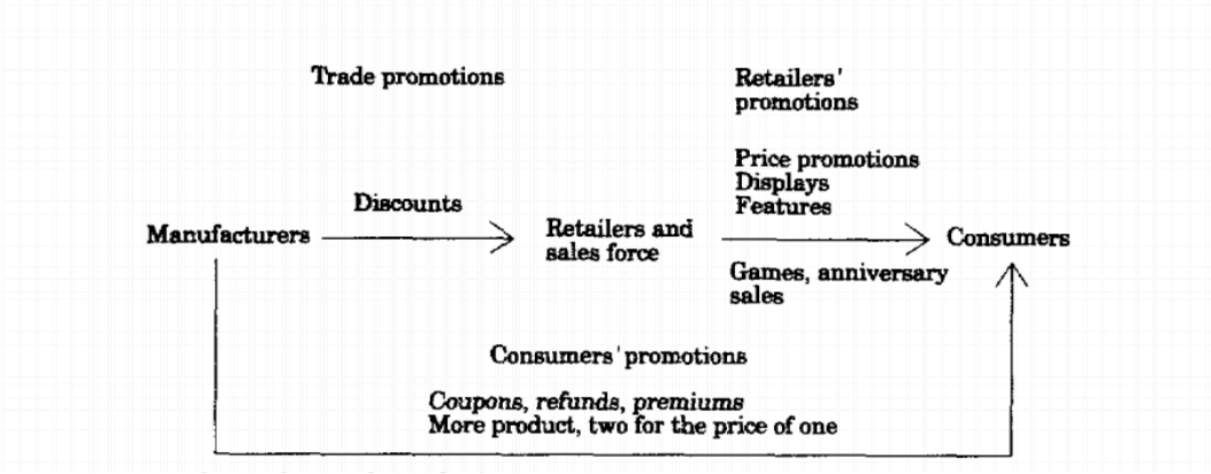

Figure 2 Promotion Topology

There are several contexts that need to be identify in the promotion model which shows in the figure below. First, the identification of frequent user on the based of risk theory

Dependent variable	Independent variable
Intention to redeem the offered coupon	Self-declared familiarity
Redemption of direct-mail coupon	Brand purchases
Increase in the repurchase likelihood of new buyers	

Table 1 Dependent and Independent Variables for Promotion

Secondly we could identify the promotion based on economic theory. In this identification process we could define two type of variables which is mediating variable which based on cost and benefits analysis and independent variables which based on a household characteristic.

Mediating variable	Independent variable
Substitution of a preferred brand for a less preferred, promoted brand	House of Ownership
Opportunity costs of time	Living in large cities
Inventory of stockpiled promoted brands.	Higher Education
Smart shopper feelings	Price Sensitivity
	Size
	Brand and store loyalty
	Children
	Working housewife
	Income
	Age

	Sex

Table 2 Variables that effect user choices of promotion

`

2.2 Recommendation System

Now days, we can see a lot of recommender system (RS) in our modern society. Mostly, the system can be found in the application which has a huge amount of data collections. The reason behind the use of the RS is the system need to provide the user with a list of items that they might buy or interested. In addition, not only providing user with the list of items but also create a personalized RS for each different user.

The basic reason behind the implementation for recommendation system is revenue. Amazon, Netflix, Spotify, and Apple Music are some example of big company who implements RS on their business model. Which proven to increase user engagement towards their application. The company revenue will increase significantly with more user engagement to the application. One of the example for this statement is Netflix. We know that Netflix provide their user with a list amounts of films and the user need to pay a subscription monthly fee in order to get the services. As the user start to select, watch, and rate each of the movie, the RS create a unique list of movie for every user, which make the user want to try to see the recommended movie.

The other simple example is “Now Touching The Void and Into Thin Air” example case. As the author already mentions it in the Chapter 1. Amazon prove that RS could increase their revenue towards relating items in a big collection amount of data.

2.3 Phases of recommendation process

The figure below shows the process of RS in order to give a recommendation towards user.

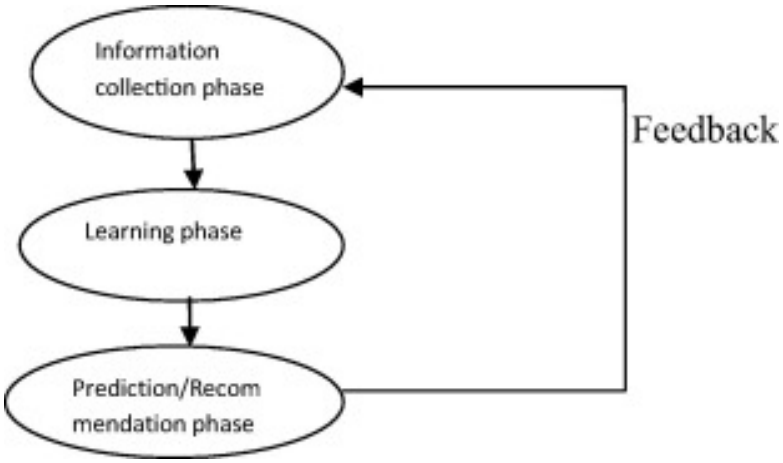

Figure 3 Recommendation process

2.3.1 Information Collection Phase

In this context, the system needs to put the relevant information in order to create a user profile or a model for the process of recommendation. This information includes the user attributes and behavior of the accessed content. The system needs to need as many attributes of the user in order to create a reliable recommendation from the data. In this context, we can observe user through different ways. Firstly, to ask the user regarding their interest or preference for a specific item. The most common feedback we can see in our society is rating system. In addition, another input can be assessed by observing user

behavior (Oard & Kim, 2000). Finally, hybrid feedback can also be collect by combining those two methods.

2.3.1.1 Explicit feedback

Like the author has briefly discuss below, the system or application will send a form of interface so the user can provide ratings to an item. This process will improve the accuracy for the user model in determining more reliable data. The recommendation system accuracy feedback will be fully determined by the quality of the feedback from the user. On one hand, using explicit feedback will bring a great impact towards the recommendation system because it gives a direct and more reliable data from the user (Buder & Schwind, 2012). On the other hand, since the explicit feedback need more effort from the user, sometimes the user does not ready to put enough information to the system or application.

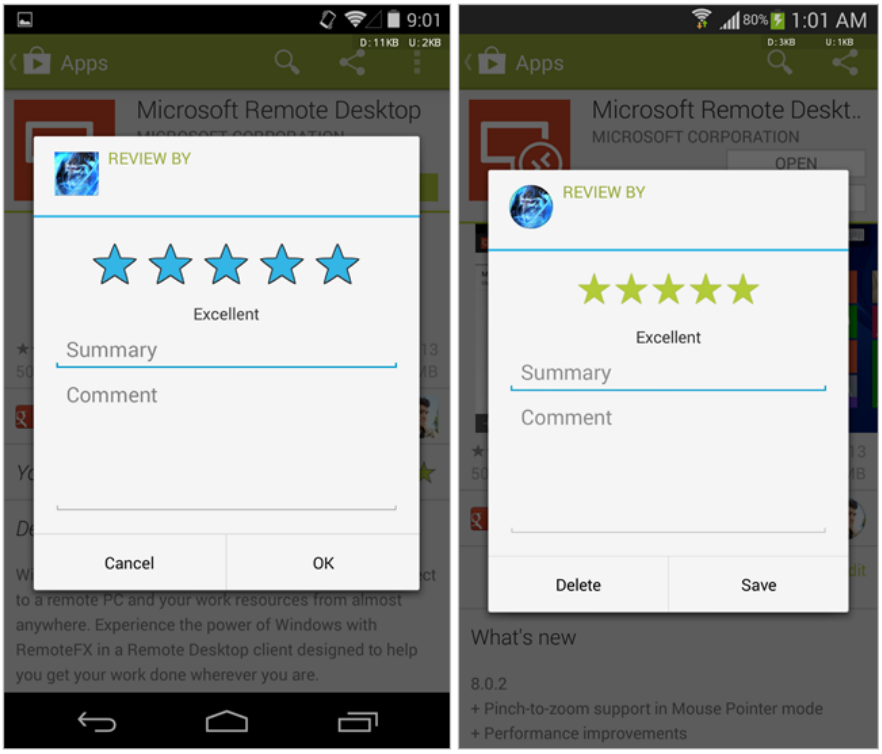

Figure 4 Explicit Feedback Example

2.3.1.2 Implicit feedback

While the explicit feedback asks directly to the user for the feedback towards the item. The implicit feedback took feedback from monitoring the actions to the users. For example, the history of purchases, time spent looking in a specific web page, kind of category the user clicked the most. Which bring great benefits towards RS feedback system, because it subconsciously took the data from the user without the user put an extra effort to the system. Moreover, a research (Gadanho & Lhuillier, 2007) shows that the data from implicit feedback may be more objective because there's no bias when the user do the actions.

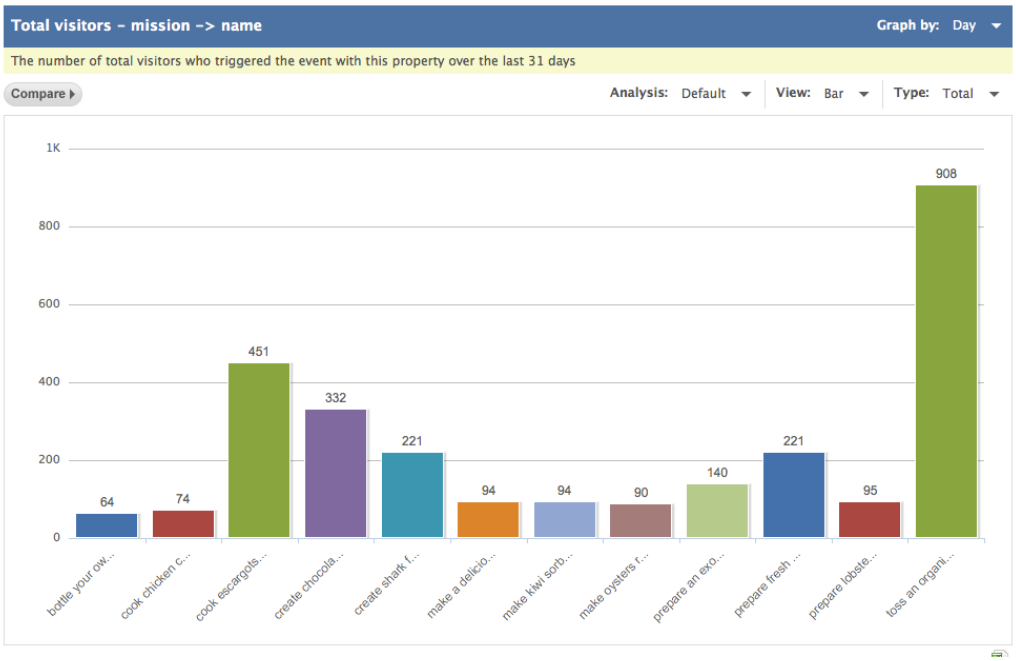

Figure 5 Implicit Feedback Example

2.3.1.3 Hybrid feedback

Both of pro and cons in doing the implicit and explicit feedback can be form into a hybrid system. The purpose of hybrid feedback is to minimize each of their weakness so the RS could perform the best feedback input from the user.

2.3.2 Learning phase

In this phase, the RS apply several algorithms to process the user feedback from the information collection process. The algorithm that common to be used are Collaborative Filtering (CF) and Content Based Filtering.

2.3.3 Prediction/recommendation phase

After RS finish doing the learning phase, it will start to provide predictions towards what item that user is preferred. This data set can be made from several methodologies which from memory based or model based.

2.4 Recommendation Algorithm

The first research recommender paper was introduced by Giles et al (1998). The paper was discussing about Bibliographic screening techniques as a part of CiteSeer project. As the time pass by, more than two hundred research article regarding RS has been published with different concepts and approach in order to create the best RS. According to by Joeran Beel (2015), half of the RS (55%) is using content based filtering as their approach, while the second most used algorithm is collaborative filtering (16%). While the other are spread for using graph database, stereotyping, and hybrid recommendations.

2.4.1 Content Based Filtering

Content Based Filtering (CBF) is one of the most widely use algorithm in RS implementation. A brief definition of CBF is to find a related dimension or domain towards an item so it can be referred to the user. The biggest benefits towards using CBF is it does not require as much user feedback to make the RS works ("The wonderful world of recommender systems | Yanir Seroussi on WordPress.com," n.d.). The CBF mostly use terms of Items, Interactions, and features in order to generate recommendation. For instance, "item" can be assume as a book and the "interaction" can be putted as buy, see,

like, vote, downloading, etc. In addition, each items have in own features or tag that can be listed to.

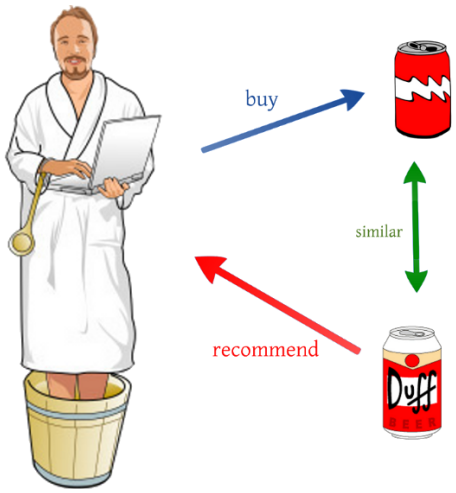

Figure 6 Content Based Filtering Illustration

This are the simple example of CBF, as we can see if a person who likes beer with label A and if there exist another beer with label B who shares the same feature with the first beer. Therefore, the person who likes to buy beer A will most likely to buy beer B.

CBF proven to be a natural approach for RS in many problems. For instance, when a person watches a Starwars episode one and two. The most logical action is to recommend the third episode of the Starwars. However, the features of each items are hardly defined, the items context is dependable towards the descriptive of data. This problem will cause some user to get a similar recommendation which already showed to their profiles. In addition, Content-based filtering also ignores quality and popularity of items (R. Dong, L. Tokarchuk, and A. Ma, “Digging Friendship: Paper Recommendation in Social Network,”

in Proceedings of Networking & Electronic Commerce Research Conference (NAEC, 2009, pp. 21–28)

2.4.2 Collaborative Filtering

This approach is probably the second best approach for RS. The Collaborative Filtering (CF) is created by building a big dataset of user preferences for items then pair it to the other user with the same interest by calculating the similarities between them (Herlocker, Konstan, Terveen, & Riedl, 2004). The usual terms that CF tend to use is neighborhood.

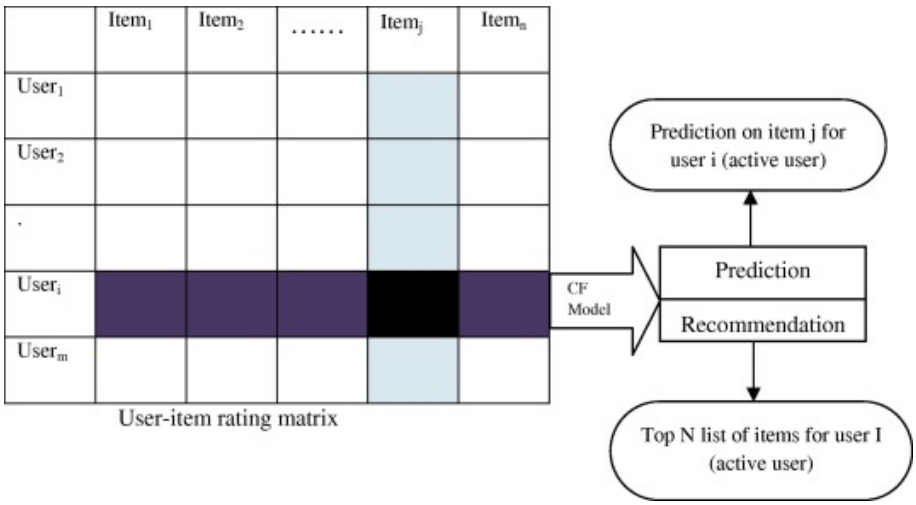

Figure 7 Collaborative Matrix Example

As we can see from the graph, CF use matrix to match targeted user from the other user perspective. By matching the matrix, the RS could create recommendation and prediction towards the user. The major limitation for CF is the reliance towards user choices and the most common problem for using this algorithm is the cold start problem scenario. In this scenario, the items which presented to the user does not enough rating which will create a

poor recommendation result to the user. By general CF can be categorize into two parts which are memory based and model based filtering.

2.4.2.1 Memory based techniques

According to Badrul M. Sarwar (2001), the idea behind this technique is to find a relation between user and items from the database. The items that already rated by the user will search a neighbor that shared the same interest with the user. The neighbors usually have a history of doing or approving with the targeted user. For example, they both share the same interest for buying the particular item. The Memory based CF can be done by using two techniques which is user or item based technique.

The first technique (User Based) will calculate the similarity between user and compare their preference to the same item. After the data has been achieve, then the system will compute the prediction of rating.

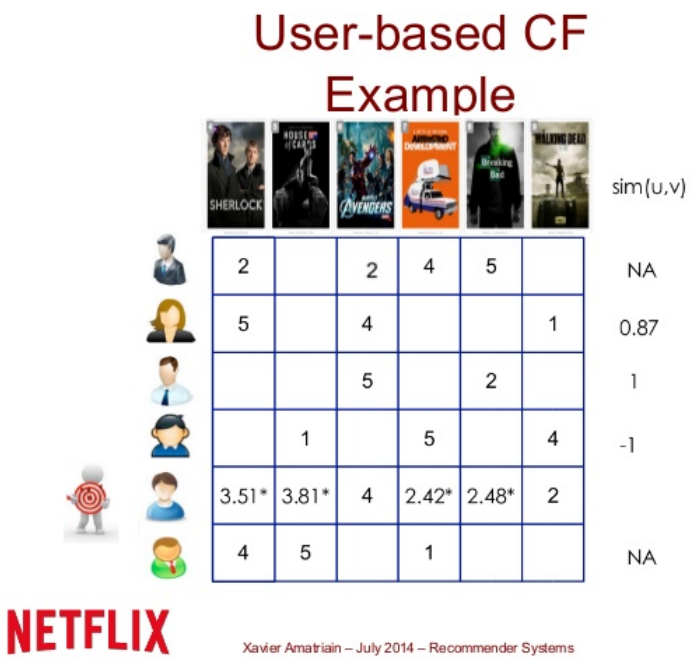

Figure 8 User Based CF Example

The illustration below shows; the user based CF was performing in the small data set of movie system. As we can see, the RS wants to predict the rate for each movie based on their neighbor who has the same interest to the same kind of movie.

The second technique for memory based recommendation is item based CF Recommendation. To find an item which the targeted user has liked before is the based foundation for this CF technique. The item based technique will retrieve all items that has been rated by targeted user. After the data has been collected then it will compare the number to the targeted items to determined the similarity between both items. Noted in this techniques takes only the known similar item ratings by the test user into account for prediction

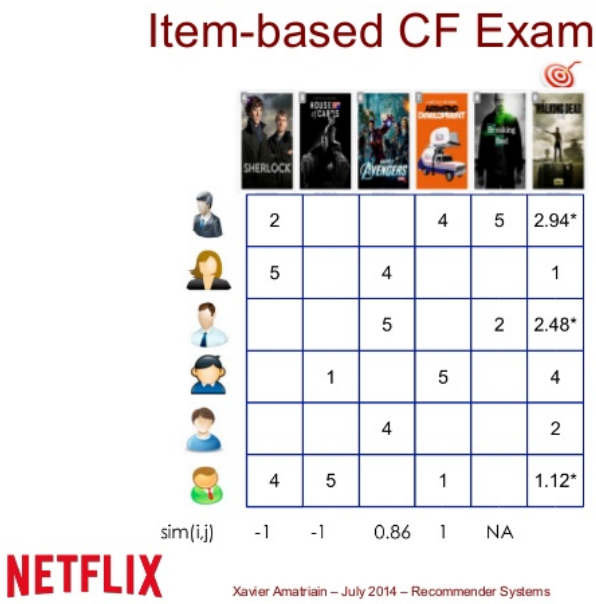

Figure 9 Item Based CF Example

According to a research (Jannach, Zanker, Felfernig, & Friedrich, 2010), in order to measure the correlation between two most popular items, Pearson correlation is used and the algorithm can be defined as

$$s(a,u) = \frac{\sum_{i=1}^{n}(r_{a,i} - \overline{r_a})(r_{u,i} - \overline{r_u})}{\sqrt{\sum_{i=1}^{n}(r_{a,i} - \overline{r_a})^2}\sqrt{\sum_{i=1}^{n}(r_{u,i} - \overline{r_u})^2}}$$

Equation 1 Pearson Correlation

From the figure below the s(a,u) explain the similarity between two user a and u while ra,i is the input rating which set from the user a. In addition, $\overline{r_a}$ is the mean rating which has been set by the specific user while n is the number of items. Moreover, we could also calculate the deviation from neighbor mean by applying this formula,

$$p(a,i) = \overline{r_a} + \frac{\sum_{i=1}^{n} (r_{u,i} - \overline{r_u}) \times s(a,u)}{\sum_{i=1}^{n} s(a,u)}$$

Equation 2 Neighbors Deviation

2.4.2.2 Model based techniques

This technique will utilize the whole database to generate prediction and recommendation. Model based use the previous rating in order to create or examine a model. This model will be revised as the system or ratings grow which will give a high quality for the user. One of the model based technique used in the real world is Neflix Prize. In brief, Netflix create a competition to public in order to improve their recommendation system. After several years pass by, in the last year of competition year (2007), Netflix Prize end with two top models which is Singular Value Decomposition (SVD) and Restricted Boltzman Machine (RBM).

Both of the algorithm are only a part from many possible algorithms that can be apply for the model based technique recommendation. The use of each algorithm is depend on the style of the customer in selecting a product. In the section below, the author will describe some of common used models when selecting model based technique for recommendation.

2.5 Algorithm Model

In this section, the author wants to explain the model type that might be possible to be used in order to implement the RS

2.5.1 Association Rule

This model was used when the data mining wants to predict the connection between items in an actions towards them. In real life, an example of study case might be like a number of customer who buy soap in a general store often buy a toothpaste at the same time. This model will calculate the connection between them and might find that 70% of the checkout section that contain soap also contain toothpaste as the same time (Rouse, Margareth, n.d) . The effectiveness of association rule often used to analyze sales of transaction and the impact towards them is already known from a quite sometime.

2.5.2 Clustering

Clustering algorithm attempts to grouped a set of data into a set of smaller group in order to create more meaningful group inside that datasets. This algorithm often used in pattern recognition, statistical data analysis, and image processing. When a cluster has been created, the user feedback to each cluster can be calculated and process in order to create a more specific cluster which fit for the next predictive in RS. A good clustering algorithm will produce a cluster where the intra-cluster similarity is high and the inter-cluster similarity is low. In the RS approach, the user must play a part in clustering process and will be calculated by using degree of participation. The most common algorithm for this algorithm is using K-means which showed in figure below

$$\arg\min_{\mathbf{S}} \sum_{i=1}^{k} \sum_{\mathbf{x} \in S_i} \|\mathbf{x} - \boldsymbol{\mu}_i\|^2$$

Equation 3 K-mean Algorithm

Given a set of remarks (x1, x2, …, xn), where each inspection is a d-dimensional real vector, this algorithm of clustering aims to partition the n observations into k (≤ n) sets S = {S1, S2, …, Sk}

2.5.3 Decision tree

This methodology was based using tree graph which created by calculating a set of collection of data with their attributes. The decision tree is trained algorithm which mean if the training was correct then it will produce a very reliable data or prediction in RS. However, there are some advantage and disadvantages for using this algorithm which is, Firstly the algorithm is capable for generate an understandable rule for human to read. In addition, it could perform classification without too many hassle on doing computation towards the system (Caruana & Niculescu-Mizil, 2006) . However, decision tree will perform poorly in small data set of case and not suitable for prediction in continuous attributes since it not capable for doing dynamic improvements (add new value).

2.5.4 Artificial Neural network

Artificial Neural Network (ANN) is using terms of neurons when applying their algorithm. The algorithm is to find how many connected neurons in the system in a logical way. The

neurons can be interpreting as a computational unit which will receive inputs, process the data in order to give an output (“Artificial Neural Networks for Beginners,” n.d.) . The connection between neurons can be in unidirectional or bidirectional.

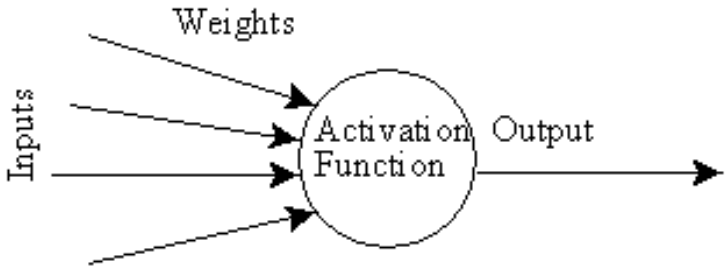

Figure 10 Artificial Neuron Network Model

In the figure below shows how ANN is modeled. There are basically consist of some fundamental attributes such as inputs which contains of several weights inside of its and by using activation function which calculate or computed the data, and then gives an output. The benefits for using ANN model is to estimate nonlinear computation and collecting complicated connection in a collection of data. While the disadvantage for using this algorithm is to come up with the correct or fit network topology for the problem.

2.5.5 Bayesian Classifiers

Bayesian Classifiers are a probabilistic framework for solving a grouping problems which took the definition of conditional probability and using the Bayes Theorem. The basic Bayes Theorem for Bayesian Classifiers is

$$p(C_k|\mathbf{x}) = \frac{p(C_k)\ p(\mathbf{x}|C_k)}{p(\mathbf{x})}.$$

Equation 4 Bayesian Classifiers

Whre p(Ck|x) is the probability of instance x in Ck and p(x|Ck) is the probability of generating instance x given class of Ck. Moreover, p(X) is the number of instance x occurring and p(Ck) is the number of instance Ck is occurring too. By using this theorem as foundation, we could apply naïve Bayesian classifiers which able to solve a record of N features (A1, A2, …, AN), the goal of the classifier is to predict class Ck by finding the value of Ck that maximizes the subsequent probability of the class given the data P(Ck|A1, A2, …, AN).

The main benefits of Naive Bayes classifiers are that they are robust to isolated noise points and irrelevant attributes, and they handle missing values by ignoring the instance during probability estimate calculations. In addition, they are fast to train and classify. However, the neutrality assumption may not grasp for some attributes as they might be associated. Which for RS the Bayesian are more suitable with a slow change of user preferences not a real time or rapid changes.

2.5.6 Matrix completion techniques

The purpose of this algorithm is to predict the unknown values in the user item matrices. One example of this problem could be seen in below where the RS wants to predict the score of one movie for specific user. In most of RS, they have a big key set value paired matrix data and the scrubby amount of data which due the reason that some users does not input the value most of items in the matrix system. This problematic issue sometimes cause the RS are unable to give a reliable and accurate output towards the user.

One of wide use algorithm for calculating Matrix completion technique is Alternating Least Square. This algorithm helps to minimize the square error over observed entries while keeping other factors fixed. Keshavan et al, used SVD technique in an OptSpace project algorithm to deal with matrix completion problem. The result of their experiment proves that SVD is able provide a trustworthy initial evaluation for bridging subspace which can be further refined by gradient descent on a Grassmannian manifold which later known as SGD.

Model based techniques solve sparsely problem. The major drawback of the techniques is that the model building process is computationally expensive and the capacity of memory usage is highly intensive. Also, they do not alleviate the cold-start problem.

2.6 Database Management System

A database management system (DBSM) is a tools for computer system in order to create and maintain databases. This system provides an organized way for user and programmers in order to do a Create, Update, Read, and Delete data (CRUD). This tools have a core function for maintaining three important parts which are the data, data schema which define the database structure, and database engine which provide an access, security, and managing the data. This three core function help to provide concurrency, safety, data integrity, and uniform management procedure. In addition, DBSM also capable for doing role back (running a previous version of a database), restarts, and recovery function.

The DBMS is probably a very useful tools in order to providing a unified view of data that can be accessed by several user using different platforms and locations. A DBSM can constraint the limit regarding the data access for each user. In addition, the end user and programs are capable for understanding where the data is stored because the DBMS handle all the requests.

2.6.1 Relational Database Management System

There are several popular DBMS model such as Relational database management system (RDMS) which adaptable to most problem in software engineering. This database supports the relational data model, which defined by the table name and a certain number of data with data types. The RDMS use record as a row in the table which contain several values for each attributes. The basic operation for this RDMS include classical set of operations such as union, intersection, and difference. In addition, the user also capable for doing a selection for a subset of record with several criteria such as contain. The RDMS also

capable for doing projection which able to select a subset of attributes in the table. Finally, the most useful operation in RDMS is join, which enable the user to combine multiple table. The example for RDMS is Oracle. However, although it could solve almost the problem, the RDMS can be quite expensive for development.

2.6.2 Document Stores

The next type of DBSM is NoSql DBSM which suited for loosely data structure that might grow in the future. NoSql also can be defined as Document Stores. This database model has several characteristic, firstly, the records does not have uniform structure (different records might have different columns). Secondly, the types of values in each column can be differ from one another. In addition, each column may have more than one value. Finally, record can have a nested structure (append values). In the real practice, the Document stores often use internal notations such as JSON. Wide column stores is another example of DBSM which could handle a very large numbers of dynamics columns. This type of DBSM can be seen as two dimensional key stores values. Since the column and record key are not fixed. This type of DBSM share the same attributes with the NoSQL DBMS but the implementation in the real practice is quite different.

2.6.3 Graph Database

According to data from DB-Engines.com, a website that tracks database gain more popularity. Graph databases were the fastest growing type of database in 2014. This database represent data in a graph projection using terms nodes and edges. By doing this,

graph database allows easy processing and simple calculation of the data. An example for graph database is neo4j graph database.

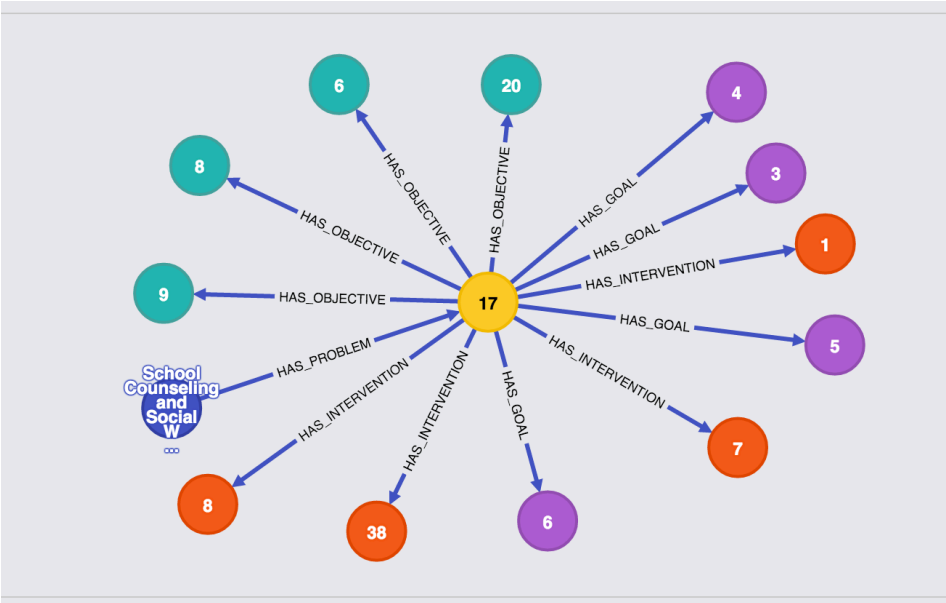

Figure 11 Neo4j Graph Modelling

In the figure below shows the representation neo4j for data (node) with their edge. As we can see each of node has different color representing its model and the arrow represent the edge as the function for each node. The pro's for using graph database according to Josep Lluis and Larriba Pey, the graph database allow programmer to deep transversal quicker than traditional database. Moreover, the graph database tends to query faster when finding a connection between nodes which shares the same preferences in their edge. The cons for graph database is the maturity for its system because it is a growing technology.

CHAPTER 3

METHODOLOGY

3.1 Research Framework

According to Cambridge dictionary, recommendation can define as “a statement that someone or something would be good or suitable for a particular job or purpose, or the act of making such a statement”. Recommendation system has become the most important tools in order to give the user a good and reliable choice of items. Most of company who sells goods will use recommendation system for increasing their revenue and knowing their customer preferences. The simple example for this scenario is Amazon (One of biggest E-commerce company)

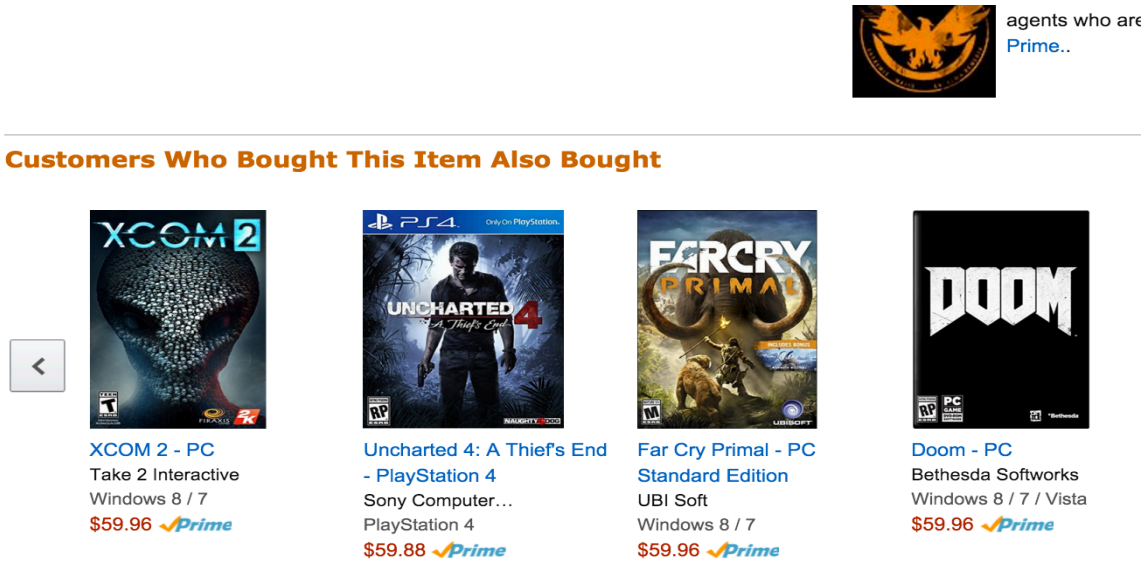

Figure 12 Example of Amazon Collaborative Filtering

The image below shows a recommendation using content based filtering which took the attributes of the selecting item to all item in the company list. In addition, the website also offers another recommendation method which offer the user what other user do after viewing or buying this items

Figure 13 Example of Content Based Filtering on Amazon

The recommendation system described above use basic algorithm for iterating the result to the user (CBF and CF).

In this research, the author will focus on the context promotion recommendation. There are several reasons behind this is. Firstly, because promotion has become one of the important tools for a company to gain their customer loyalty. Secondly, it helps customer to drive customer buying decision making for the next products. Finally, it simply will grow revenue for the company who gives the promotions.

In the literature review below, there are several basic algorithms in order to create a recommendation system from a collection of data. First of them is Content Based Filtering which took the item attributes and then try to find the attributes similarity towards other items in the data collection. The other algorithm is Collaborative Filtering which took the neighborhood value to

give the output recommendation. The author believes that the algorithm can be improve by adding context to the system.

The type of this thesis is research paper. However, in order to collecting the data, the author took one start up company as the source in data collection. PT ABC is a company who use promotion as its service. The main purpose of this company is to provide their customer with visible, transparent, and reliable promotion around them so the user could use them. In addition, PT ABC also able to help the merchant in order to increase number of customer which has the interest towards their products. Before we continue further, in the diagram below the author wants to explain the step for doing the research:

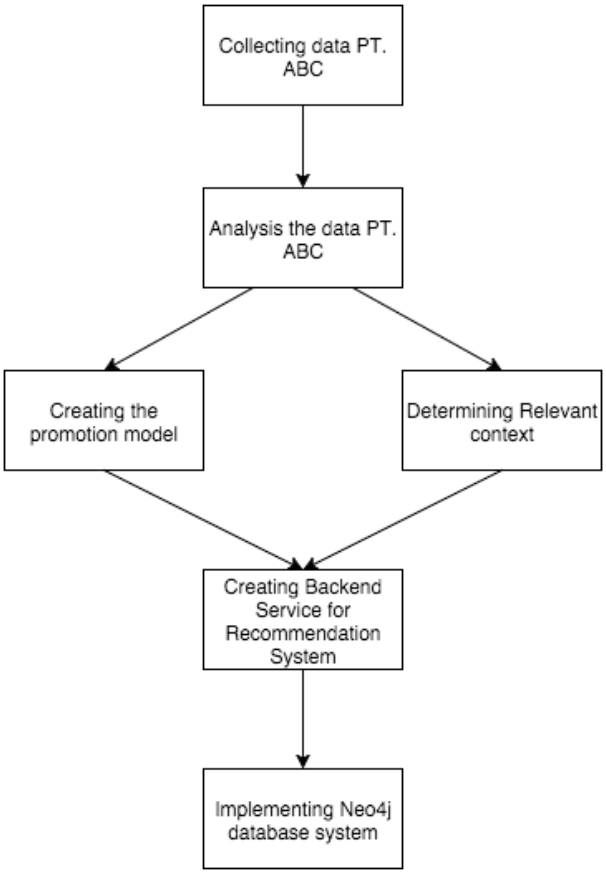

Figure 14 Thinking Structure Breakdown

3.2 System Design

From what the author has been discussed below, the evaluation of the diagram will be explaining in this chapter below.

3.2.1 Data Collection

In this data collection, the author will take the company data which based on the company Mixpanel account. According to Mixpanel.com,

> *"Mixpanel is the most advanced analytics platform for mobile & web. Instead of measuring pageviews, it helps you analyze the actions people take in your application. An action can be anything - someone uploading a picture, playing a video, or sharing a post, for example."*

With data, the author would able to retrieve all the several information inside the company data such as merchant statistic, how many redeem promo for in specific terms and who are they, which promotion has the more accessed, data of virtual check in, and many more. Figure below shows one of the example regarding what type of promotion that user most prefer,

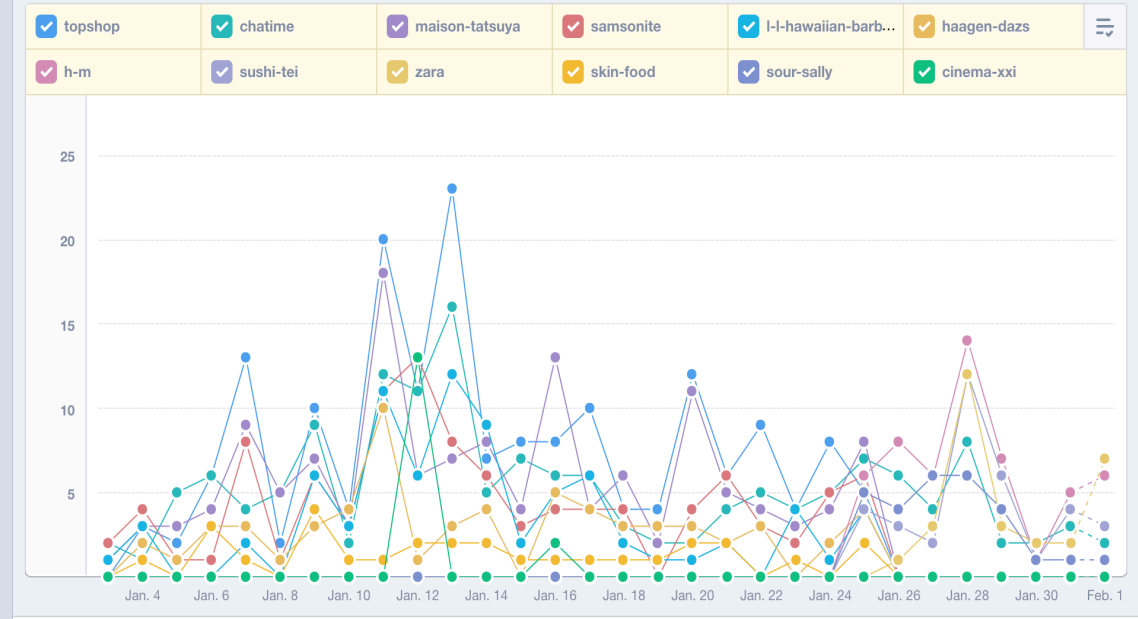

Figure 15 Mixpanel User Checking Promotion Data

As we can see from the table, there are several dots which represent the number of selected promotion in a specific day. For instance, in between January 14 and 16 the Topshop shows a highest rate of clicked. Moreover, we also could extract some data regarding which merchant who has the most popular rating around the user

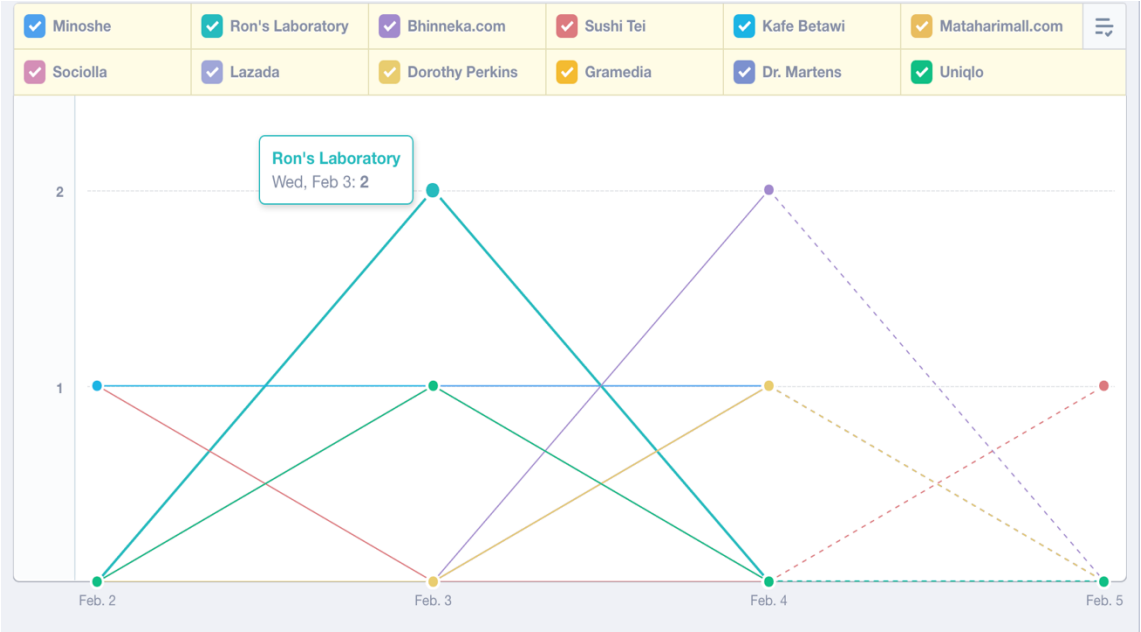

Figure 16 Mixpanel Merchant Popularity

During data processing, the ideal way for collecting data would be using automated system which took the daily data form the company server and send it to the recommendation system to be process. However, for the sake of scope satisfaction the data from the company might be taken daily using several CSV file taken from mix panel and uploaded to the recommendation to be process.

3.2.2 Analyzing Data

In this step, the author will analyze the data that has been collected from the fist step. The data analyzing will be consisting of several step. Since the research was based on doing on Content Based Filtering and Collaborative Filtering therefore there are several main models for this research purpose.

	Node	Value
Promotion	Category	Books, Dining, E-Commerce, Education, etc
	Type Deals	< 25%, 25% - 50%, Buy 1 Get 1, etc
User Profile	Gender	Male, Female
	Interest	Football, Basketball, Sports
	Education	High school, Bachelor degree, Middle School
Merchant	Category	Books, Dining, E-Commerce, Education, etc
Mall	Location	Mall Location
Outlet	Location	Outlet Location

Table 3 Model For Recommendation System

This are the draft model regarding the output from data analyzing from PT. XYZ. As it mentions below, there are main model which will play the biggest role in Content Based Filtering which is promotion model. In the promotion model, we define the promotion will consist category and type of deals. The category attributes will define what niche of the promotion is. In addition, by putting the type of deals the recommendation system will able to classify which type of deals that interest the user the most.

While in the Collaborative Filtering, the main foundation for this algorithm is the user itself. The data from PT. XYZ consist the data of the user who use the application. However, for security issue the classify data attributes such as name, age, etc will not be display in this research paper.

After the main node has been collected, then the author starts to collect contexts data which related or connected to the current main node. As it mentioned in chapter 2.1.1, there are some context which the author could took as the context which are,

	Context	**Value**
Time	Time	Moring, Afternoon, Night
Location	Location range	Near, Mid, Far
Weather	Weather	Rainy, Cloudy, Sunny
Income	Income	Low, Medium, high
City Target	City	Suburbs, City, Rural

Table 4 Context for Recommendation System

In the company ERD diagram there are several contexts that might apply for the current recommendation system. First of all is UserCheckin table. In this table we could extract where the user usually checks in and when. In addition, the author will create several extra contexts such as weather condition during the next prediction for check in. For the weather condition, it could be taken from online resource.

3.3.3 Creating Backend Service

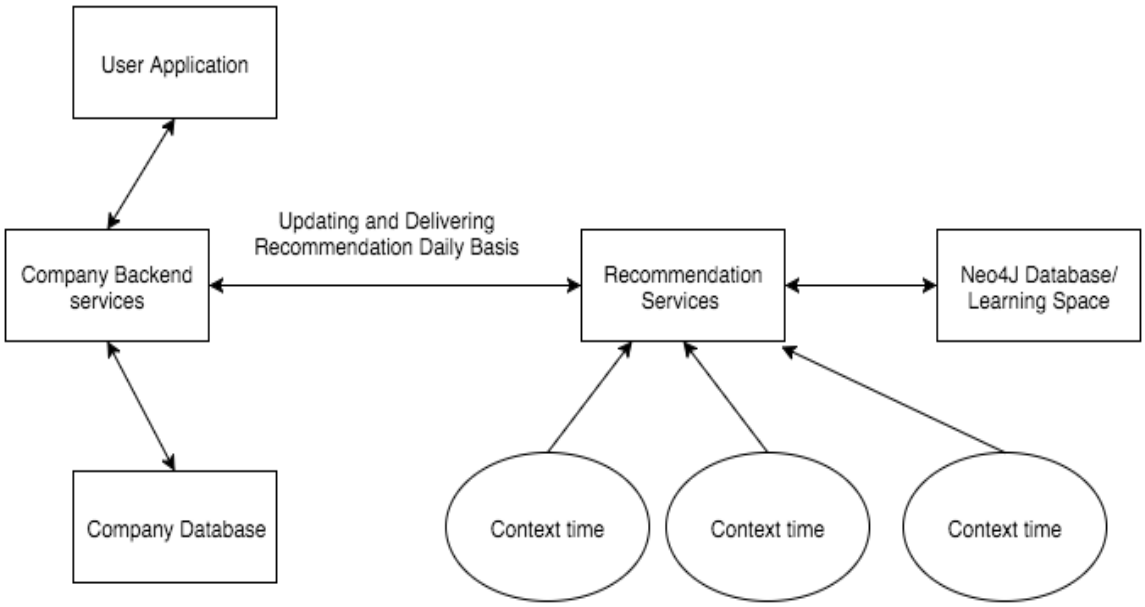

Figure 17 System Architecture

The figure below shows the basic structure of how the recommendation system will be implemented during the process. In the first process, the data will be collected by recommendation service engine in overnight time. After the data was taken, it will transform the collected data towards the neo4J database system and mapped through nodes and relationship. For the diagram below is, the current context (time, location, and weather) become an independent variable towards the recommendation system. After the recommender system is running then we might add more context in order to improve the prediction accuracy to the user.

After the data processing is done, the recommendation system will be ready for giving the recommendation towards the company backend system. The result from the accuracy and

reliability recommendation will be tracked using the Mixpanel tools and will be used as a learning tools to provide another recommendation. In addition, there might be explicit user feedback by asking the user how the recommendation benefits to them.

3.3.4 Implementing Neo4J

For this thesis purpose, the author will use graph database as the database management tools in order to process the data collection from the company. One of reason is because Neo4J offer a free community product which help the author reduce cost for developing the system. Another reason is the graph database is a new technology that still young and have a lot potential to be explored especially for this kind of purpose. Like it mentions before, Neo4j is a graph database which help user to visualize the data in terms of nodes and relation. This graph database offers a feature for load a CSV file and process it as the programmers wants to. Since in this research all the data from the backend system will be transform to the graph, therefore there are several model that will be implemented in the Neo4J database.

3.3.4.1 Implementation Promotion Node

```
1 LOAD CSV WITH HEADERS FROM
  "https://dl.dropboxusercontent.com/u/53022271/testData.csv" AS line
2 CREATE (p:PromotionTemp { name: line.banner_id })
```

Equation 5 Promotion Node Equation

This is an example query for loading the promotion test data to the Neo4J. The result from this query can be seen as node,

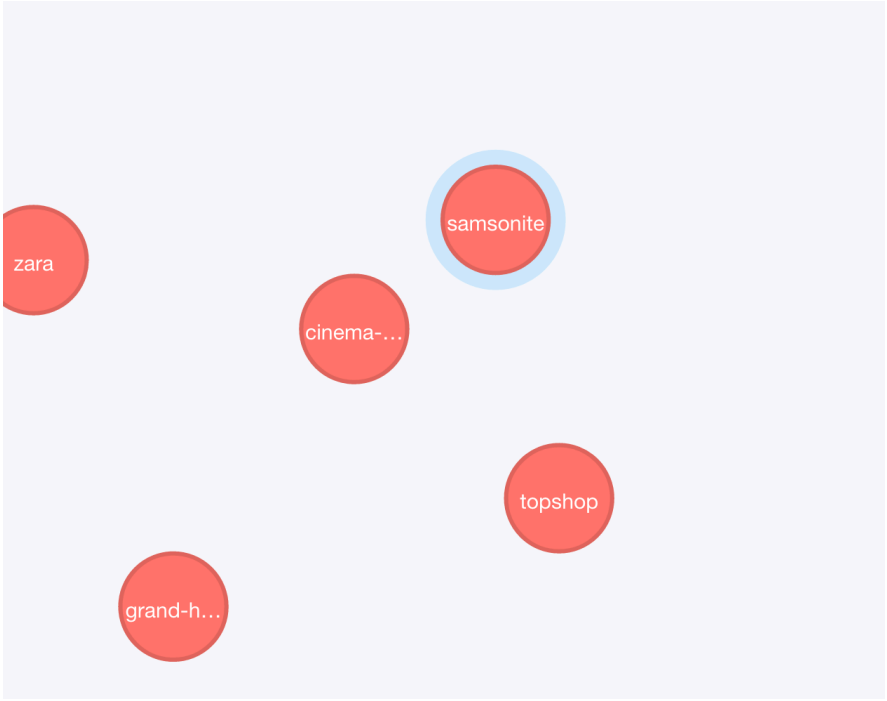

Figure 18 Promotion Node

The figure bellows shows some of nodes that has been created after imported the promotion data from the CSV file.

3.3.4.2 Implementation of Merchant Node

```
1 LOAD CSV WITH HEADERS FROM
  "https://dl.dropboxusercontent.com/u/53022271/testData2.csv" AS line
2 CREATE (p:Merchant { name: line.merchant})
```

Equation 6 Merchant Node Equation

With loaded data CSV then it will produce

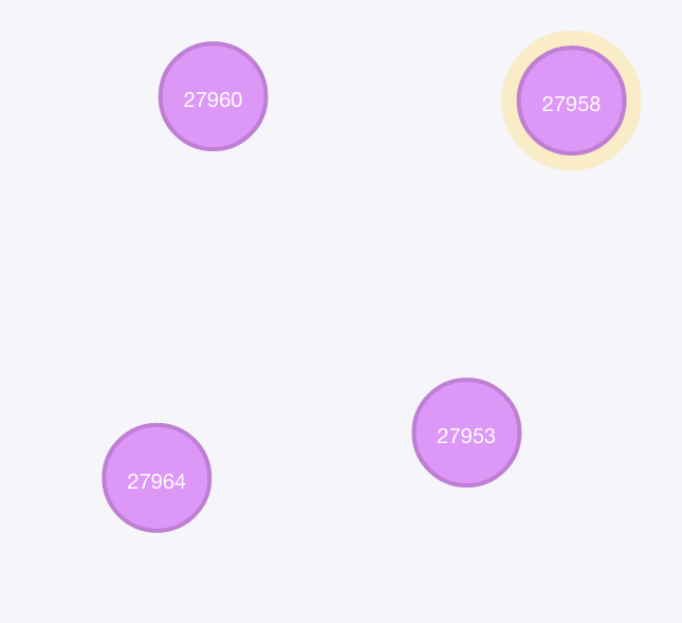

Figure 19 Merchant Node

With Attributes Node ID and Node Name

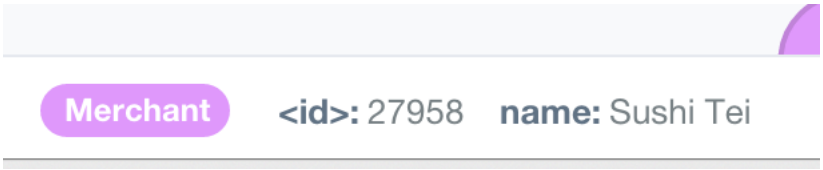

Figure 20 Merchant Node Attributes

3.3.4.3 Implementation of User Node

First we input the same test data from the previous data in order to receive the number of active user who open the merchant page on the apps by running this query,

```
1 LOAD CSV WITH HEADERS FROM
  "https://dl.dropboxusercontent.com/u/53022271/testData2.csv" AS csvLine
2 CREATE (p:Person { id: toInt(csvLine.user_id) })
```

Equation 7 User Node Equation

And the output of this process will be

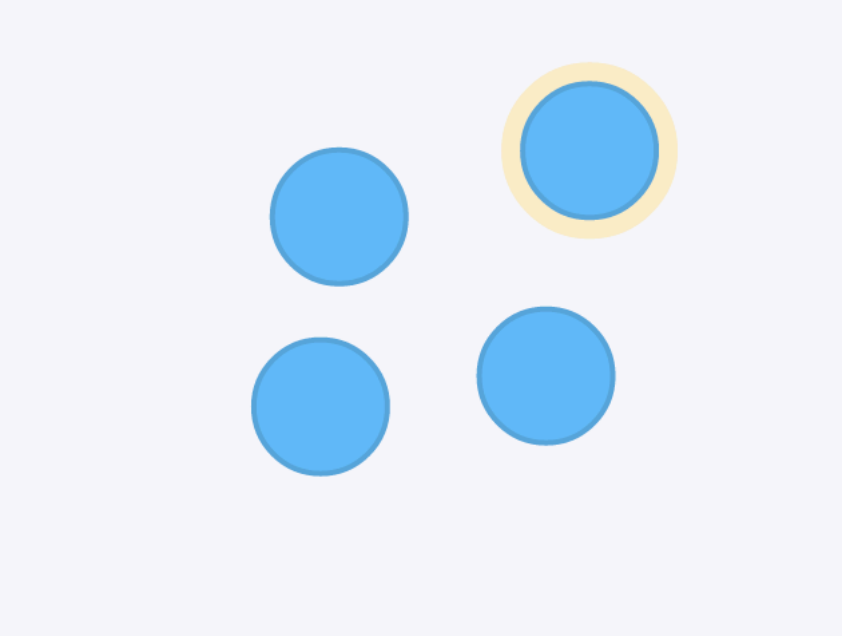

Figure 21 User Node

Figure 22 User Node Attributes

Noted from the figure below the data was not added with the title because of company policy of customer secrecy for not showing name in this research paper.

3.3.4.4 Implementation of User Promotion Relation

After the data was generated, therefore we are able to create a relation between two node such as, what kind of promotion does a specific user check to. The example of the query can be found in below,

```
LOAD CSV WITH HEADERS FROM "https://dl.dropboxusercontent.com/u/53022271/testData2.csv" AS
line WITH line WHERE toInt(line.count) > 0
MERGE (m:Merchant {title:line.merchant})
MERGE (p:Person {name:line.user_id})
MERGE (d:Date {date:line.Date})
MERGE (p)-[:CHECK_MERCHANT{date:d.date, count:line.count}]->(m) RETURN count(*);
```

Equation 8 Check Merchant Relation Equation

The result of the query would be look like,

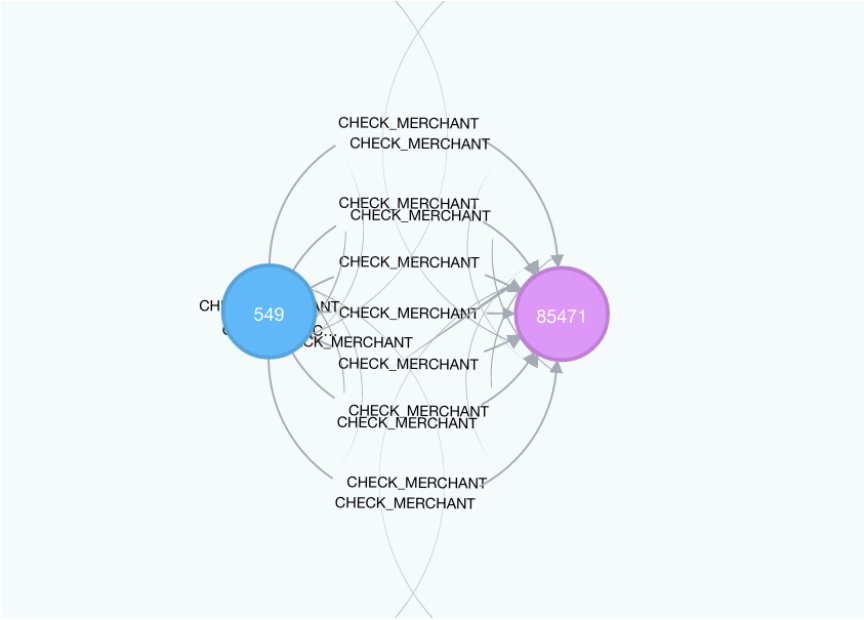

Figure 23 User Merchant Check in Relation

In this relation shows that each of relation was represented using a line, currently the user with id 549 has checked the merchant with id 85471 (Coconut Island) for 25 times in the different day and time. For example, in first January 2016, the user visits the merchant five times.

REFERENCE

Artificial Neural Networks for Beginners. (n.d.). Retrieved January 30, 2016, from http://arxiv.org/pdf/cs/0308031.pdf

Buder, J., & Schwind, C. (2012). Learning with personalized recommender systems: A psychological view. *Computers in Human Behavior*, *28*(1), 207–216. http://doi.org/10.1016/j.chb.2011.09.002

Caruana, R., & Niculescu-Mizil, A. (2006). An empirical comparison of supervised learning algorithms. In *Proceedings of the 23rd international conference on Machine learning - ICML '06* (Vol. 2006, pp. 161–168). New York, New York, USA: ACM Press. http://doi.org/10.1145/1143844.1143865

Gadanho, S. C., & Lhuillier, N. (2007). Addressing uncertainty in implicit preferences. In *Proceedings of the 2007 ACM conference on Recommender systems - RecSys '07* (p. 97). New York, New York, USA: ACM Press. http://doi.org/10.1145/1297231.1297248

Herlocker, J. L., Konstan, J. A., Terveen, L. G., & Riedl, J. T. (2004). Evaluating collaborative filtering recommender systems. *ACM Transactions on Information Systems*, *22*(1), 5–53. http://doi.org/10.1145/963770.963772

Jannach, D., Zanker, M., Felfernig, A., & Friedrich, G. (2010). *Recommender Systems*. Cambridge: Cambridge University Press. http://doi.org/10.1017/CBO9780511763113

Oard, D. W., & Kim, J. (2000). Implicit Feedback for Recommender System. Retrieved from https://www.researchgate.net/publication/2320773_Implicit_Feedback_for_Recommender_System

The BellKor Solution to the Netflix Grand Prize. (n.d.). Retrieved February 5, 2016, from http://www.netflixprize.com/assets/GrandPrize2009_BPC_BellKor.pdf

The wonderful world of recommender systems | Yanir Seroussi on WordPress.com. (n.d.). Retrieved February 5, 2016, from http://yanirseroussi.com/2015/10/02/the-wonderful-world-of-recommender-systems/