Data Quality and K-map

Master of Science Project

Final Audit

MONTCLAIR STATE
UNIVERSITY

Faculty:

Prof. H. Mark Hubey

Developement:

Student: Krupaben Patel

ABSTRACT

In the field of Information Retrieval, the important measure is Data and Information Quality. At each step of retrieval process we measure the quality of data by validating it on particular dimensions. The proper co-relation of this these dimensions is important when we consider the data quality issues. If the balance of the co-relation is inappropriate then there may occur serious issues like Challenger Disaster by NASA. All the other Data quality measure were concerned at that accept one. So, it is really important to measure the data quality with concern of all the dimensions and here I have tried to propose a project to fit all the dimensions of data quality into a K-map from which we can accurately determine the nature of data quality.

ACKNOWLEDGEMENT

I would like to express my gratitude to Dr. Mark Hubey for his suggest me idea of this master project and guide me for that. His deliberations and considerable effort to collect and come up with the comprehensive set of requirements served an important role in shaping the application.

I also would like to thank many faculty members at Computer Science department who helped me learn many topics in Computer Science, which enabled me to achieve my academic goals.

I cannot forget to express my gratefulness towards my family members including my mother, my sisters and especially to my lovely husband Mr. Alpesh Patel, without whom it would not be possible to complete my study.

TABLE OF CONTENTS

1. INTRODUCTION

1.1 Background and Motivation :

The idea of this project belongs to Dr. Mark Hubey. The idea of using K-map to determine the data quality is convenient and useful as nowadays people hold mobile phones in their hand, carry tablets to surf internet and use laptop for their work, information retrieval has become the most important field in IT industry. Thus, the Big Data has become so much big that store the data and retrieve the data has become a difficult task. Information retrieval is a process of retrieving information from some storage media. Information retrieval has several definitions, but it has a vast meaning because the field consist of a huge amount and category of information which actually comes from totally different kind of fields such as healthcare, IT industry, finance, education, etc. The fields may be different but they all have a big data and they upload it on directly or indirectly on internet. Even if they do not upload on internet either way they use some data store to store and retrieve it through networking. We can define information retrieval as "Information retrieval is the activity of obtaining information resources relevant to an information need from a collection of information resources. Searches can be based on metadata or on full-text (or other content-based) indexing" [Wikipedia]. As the name suggests, information retrieval isn't just retrieval of information from some media that you plug into the USB port and get all the information or just search on Google and get all the information you want. Information retrieval is rather a big field and it consists of several concepts and regulations to be maintained before it is delivered to the user. The database operations for information retrieval include data storage, updating, editing and validation of information being retrieval. Data being processed need to maintain its quality. Thus, information and Data Quality is one of the measure of the Information Retrieval field where quality of data and information need to maintained at all steps of retrieval process.

1.2 Problem Definition :

In order to describe the problem addressed in this paper, let us look at the definitions of information quality and data quality.

Information quality can be defined as "Information quality (IQ) is a term to describe the quality of the content of information systems. It is often pragmatically defined as: The fitness for use of the information provided" [Wikipedia].

Data quality can be defined as "This list is taken from the online book Data Quality: High-impact Strategies.

- Degree of excellence exhibited by the data in relation to the portrayal of the actual scenario.
- The state of completeness, validity, consistency, timeliness and accuracy that makes data appropriate for a specific use.
- The totality of features and characteristics of data that bears on their ability to satisfy a given purpose; the sum of the degrees of excellence for factors related to data.
- The processes and technologies involved in ensuring the conformance of data values to business requirements and acceptance criteria. Complete, standards based, consistent, accurate and time stamped" [Wikipedia].

Different organizations have different kind of data and use various categories of information. Thus, there arises a need to consider different measurements for data quality and information quality. Mostly we think that the most important measurement of data quality is its accuracy but there are 16 of them include accuracy, timeliness, ease of operation, etc. The increasing size of data on the internet and Big Data of enterprises comes up with a big problem of data quality errors and that may cause serious problems. This can be useful in all the fields which use internet and other storage media also. We all know that the space shuttle disasters by NASA caused by data quality errors. Not only that recently apple launched iPhone6 and iPhone6 plus and after a few days of its launch it came out that it get bent when you put in pocket, this is data quality error. To analyze the quality of data and information is thus a certain requirements in such enterprises. After a lot of research, researchers have defined 4 dimensions with 16 categories but they are unequally distributed and to determine the quality of data is difficult.

1.3 Problem Solution:

My approach to solution of this problem is about how we can fit this 16 categories of data quality into a Karnaugh or K-map and get an idea about the nature of quality. We know that Karnaugh map or simply K-map is used to simplify the Boolean algebra expressions. K-map can maximally consist 4 variable and it can be represented with maximum 16 different combinations of 4 variables. Thus, using K-map we can determine the nature and nurture of quality of data and information as we have exactly 16 categories of them and they are grouped into 4 dimensions. Thus, this research can be useful at most in determining the data and information quality in a better way.

2. Data and Information quality dimensions:

2.1 Data and Information Quality dimension:

A definition of Data quality dimension written in a paper by N. Aksham et.al (2013) "A Data Quality (DQ) Dimension is a recognized term used by data management professionals to describe a feature of data that can be measured or assessed against defined standards in order to determine the quality of data"[Page-2].

Many researchers have found a variety of data quality dimensions depending upon the use and understanding. Dr. Richard Wang, co-director of MIT's Total Quality Management researched about Data quality dimensions. Initially, there were around 176 categories and they continuously been reduced and at last there 16 categories with 4 dimensions.

The 4 dimensions of Data Quality include: Intrinsic DQ, Contextual DQ, Representational DQ and Accessibility DQ.

Dimensions	Measures
Intrinsic	Accuracy Objectivity Believability Reputation
Accessibility	Accessibility Security
Contextual	Value-added Relevancy Timeliness Completeness Amount of Data
Representational	Interpretability Ease of Understanding Representational Consistency Conciseness of representations Manipulability

Table 1: Information Quality Categories and Dimensions

As above figure represents the four dimensions of information quality includes 16 categories. "Intrinsic means the data is directly knowable from the data, then the quality is said to be intrinsic." Accuracy, Objectivity, Believability and Reputation comes under this dimension as all these categories are intrinsic in behavior. Accuracy is how much accurate information we get from the input data. It generally means that the data is free from errors. Many people consider this category as a substantial part of Data Quality. But, what if the user does not believe about accuracy of data? This happens a lot of time when user does not believe because their perceptive about the solution of a problem can be totally or partially different than the developer. In Software Development Life Cycle(SDLC), the similar issues occur when user needs are different than the developed version. The judgment of creating data is also a very strong part as it is objectivity of a data and it degree of objectivity affects believability the most. As time passes on, data builds a reputation about its quality and maintenance. For example, if you start searching through one search engine and you get your resulted pages which you like the most, you will be used to that search engine and if sometimes it gives some unlikely pages then also we will be using it only such as Google. Thus, the reputation matters most in data quality.

Accessibility means how we can access the data from a storage media or from other sources. Accessibility includes accessibility and security. Accessibility means how quickly the data is retrieved and accessed. Security means as the its name suggests that how one can use the data with all the security for example when we are using locks on database or passwords for administrative websites. The security measure is an essential part of data quality.

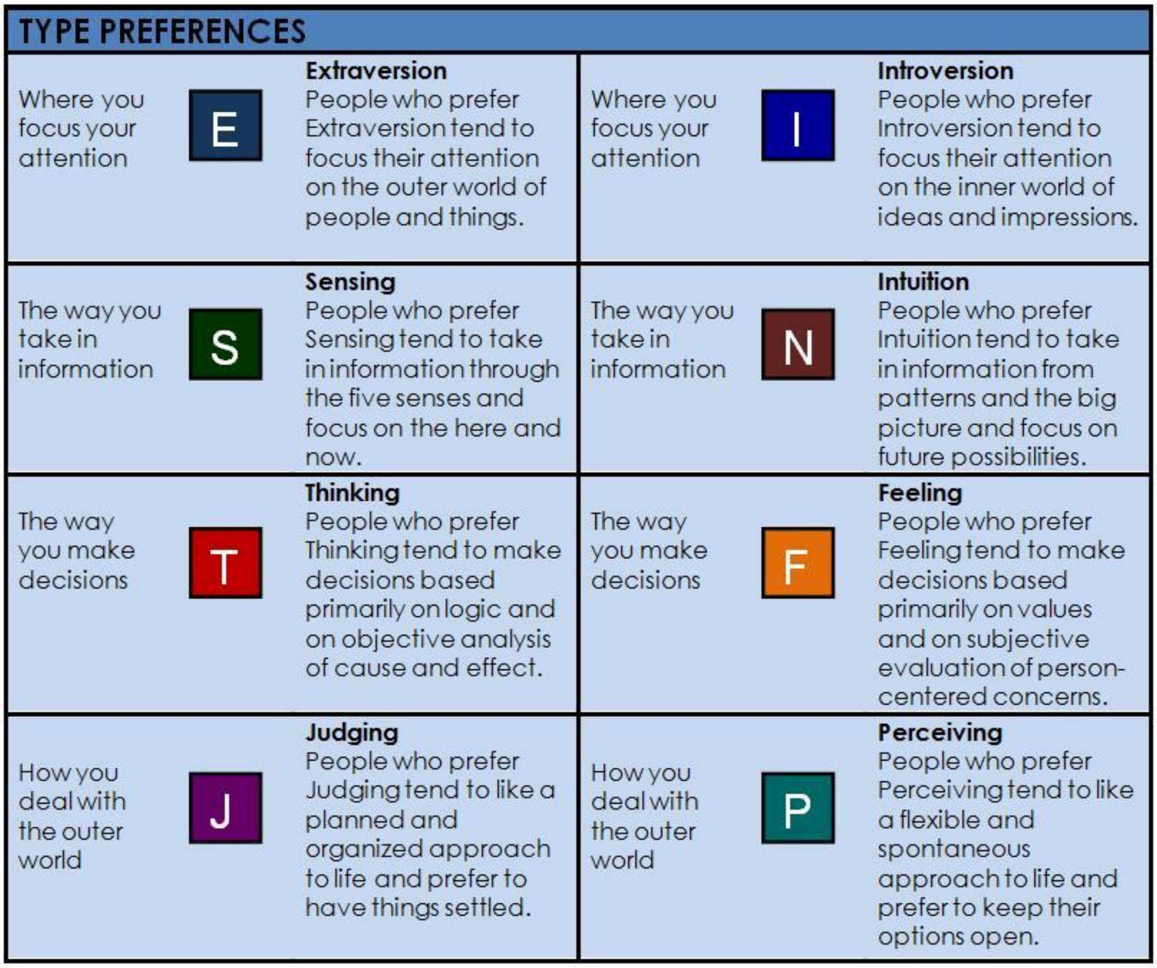
Context is very important measure in Data Quality. Without the context and data storage all the other dimensions are worthless. Contextual dimensions include value-added, relevance, timeliness, completeness and amount of data. Value-added measure is about the benefits we can get from use of the data. Relevancy measure is how data is relevant to the particular task or action. It depends on user queries as when user query arises that particular problem can be solved if the data is relevant. Timeliness refers to the amount of time the data is residing in the database. If the data is too old and is not useful then it is wasting the space or if a particular task requires only latest data, then too old data can be useless. Completeness refers to the values of the data present in the storage. If total value of data is complete than the data can have completeness measure. Amount of data refers to how much amount of data is getting from the storage when the user queries arise. The amount of data should be dependent upon the task, some task requires large data and some task requires less data.

Representational measure is also an important dimension of the data quality as it includes, interpretability, ease of understanding, consistency, conciseness and manipulation. Interpretability means that data should be clear in terms of language or terms. Ease of understanding refers that data should not be ambiguous and should be easily comprehended and the data should be much clear to be understood. Consistency dimension refers to the common formats, which means that if machines differ the data should have a common format which can be accessed from each machine on the network. The conciseness dimension means the data retrieved should be brief enough that user can access and understand it. It does not mean the shorter data but data should not be unnecessarily large that user do not want.

3. K-map and Information Quality Dimensions:

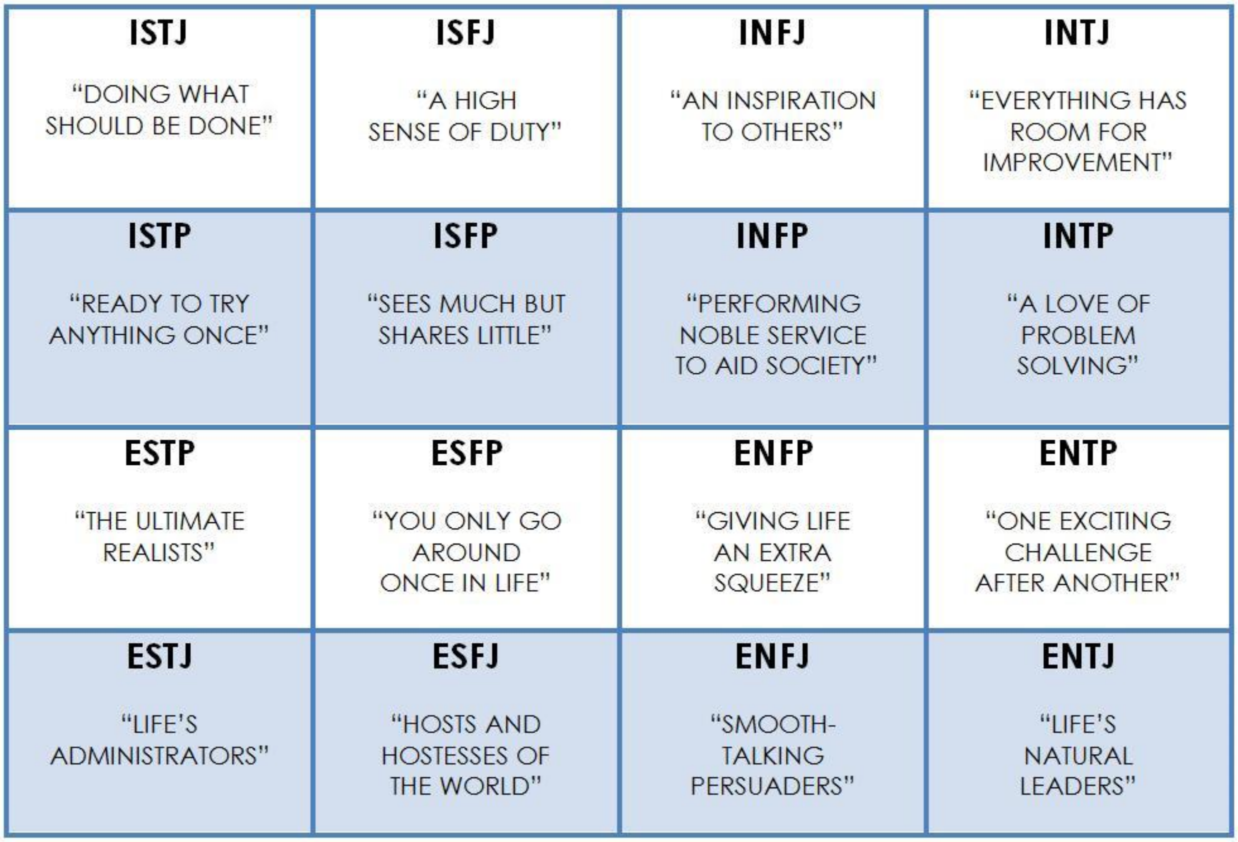
K-map: K-map is generally used to simplify Boolean Algebra expressions. K-map can have 2,3 and 4 variables. We can represent 4 Boolean variables in a K-map and represent one equation related that.

For example, we already have a personality test using K-map. In that there are 8 types of preferences with total as 16 personalities. One can determine a person's personality depending upon the quality he/she has.



TYPE PREFERENCES					
Where you focus your attention	E	**Extraversion** People who prefer Extraversion tend to focus their attention on the outer world of people and things.	Where you focus your attention	I	**Introversion** People who prefer Introversion tend to focus their attention on the inner world of ideas and impressions.
The way you take in information	S	**Sensing** People who prefer Sensing tend to take in information through the five senses and focus on the here and now.	The way you take in information	N	**Intuition** People who prefer Intuition tend to take in information from patterns and the big picture and focus on future possibilities.
The way you make decisions	T	**Thinking** People who prefer Thinking tend to make decisions based primarily on logic and on objective analysis of cause and effect.	The way you make decisions	F	**Feeling** People who prefer Feeling tend to make decisions based primarily on values and on subjective evaluation of person-centered concerns.
How you deal with the outer world	J	**Judging** People who prefer Judging tend to like a planned and organized approach to life and prefer to have things settled.	How you deal with the outer world	P	**Perceiving** People who prefer Perceiving tend to like a flexible and spontaneous approach to life and prefer to keep their options open.

Figure 1: Personality Preferences



ISTJ "DOING WHAT SHOULD BE DONE"	ISFJ "A HIGH SENSE OF DUTY"	INFJ "AN INSPIRATION TO OTHERS"	INTJ "EVERYTHING HAS ROOM FOR IMPROVEMENT"
ISTP "READY TO TRY ANYTHING ONCE"	ISFP "SEES MUCH BUT SHARES LITTLE"	INFP "PERFORMING NOBLE SERVICE TO AID SOCIETY"	INTP "A LOVE OF PROBLEM SOLVING"
ESTP "THE ULTIMATE REALISTS"	ESFP "YOU ONLY GO AROUND ONCE IN LIFE"	ENFP "GIVING LIFE AN EXTRA SQUEEZE"	ENTP "ONE EXCITING CHALLENGE AFTER ANOTHER"
ESTJ "LIFE'S ADMINISTRATORS"	ESFJ "HOSTS AND HOSTESSES OF THE WORLD"	ENFJ "SMOOTH-TALKING PERSUADERS"	ENTJ "LIFE'S NATURAL LEADERS"

Figure 2: Personality types

The different combinations of personalities represent a person's personality. All the information we can see from figures that how dimensions were converted into K-map and one can determine his/her own personality type.

The similar way we can represent 16 categories of data quality in a K-map and determine the quality of data from the chart. I will develop the chart for the data quality dimensions and it will be really useful in big enterprises where we have big data.

4. EXPERIMENTAL EVALUATION:

We discuss the implementation of our approach along with a summary of our experimental evaluation.

4.1 Implementation:

As per we have four dimensions of data quality here which are intrinsic, contextual, representational and accessibility. All four dimensions include measures as per their characteristics. As per stated we make experiments to calculate data quality by mapping its dimensions on Karnaugh map. For that, we need to make four Karnaugh map each for one dimension as we have total 16 measurements to be mapped. So, we have decided to implement one Karnaugh map for one dimension and we can implement the same way the other dimensions. And here, we are going to implement the accessibility dimension which has ease of operations, accessibility and security measures. To map this dimension onto K-map, we have to make more analysis on the measures as we must have 4 measures in each dimension.

Thus, we need to look at the measures again. Intrinsic DQ include accuracy, believability, objectivity and reputation which seems fair where all these measures have intrinsic property and thus they should be categorized into the same category. Next contextual DQ, which contains value-added, relevancy, timeliness, completeness and amount of data. Contextual DQ is related to the context of the data. Here, there can be a discussion about timeliness measure as timeliness refers to the most recent and relevant data stored. If it is considered in terms of context, we can think that it lies under contextual DQ, but there is also another point that if the access time of data becomes much larger or it is not possible to get the most recent and relevant data then it comes to access. This point is very important as at the end, user wants appropriate, accurate and on-time data. If the access time takes even some minutes more, the user leaves to try and move forward to another work. Thus, here for assumption we consider "timeliness" measure under Access category. Representational category includes interpretability, ease of understanding, manipulability, representational consistency and conciseness of representation. Representation category involves the user for representation. From all of the measures from representation category, manipulability means how the user manipulates which seems related to Access category. Thus, here we take another assumption that "manipulability" come under Access category and we have total four measures in Access category which are accessibility, security, manipulability and timeliness.

Calculation of Dimensions:

K-map can have four dimension each with 0 or 1 value. Thus, we need to calculate each dimension with 0 or 1 value. But, here some complications arise as quality measures always cannot have binary value.

Method: We can calculate by rating the dimension from 0 to 4. For example, if user can rate the dimension from 0 to 4. But, we can say that rating 0 and 1 can be value 0 and 3 and 4 can be value 1. So, we have to assign value to 2 between 0 and 1. We can think of "fuzzy logic" here. Fuzzy logic is a way to find the truth beyond the binary value.

First, we take Accessibility dimension: Accessibility is that we can retrieve data or not. So, it can be assigned binary value where 0 means not retrieved/not accessed and 1 means retrieved/accessed.

Security: Security is based on authorized access. So, there can also be binary value where authorization and authentication applied or not and 0 or 1 value depending upon that.

Manipulability: Manipulability is based on how easily data can be modified. Here, it is necessary to rate the dimension. So, the value can be defined between 0, 1 and fuzzy value.

Timeliness: Timeliness refers to the age of data. Timeliness depends on user and other circumstances. Thus, we need get rating for this dimension too, which is same as manipulability.

K-map:

To draw k-map for Access category, we need to make assumptions for values.

- Every dimension will have binary value for easily getting the overview.

K-map of access category:

		Manipulation, Timeliness			
		00	**01**	**11**	**10**
Accessibility, Security	**00**	0000	0001	0011	0010
	01	0100	0101	0111	0110
	11	1100	1101	1111	1110
	10	1000	1001	1011	1010

Table 2: K-map of Access Category dimensions

Analysis of K-map:

K-map is used to simplify the Boolean algebra. To do analysis of K-map we need to get the Boolean expression from K-map.

Here, from K-map we can say that the best case of data quality is where the value 1111 occurs. 1111 value means that all the data quality dimension is rated as the best and all of them are present within the retrieved data. But, we need to discuss the average case too.

Average Case:

To consider average case, I have used Hamming distance. We can find hamming distance from a binary field to another field which differ by only 1 bit. By doing this, we can conclude with all the possible best cases.

Here, the best case 1111 has nearby 0111, 1110, 1011 and 1101.

Manipulation, Timeliness

		00	**01**	**11**	**10**
	00	0000	0001	0011	0010
Accessibility, Security	**01**	0100	0101	0111	0110
	11	1100	1101	1111	1110
	10	1000	1001	1011	1010

Table 3: K-map of Access Category dimensions with Hamming Distance bits

Analysis of K-map with characteristics:

Interpretation:

1111 : Accessible, Secured, Manipulable and Timed(up-to-date data), this case is best case where all the dimensions present; hence best

0111: Not accessible, Secured, Manipulable and Timed(up-to-date data), here the data is secured, easily modified, latest but not easily retrieved; hence not useful

1011: Accessible, Not secured, Manipulable and Timed(up-to-date data), here the data is easily retrieved on time and manipulable but not secured; not reliable data

1101: Accessible, Secured, Not manipulable and Timed(up-to-date data), here the data is easily retrieved on time, secured but not manipulable; hence static data

1110: Accessible, Secured, Manipulable and Not Timed(old data), here the data is easily retrieved, secured, manipulable but the not recent data is retrieved; hence old data

Other cases: The behavior of other cases are shown in the table given below.

In table,

Red color: The most inferior quality

Light red color: Inferior quality

Light green color: Good quality
Green color: Best quality

		Manipulation, Timeliness			
		00	**01**	**11**	**10**
Accessibility, Security	**00**	Worst case	Only latest data	Latest and Modifiable	Only Modifiable
	01	Secured, not retrieved, old and static data	Secured, not retrieved, latest and non-modifiable data	Not Accessed	Secured, modifiable, not retrieved on-time data
	11	Secured and static old data retrieved	Static	Best case	Old data
	10	Data retrieved without security, time and modifiability	Latest data retrieved without security and modifiability	Not Reliable	Old data retrieved without security

Table 4: K-map of Access category with analysis

This way we can map other categories onto K-map and analyze the same way.

Benefits:

- Using hamming distance and K-map, we can map more dimensions and reduce the time complexity
- All other categories can be analyzed same way and thus easily achieve the result

Limitations:

- We can't map fuzzy values onto K-map
- The rating can't be accurate every time and thus can't predict the data quality accurately

5. CONCLUSION:

We have addressed the issue of Data Quality and its dimensions. In my work, I have completed fitting the Access category dimensions into a K-map and determine the nature of Data Quality and thus we can improve the quality of data easily and quickly. This can be mostly useful as we all are using internet and it can be specifically useful in Information Retrieval field where the Data Quality and Information Quality measures mean much important. This can be my contribution toward the fields in database, information retrieval and website maintenance.

:List of Figures:

:List of Tables:

:References:

1) Wikipedia.org

2) N. Askham, D. Cook, M. Doyle, H. Fereday, M. Gibson, U. Landbeck, R. Lee, C. Maynard, G. Palmer, J. Schwarzenbach (October, 2013), "Defining Data Quality Dimension".Retrieved from http://www.damauk.org/RWFilePub.php?&cat=403&dx=2&ob=3&rpn=catviewleafpublic403&id=106193.

3) L. Ng (2012). Retrieved from http://www.genycoaching.org/2012/10/24/post-1/ [Figure- 1, 2]