Classification

1. *method background*

Classification research method is used to classify or group entities with similar properties and behaviors together. In computer science classification method could be used in identifying threats or level of importance of entities. The method is applicable in areas where there is a need to cluster entities with similar properties and behavior together for the purpose of identifying or classifying them for later use. It could be applied to provide better decision support. For example in computer science email classification is one application in sorting out emails as Spam, promotions or normal emails, based on the similarities shared among the other classes.

Both inductive and deductive methods could be applied in classification research method. We could use classification to prove whether a theorem is applicable for a particular area of interest or to make a general conclusion based a result obtained from sample data.

In our view, classification method is related to post-positivism and pragmatism philosophical approaches. Depending on the kind of research the classification method is used, whether to find out the existing truth as targeted by post-positivism or exploring what is important, practical and useful. It has also a strong affiliation to post-positivism, as researchers tries to test, verify or redefine a theory based on the result obtained from classification research method. The method also doesn't give an absolute truth about the subject it researches; as different classification criterion could yield different outcomes. Its association with the philosophical approach of pragmatism could be mentioned as, when we try to research new knowledge from a data without prior hypothesis or assumptions and purely based on the results obtained from the research method. The research method could also fit in the Jarvinen's taxonomy of research methods as researchers could stress on the approach of building or evaluating artifacts.
Reference is Wang, H. Bin, Yang, H. L., Xu, Z. J., & Yuan, Z. (2010). A clustering algorithm use SOM and K-Means in Intrusion Detection. Proceedings of the International Conference on E-Business and E-Government, ICEE 2010, (2007), 1281 – 1284. DOI is 10.1109/ICEE.2010.327.

Archival science which is the father of classification is the study and theory of building and curating archives, which are collections of recordings and data storage devices.

To build and curate an archive, one must acquire and evaluate recorded materials, and be able to access them later. To this end, archival science seeks to improve methods for appraising, storing, preserving, and cataloging recorded materials.

2. Data collection

There are several data collection methods such as experiments, surveys, interviews, questionnaires, observations, archives. For classification, numerical data and text files are normal. As for data sources, it can be listed as below

University:

Data and story library, Data lab

Normal:

Freebase, Numbrary, Many eyes, Infochimps, Swivel, Amazon public data sets, DBpedia, Wikipedia.

Getting it from API:

Plenty of sites and applications make their data freely available via APIs. Twitter has an API (duh). Google has lots of APIs. Yahoo does too.

In general, we will write a program using java or python to catch the data such as reviews of user, ranking in one platform.

3. Method Implementation

E-mail classification
Let's assume we are interested in classifying emails as Spam or non-Spam.
To start with the research, a huge email dataset with emails labeled as Spam and non-Spam is needed. The data could be collected from an email service provider, usually the provider who's interested in utilizing the result. Then depending on the size of the data, if it's too large to process, random sampling could be used to determine which emails need to be included in the sample. After applying the classification method it could produce a knowledge which could be used to identify and classify incoming email messages as Spam or non - Spam.

In our opinion classification method is useful in our discipline in classifying entities with similar properties/patterns into clusters or groups, to further research/investigate them in order to generate new knowledge or to verify existing ones.
The reliability and validity of a research using classification research method depends on different factors but the main factor which determines the outcome is parameter settings. It could also determine how closely or distantly related entities should be classified together. Therefore by changing the parameters setting we could get different results that needs to be put into consideration while using classification research method. Finally, we couldn't identify any ethical issues needs to be considered while implementing classification research method.