CHAPTER 1: INTRODUCTION

This chapter presents a general overview of this thesis. Section 1.1 and 1.2 describes the motivations and problems that need to be solved. Sections 1.3 provides the objectives of the work. Section 1.4 provides the hypothesis and Section 1.5 and 1.6 describe the research methodology and overview of the thesis. Finally, Section 1.7 attempts to indicate the contribution of this work.

1.1 Motivation

On arrival of low-priced high-speed connection to the home and the availability of budget smartphone, more people become connected to the internet and use web application on regular basis. The web application now gives more attention to their users, lowering barriers for inexperienced people to participate in web content creation. A special kind of web application that allows its users to create content put a label on the content with keywords and share them with other users is called Social Tagging System (STS for short). The key words attached are called tags, they are freely chosen by users and have a purpose as a classification and simple search. Some examples of STS are websites like Delicious[1], RainDrop[2], Lastfm[3], BibSonomy[4], and CiteULike[5]. Delicious and RainDrop allow the sharing of URL bookmarks, Lastfm the sharing of music, BibSonomy, and CiteULike the sharing of scientific literature references.

Today, STS becomes very popular, millions of users join the systems and share their contents. These enormous number of users floods STS with contents and tags in an unrestrained way in that threatening the capability of the system for relevant content retrieval and information sharing. This is known as information overload problem, this makes STS less effective since too many nonrelevant information hinders the user to get the relevant one. Recommender Systems (RS) is a successful method for overcoming information overload problem by filtering the relevant contents over the non-relevant one that continuously increase

[1] https://del.icio.us/about
[2] https://raindrop.io/
[3] https://www.last.fm/about
[4] https://www.bibsonomy.org/
[5] http://www.citeulike.org/

as more and more content become available online [1]. STS needs to implement RS to help its users get more valuable contents and have a better experience with the system.

To find a relevant content for its users, STS can apply a method used in information retrieval field, that is a search engine. The drawback is that the user must supplies keyword as an input query before the system can provide contents that match the query. Unfortunately, the user of STS usually does not provide keywords query, therefore another method must be used to automatically provide relevant content for its users.

The exposure of contents and tags to users set up an essential foundation for communication and sharing. It facilitates more cooperation and contribution to the creation of a collaborative classification system called folksonomies. This classification is created and maintained by ordinary users. This makes the tagging process more personal for each user and can be an essential information to learn user's taste. The tags used by the STS user can be used to create user's profile [2] - a collection of user's preferences. Thus, RS can find resources that similar with user's profile. For example, a user that frequently used word "robotic", "AI", "technology" as his/her tags presumably would like resources that tagged with "robotic" too.

Besides using folksonomy at its core, STS also employs social network features, allowing its users to create a relationship with other users in the form of friend relationship. According to [3], social network information can be utilized to learn the user preferences. A friend of user might want to use his/her resources. This additional information can help RS to produce better recommendations.

The data heterogeneity owned by STS create a new challenge for RS [4], [5]. Most present RS is intended for specific area and application, utilizing a solitary sort individual information and without expressly tending to the heterogeneity of the current individual data. Data heterogeneity can be recognized in any of the backbones of RS: user resource model and resource rank algorithm.

1.2 Problems and Research Question

In [5], the recommendation problem is defined as follows. Supposed U is the set of all users and R is the set of all possible resources that can be recommended, and function f is a utility function that measures the usefulness of resource r to user u, i.e. $f: U\ x\ R \rightarrow S$, where S is the amount or score (usually in real number). Then for each user $u \in U$, select resource $r' \in R$ that maximizes the user's utility and be formulated below:

$$\forall u \in U, r'_u = \max_{r \in R} f(u,r) \qquad 1.1$$

The finding of applicable function f and the appropriate representation for user and resource are some of the subjects in recommender system research.

In [6], a content-based approach is used to propose resource recommendation for a user in STS. They experimented with some weighting scheme ranging from the simple term frequency (TF) to inverse document frequency (IDF) and Okapi BM25. A dataset crawled from Delicious, (bookmark sharing) and Lastfm (music sharing) are used for their analysis purposes. From their experiments, it is reported that weighting scheme Okapi BM25 and TF-IDF performed best using cosine similarity. Another method based on graph ranking is reported by [7]. They collected friendship data and sharing data from Lastfm. They ran the algorithm with and without friendship (social network) data to observe the difference. They reported that by using additional social network data the accuracy achieved higher.

The method in [8] uses combination of content-based method and graph ranking-based method to generate resource recommendation for the user. They use tags as the source of user resource representation and utilizing simple multiplication between graph similarity and cosine similarity to rank resources. This combination is used as the baseline for their group aggregation recommendation. A dataset of Delicious and Lastfm provided in [4] are used to test their method.

This thesis attempts to develop a hybrid recommender system, a combination of content-based, graph rank-based and collaborative filtering approach. The result are compared to the result of other approaches as reported in [8], including pure content-based [6] and pure graph rank-based [7]. The work in [8] combines a content-based and random walk with restart, whereas in this hybrid additional collaborative filtering approach also used. In [7], they use an only random walk with restart for their recommender. To capture the user's tastes and preferences better hybrid recommenders can use more data available to them. The current recommender accuracy is still considered low, around 0.1 [8], and this gives us room for an improvement. So, the research question is "Can additional social network information be utilized by hybrid recommender that combines content-based, graph rank-based and collaborative method enhance the recommendation accuracy?"

1.3 Objectives

The objectives of this thesis are:

1. To combine a content-based, collaborative filtering and random walk with restart method into hybrid recommender and find values of the parameter to control the contribution of each method.
2. To perform an evaluation on proposed hybrid method and examine the result of the previous method.

1.4 Hypothesis

Earlier researchers have built hybrid recommender systems for social tagging system using a combination of content-based and collaborative filtering that work solely on the input of user-resource-tag information [9], [10]. Besides those data, social tagging system also has user-user relation information. However, content-based and collaborative filtering can not utilize this kind of data.

Graph ranking-based recommendation technique works on the network of users and resources as its input. Random Walk with Restart is a popular example of this technique, some experiments are reported that they have good performance by [7] and [11]. The hypothesis of this thesis which incorporating user-user relation network and add random walk with restart method into hybrid recommender system will increase the accuracy of recommendation.

1.5 Research Methodology

The steps to conduct the research methodology in this thesis are described as follows:

a. Problem identification, the purpose of this step is to identify potential improvement on recommending a resource on the social tagging system.
b. Model design, this step describes the design of recommender system in social tagging system used in this thesis. Parts of model design activity are:
    - Understanding dataset. In this thesis, a dataset is used from social bookmarking service Delicious and music sharing service Lastfm. These datasets are prepared and made available online by [4].
    - Content-based similarity. This step describes the content-based method for creating user resource profile and measuring similarity.
    - Graph rank-based. This step describes representation in the graph and random walk with restart algorithm for measuring similarity.
    - Collaborative for neighborhood forming. This step describes k-nearest algorithm to find similar users in the collaborative system and collect candidate resources.
    - Recommendation model. This step describes how to select a resource to recommend from a list of candidates.
c. Hypothesis definition and experiment design. This step states the hypothesis clearly and provide the design of the experiment to prove the hypothesis.
d. Implementation, this step is to build software of recommender system based on technique and algorithm explained in the model design.
e. Experiment. In this step, some scenarios of the experiment are conducted and the results are gathered to be processed later in the next step.

f. Result analysis. In this step, the experiment results are evaluated and detail analysis regarding the outcomes are performed.

1.6 Thesis Overview

This thesis presents how additional social network data can help recommender system enhance their accuracy. The organization of this thesis is as follows:

a. Chapter 1 Introduction. The motivation, problem and research question of the research is described, as well as the objective, hypothesis and research methodology used.
b. Chapter 2 Literature Review. A survey of some related work and background literature on recommender system and their evaluation.
c. Chapter 3 Research Methodology. This chapter describes research objective and detail description of the methodology used to reach the objective.
d. Chapter 4 Experiment and Analysis. In this chapter, the purpose and scenario of the experiment are described. The result obtained from running the experiment are collected and detail analysis regarding those results are described.
e. Chapter 5 Conclusion. This chapter provides the conclusion of this research, answers to the research questions and suggestions to further this research.

1.7 Contribution

The main contribution of this thesis is a hybrid recommender that combines a random walk with restart, content-based method, and collaborative filtering method in a way the contribution of each method can be controlled by the parameter.