**Thesis Summary**

**Semantic Path Based Personalized Recommendation System (SemRec) in Weighted Heterogeneous Information Network (WHIN)**

1. **Motivation**

The main disadvantage of the traditional collaborative filtering recommendation system is its scalability. This might be slow, extremely time-consuming, or unable to execute, as the number of tens of millions data.

Therefore, we want to compare between traditional collaborative filtering method (data storage by Matrix) and graph - based collaborative filtering method (data storage by Graph) in case of the huge datasets to prove the scalability of the graph which can replace the traditional method in the future.

1. **The Core Knowledge**

In this thesis, we only focus on the two similarity measure ways in User – based collaborative filtering recommendation system include Cosine similarity for storing data by Matrix and PathSim (Path-based similarity) for storing data by Graph.

* **Cosine Similarity**:

We have:

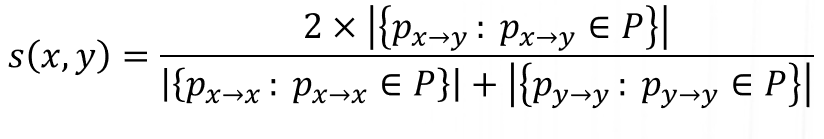
The reason data storage by Matrix is to perform the users as vectors. Therefore, we can easily apply the Cosine formula to calculate the similarity between users.

* + If the angle between two vectors is 0 degree,
  + If the angle between two vectors is 90 degree,
  + If the angle between two vectors is 180 degree,

A value close to 1 indicates similarity, while a value less than 0 indicates just the opposite. Therefore, we can see:

With represents the users irrelevant or less relevant with the target user. Hence, we call these values is the negative effect. Because it will negatively affect the recommended results (the accuracy will be reduced).

* **PathSim (Path-based Similarity)**

****

We have:

**The important concepts:**

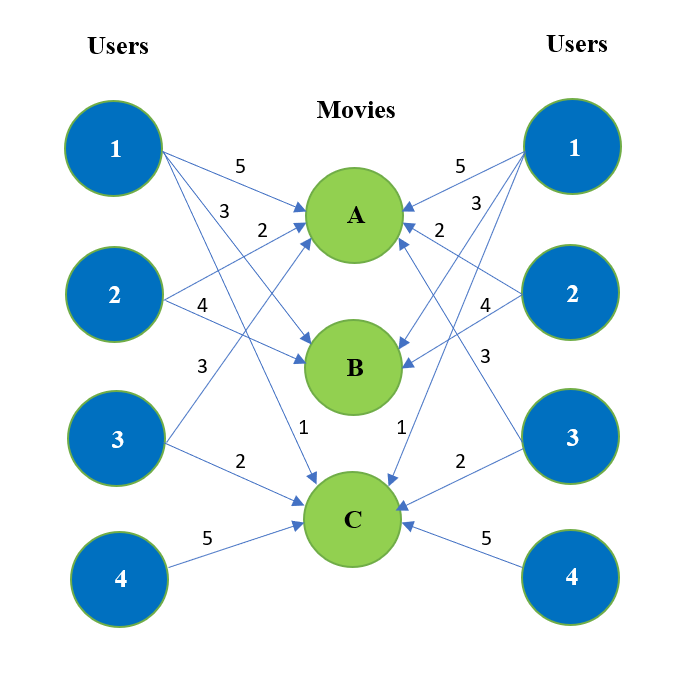
* **Path count**: the number of path instances between two nodes.
* **Symmetric meta path**: a meta-path is symmetric if the relation R is symmetric

Ex: if a relation exists from type A to type B, denoted as  
, and the inverse relation for. If the two types are the same, R is symmetric; otherwise, it is not symmetric. It means if it is symmetric, .

The below are three examples about symmetric meta path:

m1: King PersonQueen

m2: King Royal Queen

m3: King Country Queen

With:

* PI: Path Instances
* The rating score is the corresponding number of path instances between 2 nodes

1. **Experiment**
   1. **Datasets:** we use 2 datasets include:
      * **Movielens – 100k**: 100,000 ratings (1-5) for 1682 movies by 943 users (100K dataset which we used it in the first step of our experiments). Each user has rated at least 20 movies.
      * **Movielens – 1M**: 1 million ratings (1-5) for 3900 movies by 6040 users
   2. **Criteria:** In this thesis, we don’t test the executed memory because we can configure the executed memory in Neo4j.
      * **Accuracy:** We split the u data into a training set and a test set with exactly 10 ratings per user in the test set, include in 2 smaller datasets: ua.base, ua.test, ub.base, and ub.test. The sets ua.test and ub.test are disjoint.
      * **Time:** We test directly on the two original datasets
   3. **Environment:**
      * Processor: Intel Core – i7

Memory: 8GB 1866MHz LPDDR3

Storage: 512GB SSD

Operation System: Windows 10

Language: Python

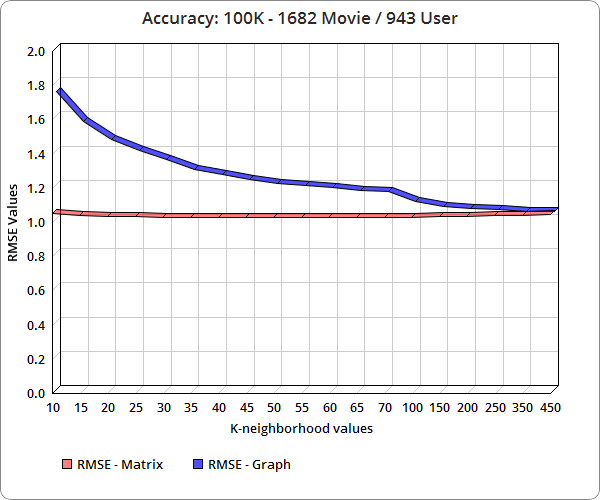
Tools: SQLite, and Neo4j Desktop (free version)

* 1. **Results**

**The RMSE accuracy is tested on MovieLens 100k dataset**

|  |  |  |
| --- | --- | --- |
| K-neighborhood values | RMSE-Matrix | RMSE-Graph |
| 10 | 1.04861122 | 1.742388 |
| 15 | 1.037894548 | 1.565377 |
| 20 | 1.032763404 | 1.460149 |
| 25 | 1.030238442 | 1.395637 |
| 30 | 1.028444872 | 1.340109 |
| 35 | 1.027688872 | 1.286397 |
| 40 | 1.027760522 | 1.255153 |
| 45 | 1.02776354 | 1.225625 |
| 50 | 1.027433409 | 1.204477 |
| 55 | 1.027637549 | 1.190593 |
| 60 | 1.028241791 | 1.177809 |
| 65 | 1.0286041 | 1.163393 |
| 70 | 1.028804371 | 1.152236 |
| 100 | 1.029211123 | 1.094498 |
| 150 | 1.029826424 | 1.065008 |
| 200 | 1.032013414 | 1.053908 |
| 250 | 1.036295234 | 1.047114 |
| 350 | 1.039103812 | 1.03857 |
| 450 | 1.041223454 | 1.036322 |

Let’s see result in graph below:



**Accuracy in 100k MovieLens dataset**

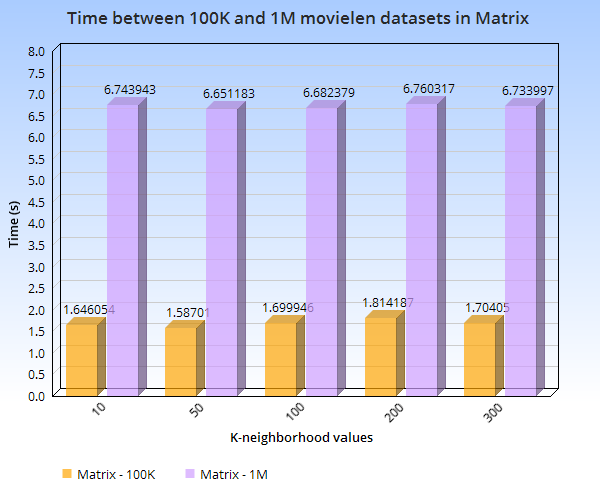
**Accuracy:** The accuracy of the traditional recommendation system is highest at the position k = 50 (the smallest value of RMSE is 1.027433409). If the out of that scope is the accuracy will be reduced. For the graph-based recommendation system, k is directly proportional to the accuracy. Therefore, the graph-based recommendation system will have the advantage of a quite high accuracy if that is a huge dataset (big data) and k gets the big value.

**K-neighborhood:** The size of the neighborhood has a considerable impact on the prediction quality. To determine the sensitivity of this parameter, we performed an experiment where we varied the number of neighbors to be used to find out the position k which the two recommendation systems will have the relatively highest accuracy

Let’s see result in MovieLens datasets:

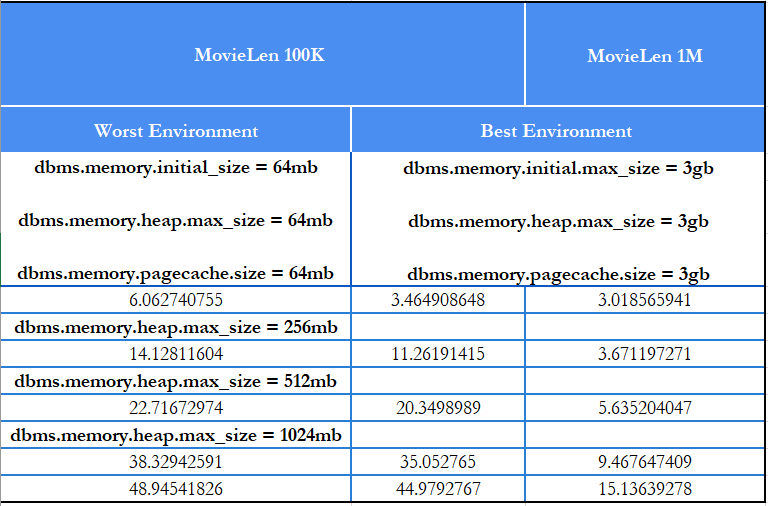
**Table values of the Time between 100K and 1M movielen dataset in Matrix**

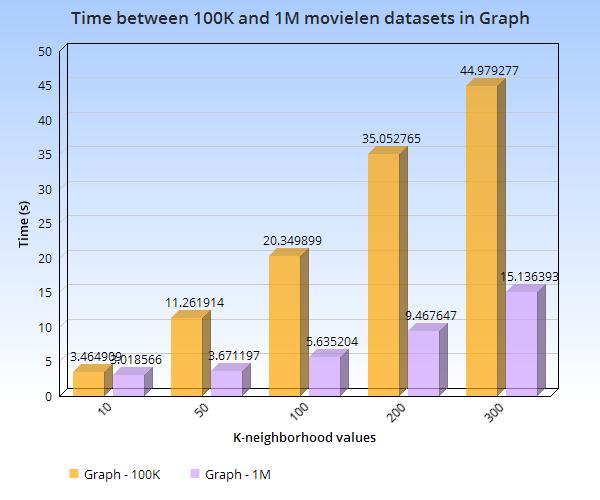
|  |  |  |  |
| --- | --- | --- | --- |
| **K neighbor** | **MovieLen 100K** | | **MovieLen 1M** |
| **Time** | | |
| 10 | 1.646053839 | 6.743942642 | |
| 50 | 1.587010145 | 6.65118261 | |
| 100 | 1.699946295 | 6.682378769 | |
| 200 | 1.814186573 | 6.760316515 | |
| 300 | 1.704050207 | 6.733996997 | |

****

**Chart of The Time between 100K and 1M movielen dataset in Matrix**

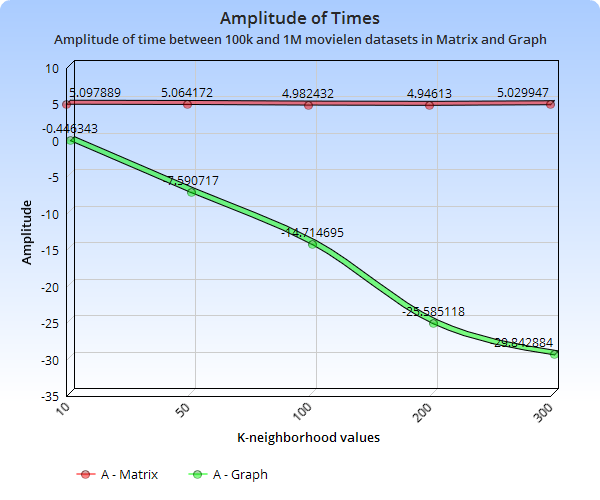
**Table values the Time between 100K and 1M movielen dataset in Graph**

****

****

**Chart of the Time between 100K and 1M movielen dataset in Graph**

Figure below is Amplitude of time between 100K and 1M movielen datasets in Matrix and Graph.

****

As we can see, the different datasets are the differently recommended time. It usually depends on the number of ratings, the sparsity of the rating values, e.g. After the experiment on two above datasets, the recommended time of the graph-based RS is always faster than the traditional RS.

1. **Conclusion**

As we can see the above results, Graph-based collaborative filtering method can execute better than the traditional collaborative filtering method both accuracy and time in the huge datasets.