Scientific Report

First name / Family name	BIDYUT KUMAR / PATRA
Nationality	INDIAN
Name of the *Host Organisation*	VTT Technical Research Centre of Finland, Finland.
First Name / family name of the *Scientific Coordinator*	Raimo / Launonen
Period of the fellowship	01/05/2013 to 31/10/2014

I – SCIENTIFIC ACTIVITY DURING YOUR FELLOWSHIP

Recommender System (RS) techniques have been successfully used to help people cope with information overload problem and they have been established as an integral part of e-business domain over last decades. The primary task of a recommender system is to provide personalized suggestions for products or items to individual user filtering through large product or item space. Many recommender system algorithms have been developed in various applications such as e-commerce, digital library, electronic media, on-line advertising, etc. These algorithms can be categorized into two major classes, *viz. content based filtering* and *collaborative filtering*.

Collaborative Filtering (CF) is the most successful and widely used recommendation system. There are two main approaches for recommending items in CF category, *viz. neighbourhood based CF* and *model based CF*.

Generally, neighborhood based CF uses a similarity measure for finding neighbors of an active user. Traditional similarity measures such as pearson correlation coefficient, cosine similarity are frequently used for this purpose. However, correlation based measures perform poorly if there are no sufficient numbers of co-rated items in a given rating data. As a result, correlation based measure and its variants are not suitable in a sparse data.

- In this research programme, we develop a *novel approach for finding similarity between a pair of users in sparse data*. Unlike existing measures, proposed measure can compute similarity between two users in the absence of co-rated items and the approach uses all ratings made by a pair of users.
- As we know the traditional similarity measures are not effective for *user cold-start problem* (user with few ratings) in CF. We develop a RS which is effective in this scenario.

Diversity in RS: Classical algorithms in recommender system mainly emphasize on recommendation accuracy in order to match individual user's past profile. However, recent study shows that novelty and diversity in recommendations are equally important factors from both user and business view point. In this research programme, we introduce a knowledge reuse framework to increase novelty and diversity in the recommended items of individual users while compromising very little recommendation accuracy. The proposed framework uses features information which have already been extracted by an existing collaborative filtering. We do not access original rating data in order to improve novelty and diversity in recommendation. Having applied a model based CF such as Regularized SVD for predicting ratings of non-rated items of an active user, we employ clustering technique to the items which received predicted rating more than a predefined threshold. Finally, recommended list is generated by selecting items from clusters to provide maximum diversity in the list.

Incremental Clustering: Recent advances in storage, network and computer technology result in increasing the size of database day by day. To analyze these dynamic datasets, Incremental clustering is found to be a good tool instead of classical clustering methods.

In this research programme, we develop an incremental clustering in dynamic scenarios such as purchasing patterns of a new product, financial transactions, early detection of epidemic, etc.

II – PUBLICATION(S) DURING YOUR FELLOWSHIP

1. **Bidyut Kr. Patra**, Raimo Launonen, Ollikainen Ville, Sukumar Nandi. *A new similarity measure using Bhattacharyya coefficent for collaborative filtering in sparse data.* **Knowledge Based System, March, 2014. (Pending)**

2. **Bidyut Kr. Patra** and Sukumar Nandi. *Effective data summarization for hierarchical clustering in large datasets*, Knowledge and Information System (KAIS), 2014. (**Accepted)**

Abstract: Cluster analysis in a large dataset is an interesting challenge in many fields of Science and Engineering. One important clustering approach is hierarchical clustering, which outputs hierarchical (nested) structures of a given dataset. The single-link is a distance-based hierarchical clustering method, which can find non-convex (arbitrary)-shaped clusters in a dataset. However, this method cannot be used for clustering large dataset as this method either keeps entire dataset in main memory or scans dataset multiple times from secondary memory of the machine. Both of them are potentially severe problems for cluster analysis in large datasets. One remedy for both problems is to create a summary of a given dataset efficiently, and the summary is subsequently used to speed up clustering methods in large datasets. In this paper, we propose a summarization scheme termed *data sphere (ds)* to speed up single-link clustering method in large datasets. The *ds* utilizes sequential leaders clustering method to collect important statistics of a given dataset. The single-link method is modified to work with *ds*. Modified clustering method is termed as *summarized single-link (SSL)*. The SSL method is considerably faster than the single-link method applied directly to the dataset, and clustering results produced by SSL method are close to the clustering results produced by single-link method. The SSL method outperforms single-link using data bubble (summarization scheme) both in terms of clustering accuracy and computation time. To speed up proposed summarization scheme, a technique is introduced to reduce a large number of distance computations in leaders method. Experimental studies demonstrate effectiveness of the proposed summarization scheme for large datasets.

3. **Bidyut Kr. Patra**, Raimo Launonen, Ollikainen Ville, Sukumar Nandi. *Exploiting Bhattacharyya similarity measure to diminish user cold-start problem in sparse data.*, The 17th International Conference on Discovery Science (DS-2014), University of Ljubljana, Ljubljana, Slovenia, October, 2014. **(Accepted)**

Abstract: Collaborative Filtering (CF) is one of the most successful approaches for personalized product recommendations. Neighborhood based collaborative filtering is an important class of CF, which is simple and efficient product recommender system widely used in commercial domain. However, neighbourhood based CF suffers from *user cold-start* problem. This problem becomes severe when neighborhood based CF is used in sparse rating data. In this paper, we propose an effective approach for similarity measure to address user cold-start problem in sparse rating dataset. Our proposed approach can find neighbors in the absence of co-rated items unlike existing measures. To show the effectiveness of this measure under cold-start scenario, we experimented with real rating datasets. Experimental results show that our approach based CF outperforms state-of-the art measures based CFs for cold-start problem.

4. **Bidyut Kr. Patra**, Ollikainen Ville, Raimo Launonen, Sukumar Nandi and Korra Sathya Babu *A distance based incremental clustering for mining clusters of arbitrary shapes*, The Fifth International Conference on Pattern Recognition and Machine Intelligence (PReMI 2013), Indian Statistical Institute, Kolkata, India, December, 2013. **(Accepted)**

Abstract: Clustering has been recognized as one of the important tasks in data mining. One important class of clustering is distance based method. To reduce the computational and storage burden of the classical clustering methods, many distance based hybrid clustering methods have been proposed. However, these methods are not suitable for cluster analysis in dynamic environment where underlying data distribution and subsequently clustering structures change over time. In this paper, we propose a distance based incremental clustering method, which can find arbitrary shaped clusters in fast changing dynamic scenarios. Our proposed method is based on recently proposed *al*-SL method, which can successfully be applied to large static datasets. In the incremental version of the *al*-SL (termed as *IncrementalSL*), we exploit important characteristics of *al*-SL method to handle frequent updates of patterns to the given dataset. The IncrementalSL method can produce exactly same clustering results as produced by the *al*-SL method. To show the effectiveness of the IncrementalSL in dynamically changing database, we experimented with one synthetic and one real world datasets.

5. Apurva Pathak and **Bidyut Kr. Patra**. *A knowledge resuse framework for improving novelty and diversity in recommendation,* **Communicated to 2nd ACM IKDD Conference on Data Science (CoDS 2014), Bangalore, India, March, 2015 (Pending)**

III – ATTENDED SEMINARS, WORKHOPS, CONFERENCES

1. The 17th International Conference on Discovery Science (DS-2014), **University of Ljubljana, Ljubljana, Slovenia, October, 2014**.

2. Mini-workshop on Multiple Dimensions of Relevance, **CWI, Netherlands, October 13, 2014.**

3. ABCDE Seminar III, **Athens, Greece, October, 2013**.

IV – RESEARCH EXCHANGE PROGRAMME (REP)

REP 1:

Organization: SICS Swedish ICT AB
Country: Sweden
Local Scientific Coordinator: Dr. Anders Holst, Associate Professor.
Duration: October 21-October 27, 2013

In my short visit at SICS , I explored a topic called "Outlier Detection", which is related to my Ph.D. topic. Local coordinator Dr. Anders Holst explained their project on titled Anomaly Detection for Train Maintenance. In the project, they used a statistical approach for early detection of some serious events in train operation. I presented my research plan and part of Ph.D. work in an open seminar at SICS, Sweden.

REP 2:

Organization: Universitat Politècnica de València (UPV)
Department: Dpto. Sistemas Informáticos y Computación (DSIC)
Country: Spain
Local Scientific Coordinator: Dr. Paolo Rosso, Associate Professor.
Duration: November 19, 2013-November 29, 2013.

My visit at UPV was effective in the sense that I was given chance to explore natural language processing algorithms with clustering techniques to solve the problem of opinion spam. Because I stayed at UPV for a long time, I did some effective work towards the opinion spam. Here also I presented my research plan in a departmental seminar.

REP 3:

Organization: Centrum Wiskunde & Informatica (CWI), Amsterdam
Department: Information Access
Country: The Netherlands.
Local Scientific Coordinator: Prof. Arjen P. De Vries, Full Professor.
Duration: October 13-October 20, 2014

Just before the completion of my 18 months ERCIM fellowship period, I planned to visit CWI, Amsterdam. I am really thankful to my host Prof. Arjen for accepting my request in short time. In this short visit, I attended a mini workshop on Multiple Dimensions of Relevance at CWI.
Here, I discussed my current research topic "Recommender System" with more than two researchers apart from Prof. Arjen. Discussion with them is very effective for my current and future research.