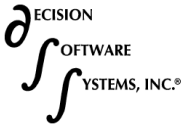

((U//FOUO) Recommender Systems

(U) From: Decision Software Systems Inc
(U) 176 Hart Ave, Doylestown PA 18901
(U) www.decision-software.com
(U) 800-682-0794 dssinc@decision-software.com

1 Table of Contents

Contents

2 (Abstract

The proposal provides baseline solution for incorporating into an integrated recommender system (RS). The Big Data Recommender will operate in the HDFS/Hadoop, MapReduce using Mahout Technologies working with Amazon Public DataSets in an EC2 Cloud Computing Platform. A front end user interface will allow the operator to dynamically customize various parameters of the recommender system. This will include integrated technologies:

1. Data Capture and Curation
2. User Interface Preferences
3. Data Analysis of Large DataSet(s) with Sparse information on the Target Behaviors
4. Predicative Analysis based on Default or User Specified parameters of the Target
5. Analytics based on incomplete data
6. Feedback Functions to be used improve future automatic recommendations and/or predictions.

3 (U) Executive Summary

This proposal provides for the development of a Recommender System based on the Apache, Hadoop/HDFS/MapReduce and Mahout Framework. The Cost Proposal provides a summary of the Deliverables, Invoicing and Project Schedule based on deliverables at 30 days, 60 days and final evaluation. The proposed solution is for a Apache Framework Mahout Recommender System based (U) State of Art

The proposed Recommender System is based current BigData Technologies using AWS/EC2, Eclipse/Java/Maven IDE, Hadoop File System (HDFS). In a Multi-Node Cluster using a Recommender System based on mahout technologies.

4 Description of the Recommender Prototype

(U) The system will provide an integrated solution for processing AWS Dataset within the EC2 platform. Mahout is a Machine Learning System the will generate Recommendations based on Collaborative Filtering of user-based and item-based filter results obtained from the underlying MapReduce/HDFS (key,value) pairs results merged into sets of user and item filters and inputs to the Recommender. The solution provided here is an integration of the these technologies via a set of User Interfaces Menu to Configure the parameters to effectively manage an end to end solution from DataSets to Recommendations. This approach integrates into a single Recommender, the Functions of Track 7.11 Topics 1-6 Requirements. Once the Recommender completes the Training cycle. The Recommender can operate in manual or automatic mode. In automatic mode, a set Recommenders can be setup in Parallel on EC2 using different configuration parameters to produce sets of Recommender Results. After generation of the results, further processing can be used for Feedback Analysis to improve the Recommender Engine Algorithms. The Recommender will use an internal database to manage

configuration, metadata formats for input and output results, feedback parameters and algorithm processing parameters.

5 Facilities, Software Frame Work/ Data Sets

Our Local Facilities are to be utilized for the purposes of Development of the Proposed Recommender System. The Datasets to be used will be provided using an Amazon EC2 Account and the DataSets Identified in the proposal. The Recommender System product will be based on Eclipse/Java/Mahout Software that executes in Amazon Web Services AWS Elastic Cluster 2 (EC2).

For this project, the Recommender System will utilize the Apache/Hadoop/Mahout Framework executing on AWS EC2. The framework provides for a scalable Machine Learning System with Computing Resources to process Large DataSets using Clustering/Grouping, Classification Training Sessions. Mahout uses MapReduce for the processing of (Key,Value) Pairs to capture and Map DataSet(s) Parameters and Reduce to merge values according to specifications of the Mahout Algorithms. Mahout uses the MapReduce Results as Inputs to the Recommender System.

Mahout technology uses a Collaborative filtering Framework. The Recommender Engine Algorithms are based on user-based and item-based filters in which Software written in Mahout associates sets of user-based behaviors and item-based datasets are statistically (Behavioral Statistics) associated to produce a set of recommendations based on the type of analysis applied to the large dataset(s).

6 Technical Data Computer Software

No Technical Data has been identified that will limit government usage rights. The deliverables are base on Open Source Apache Framework parameters.

7 Sub-Contracts or Relevant Collaborations

Due to the constraints of time and budget no subcontracts or collaborations will be utilized in the initial phase of the project. If a follow on project comes about, Identification of additional resources will be identified.

8 (U) Deliverables

The Cost Proposal Identifies the Deliverables.

The deliverables shall be in compliance with bid requirements.

- Source code of all non-commercial software used in the prototype (Mahout Recommender Developed in Java/Apache/Hadoop/MapReduce/Mahout)
- Installation instructions
- Software and Documentation in soft form on physical media of CD or DVD.

- Monthly Status Reports and Invoices (Three)

9 Licensing

No additional licensing has been identified. The deliverable products are based on open source products licensing constraints.

10 Program Proposal

As identified in the RFP, the approach is to develop the Recommender System based on existing state of art core technologies and platforms. Hence to support the requirements of the RFP, the approach is to build a solution based on existing platform technologies (Amazon/EC2, Rhat Linux/Eclipse/Java/Hadoop/Mapreduce/r/Mahout Recommender Technologies).

Software will be developed in the Eclipse Juno IDE using Java,Maven,HDFS,MapReduce and Mahout as the programming language. Using http://svn.apach.org/repos/asf as the download source for the Mahout Development platform.

Our Approach is to build a Mahout Based System to provide a Recommender System that integrates the functionality identified in the bid. The objective is to provide a modular approach that provide that address the requirements of the track.

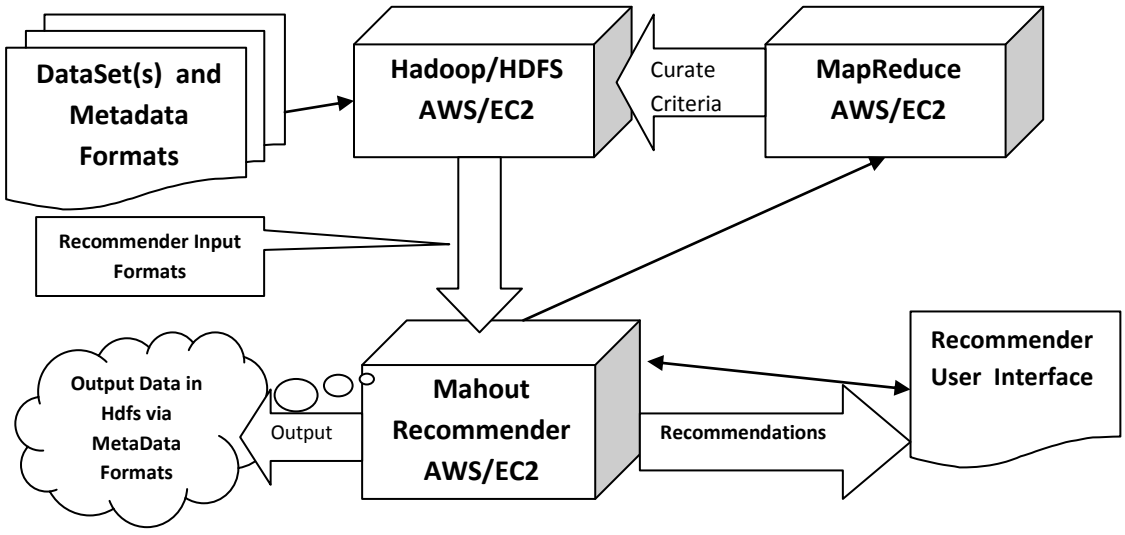

Figure 1 Top Level Mahout Recommender

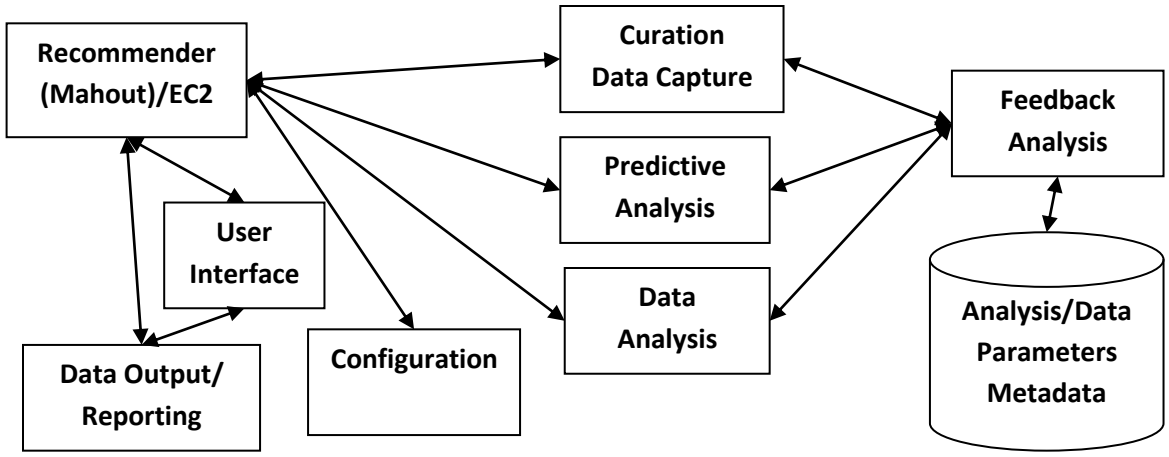

Figure 2 Recommender Block Diagram

10.1 Recommender System

The proposed Integrated Recommender System will execute in AWS EC2. The baseline system will interface with AWS DataSet(s) aws/amazon.com/publicdatasets/. When the RS starts via the User Interface one can select a set of options of how to import the dataset(s). The Datasets are to be installed or pre-installed into a Hadoop/HDFS file system. The Recommender inputs will be a set of MetaData Format Options to be used to Curate the Dataset(s). The RS will use an internal Database for management of configuration parameters, Feedback Parameter, Data Analysis MetaData Formats, Feedback/Predictive/Data Analysis Parameters and MetaData Formats for Reporting and generation output data into standardized formats compatible with HDFS/EC2 Data (XML Based MetaData Formats). The System is to be capable to be configured for Manual or Automatic Operation. The RS when configured can become a middleware component that executes in the background to process Dataset(s) and output based on formats selected. Within a Large network the RS can be configured to execute automatically as a set of parallel processes each with distinct input/outputs and configuration parameters.

(U//FOUO) In Manual Mode the User can select configuration options and define desired predictive output for review in a report or generate output Dataset(s) and MetaData Formats.

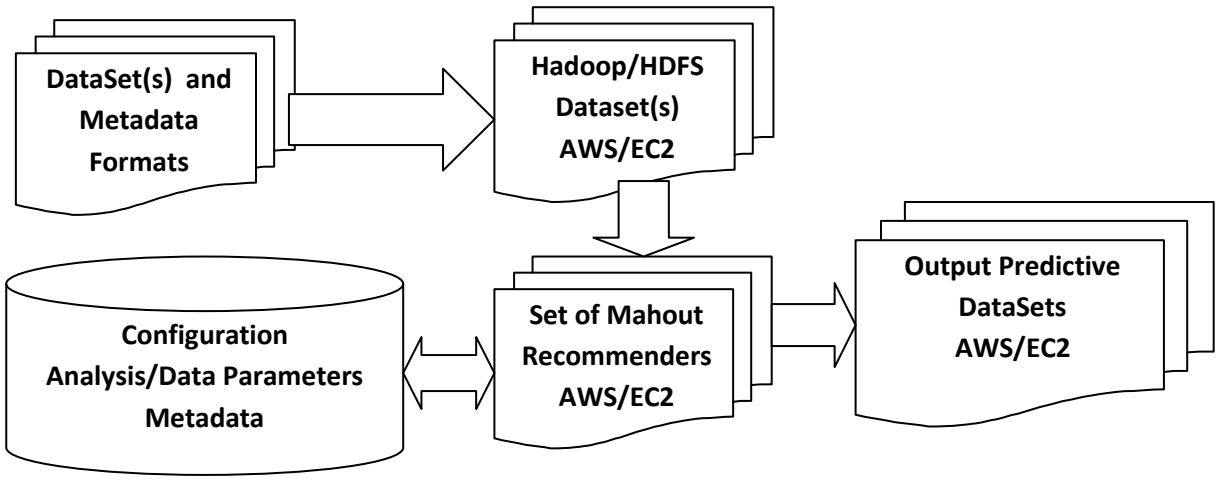

Figure 3 Recommender Automatic Operational Mode

Note: These Specifications identify a full production version of a recommender system. This level of effort is out of scope for a baseline product. In the initial system only a small set of configuration parameters will be available. If a follow on is provided additional capabilities will be incorporated.

10.2 Develop a Mahout-Recommender System based on Requirements of Topics.

(U//FOUO) Objectives for the prototype is to develop in a phased approach a Mahout based Recommender System developed in Eclipse/Hadoop/MapReduce/Mahoot with a simple user interface. Only if time allows will we consider upgrading the user interface to Ozone Widgets.

10.2.1 Establish Development Platform for Mahout Based Recommender

(U//FOUO) Initial effort is to setup the development platform(s) in Amazon/EC2, and a local PC/Rhat/Linux (Virtualized).

10.2.2 Topic 1 Requirements

(U//FOUO) The Mahout-Recommender will manage and process resulting data in a format that is usable to interface to other Recommender Systems and to other systems. The methods and mechanisms need to be adaptable and capable to accommodate new data types or process existing data identified. The RS will allow user to Select Dataset(s) to import into the Hadoop File System (hdfs) within the AWS EC2 platform. However in the initial prototype only the AWS public dataset(s) will be used.

10.2.3 Topic 2 Requirements

(U//FOUO) The Mahout Recommender requires input data on which it will base its recommendations. This data take a form call “preferences” in Mahout terms. The preferences associate user/item sets of filtered data. The input preferences can specified to the Recommender as manual user inputs, Configuration Parameters or a set of input file(s). The Recommender is to provide a mechanism for the user configure and install “preference” parameters

10.2.4 Topic 3 Requirements

(U//FOUO) The Mahout-Recommender is to make predictions and/or recommendations based on integrating and evaluating data from multiple disparate data sources. The mechanisms need to be able to incorporate data from new sources. The Mahout Recommender is to provide an indicator of the value of each data source in the generation of prediction and/or recommendation. A numeric data quality indicator is to be generated based on the threshold of user/item found in the data in a volume unit of input data processed. The user may configure the Recommender to specify data quality.

10.2.5 Topic 4 Requirements

(U//FOUO) The Mahout-Recommender is to predict the data the user is likely to need or want next. Predictive Analytics developed here shall provide objective evidence the proposed analytics is better that what is commonly used. The Recommender is to compute results of "preference" criteria applied to input data and generate a list of recommendations

10.2.6 Topic 5 Requirements

The Mahout-Recommender is to provide mechanisms to develop predictions and recommendations based on incomplete data and refine the predictions and recommendations as more data becomes available. The objective is not to produce wrong recommendations while producing recommendations of increasing value and correctness. The RS is use "preference" criteria to be applied to specific dataset(s) using collaborative filtering user-base and item-base filters. The Recommender Engine is to be setup to define threshold criteria based on predictive analytics.

10.2.7 Topic 6 Requirements

he Mahout-Recommender mechanisms are to identify and collect feedback and use the information to improve future automated recommendations and predictions. With respect to the Recommender Engine The Feed Back Parameters an be used to adjust coefficients of the algorithms to improve the Engine computational results like in a Neural Net, Feedback data can be used to change the threshold level on pass/fail criteria. Feedback can be used in the training phase of the Recommender also.

11 Personnel

All parties assigned to this effort are US Citizens. A dedicated Senior Software Engineer of Decision Software will be assigned to this project for the purposes of development of Mahout Recommender System Art Larson has over 30 years of working in the information technologies. With expertise in Unix/Eclipse/Java/Maven and Java Server Web Pages. He holds an MBA in Finance and MIS and a BS in Physics. In addition, He has expertise in Statistical analysis and a minor in Psychology. He has over 15 years of expertise in building Encryptors (Systems/Software), Unix/MLS, High Assurance, Comsec/Crypto, Database Systems (DBA,ETL,Data Modeling). The amount of effort is estimated to be provided about be about 500 Manhours of a Senior Technical Professional.

11.1 Personnel Resume – Art Larson

Mr. Larson has Extensive experience in software development. Expertise in Systems Engineering, Defense, Financial and Real Time systems. Programming expertise in Requirements Engineering, UNIX/C(++), Ada, Sybase/Oracle/Informix, X-Windows/Motif/UIL, Communications, Computer Security Engineering Object Oriented Design/Analysis, Client/Server

Computing, Database Systems development. Win/NT Systems Development MS VC++, GUIs, OOA/OOD. E-R Definitions, Modeling etc. Business and Management skills in Defining Organizational and Technical Solutions to business issues. Skilled at working in both engineering and data processing environments. Expertise in Marketing, Proposal Development, Studies and Training. Expertise in Web Development Technologies including ASP, Java, EDI and Middleware Database Development Projects for large scale operations.

- Extensive Real Time Operating Systems (Including writing operating systems for embedded applications)
- Expertise in INFOSEC Systems Engineering, EKMS Operations utilized in Red/Black Encryptors
- Expertise in real time embedded systems development. Worked with over 30+ assemblers
- UNIX/C/C++, Kernal Hacking, Device Drivers.
- Extensive Data Communications Background (AIN, IN1+, SS7, Telephony, Satellite, Many Special Military Communications Projects, Video on Demand Server for Formation).
- In Depth Background in security worked in all aspects of system development from special government code machines to the entire spectrum of the National Security Agency's Orange Book. This include Mult-Level Operating Systems, Trusted Network Communications, Database Security, Encryption Technology, Security Management etc.
- Extensive Background in all aspects of the full software life cycle development process.
- Worked on Win/7/Vista/XP Software Development Projects using Visual Studio and ASP.Net development of software in VC++, COM/DCOM, OPC (Power Plant Process Control System, Process Control System for Road construction industry and Commercial Mortgage Financial Management System)
- Extensive knowledge in all aspects of Object Oriented Design and Analysis Methods.
- Over Six years of software development for real time embedded Ada Projects.
- Around Seven years of Expertise in Data Base Systems, Oracle, Sybase, MS SQL Server, ErWin. Data Modeling, Design, Requirement and Performance Analysis, Business Process Analysis.
- Extensive expertise in the computer security industry. In the past worked on UNIX/MLS Systems (Multi-Level Security), Defense Systems Comsec/Crypto/Sigint e.g. KGV-84, Secure Communications (Milstar, GPS), Orange Book and Protection Profiles, Secure Communication (Milstar, GPS) etc.

Clearance: DoD Secret Comsec/Crypto

Education:

- *MBA in Management Information Systems and Finance, West Coast University, Los Angles CA., Minors in Computer Science including work in compiler design.*
- *B.S. Physics, Minors in Mathematics, Computer Science and Psychology., Temple University, Philadelphia. PA*

Technical Training:

- 2011 – SharePoint Programming and Administration Certificate Penn State Continuing Education Program.
- ASP.NET (VS2010, VB, C#, ASP.NET,) Penn State Continuing Education Certificate Program
- Java (Eclipse IDE, JSP, Servlets, Apache) Penn State Continuing Education Certificate Program
- Java Programming/Swing/XML/ODBC Temple University Continuing Education
- 1994 Seminar "Client/Server Computing"
- 1991 IEEE Neural Networks for Signal Processing Princeton NJ IEEE Conference on Neural Networks for Ocean Engineering Ninth Annual National Conference on Ada Technology
- 1989 Seventh Annual National Conference on Ada Technology/Object Oriented Programming
- 1988 11th National Computer Security Conference, Update Seminar on Computer Security
- 1987 Fifth National Conference on Ada, Update Seminars on OOD/OOA
- 1985 Second National Conference on Ada, training seminar on Ada Development by Richard Boltz and Dr. McKay IEEE Training Seminar on the use of "PROLOG"

Experience Profile:

- Extensive experience in Real Time Systems Programming for both Engineering and Data Processing environments.
- Worked with **UNIX/C/C(++)** Internals and **Ada** systems., Object Oriented Design/Analysis
- Expertise in software development methodologies for the life cycle maintenance of software.
- Excellent Technical Communication and Writing Skills
- Systems Programming of software such as operating systems design and compiler language development

- Communications software development and management of protocols, (X.25, OSI Standards, DECNET, Ethernet TCP/IP, IBM Bisync etc., Satellite and Packet Switching Networks.
- Database designs for Relational and Hierarchical DBMS (**MS SQL Server**, **Sybase**, **Oracle**, Powerbuilder, **Informix**, VAX/RDB, etc).
- Extensively used Sybase ISQL, DWB, dBlib and Remote/Stored Procedures, ORACLE
- Skilled at Hardware and Software Systems Applications
- Experienced with many of the software development documentation standards such as Mil-STD 2167A, 1679A 483/490, Orange Book and IEEE specifications.
- Extensively used Unix Utilities: LEX, YACC, AWK, SCCS, MAKE **X-Windows/Motif**.

Capabilities:

- Self Starting, goal oriented to meet the objective of the business in a productive manner.
- Systems Engineering Skills to work with a requirement and to translate the needs into technology.
- Works well in a team environment or in a Project Management role.
- Experienced in Systems and Application Software Development
- Marketing and Sales Support Skills (Established a Consulting Business Decision Software Systems Inc. have hired about 200+ Consultants over twenty years and generated a sales volume of over 40 million in this timeframe). Established direct vendor relationships with many major corporate accounts. During this period also established expertise in Contract Management , Bids/Proposals along with Customer Relationship Management (CRM).
- Involved with Technical Management and Project Leadership Roles for several clients including management of small teams of people for various technical projects.
- Can perform evaluations, studies and definition of technical requirements based on the needs of the business or engineering environment
- Embedded Radar Systems Software Development Applications for Air Defense Application.
- Financial Transaction Processing Systems.

12 Capabilities

Decision Software Systems is a small business concern with expertise in development of High Assurance, Multi-Level Security (MLS), Encryption, Communications, Java Development, BigData. See On-line brochure at www.decision-software.com. Our company has been in business since 1984. We have provided consulting services to various organizations over a 20 year period with over $40 million in sales volume during this time and management of over 250 labor resource for consulting assignments

13 Administrative

(U//FOUO) For the purpose of administration this program the Functions of Program Management , Technical Management and Product development will be integrated into the activities of a single person. The scope of the project with respect to Administration is not the large to require multiple personnel to manage these activities. Art Larson the President of Decision Software System will manage the roles of Program Management and Technical Development.

14 Meetings

All Meetings are to be held via teleconference using an Offerer provided phone number. For the final deliverables, an on-site meeting with the Contractor Representatives will be held at a Vendor's facility.

15 Appendix Recommender System Requirement Summary

According to the Request for Proposal, the Solicitation defines in summary the Tracks and Topics for the Bid Response. This Applies to the concepts of Recommender Systems (RS) for usage within Intelligence Analysis and Systems with a focus on:

- Data Capture and Curation
- Establish User Preferences
- Disparate Data Analysis
- Predictive Analytics
- Analytics Based on Incomplete and Evolving Data
- FeedBack Mechanisms

15.1 Objectives

Adapt/Extend/Apply Concept of Recommender System to Intelligence Analysis and Intelligence Systems with focus on:

- **Target Behaviors** – Analytics about targets (MetaData/Content) Recommend other facts to Analyst
- **System Behaviors** – Analytics to use information about system to Recommend Capabilities to Analyst
- **Analyst Behaviors** – Analytics on Analyst Behaviors to recommend additional workflows to them (Very Interesting Point possible link to SharePoint WorkFlow Systems as an example)
- **Customer Behaviors** – Analytics on customer activities to recommend reports and information to customer

15.2 Constraints

The following identifies the constraints on the proposed solution for bid:

- **DataSets** - Data Volume can be large. However data about individuals can be sparse, incomplete and dirty. The System should be able to use data from Cloud or Relational Database Datasets (Proprietary shall not be used).
- **Computing Resource** – The Recommender System Computations should consider optimum utilization of Computing Resources
- **Real Time Results** – The Recommender System needs to generate results in a near real-time mode.
- **Recommender Algorithms** - The Recommender System Algorithms used to generate results should be available to the user (implicit allow user to use default or select recommender system algorithms).
- **Computing Platform** – The Recommender System is to utilize Industry Standard Hardware/Software (Windows and/or Linux etc.

15.3 Topics Requirements

The following is the high level requirements for the proposed Recommender System (RS) based on Amazon/EC2, Hadoop/MapReduce/HDFS, mahout using Amazon DataSets.

15.3.1 Recommender System Data Capture and Curation

The RS is to corporate methods/mechanism to identify and capture data relevant to RS, Curate DataSet(s) into a format usable as input to the RS. The mechanism(s) need to be adaptable and/or adjustable as new dataset types become available.

15.3.2 Recommender System Establish User Preferences

(U//FOUO) The RS is to include methods/mechanism(s) to establish user preferences (User Interface with Ozone Widget only if time permits) otherwise use a simple console user interface. The purpose is to allow the user to adjust the operations of the recommender system.

15.3.3 Management of Disparate Data Analysis

The RS mechanisms are to make predictions and/or recommendations based on integrating and evaluating data from multiple disparate data sources. The mechanisms need to be able to incorporate data from new sources. The RS is to provide an indicator of the value of each data source in the generation of prediction and/or recommendation.

15.3.4 Predictive Analysis

The RS mechanisms are to predict the data the user is likely to need or want next. Predictive Analytics developed here shall provide objective evidence the proposed analytics is better that what is commonly used. If time allow incorporate utilization of Predict.org. The proposed Analytics of the Recommender System shall provide results showing technology used is better then existing capabilities. The RS solution should identify which predictive analytics work best under different conditions. Eg Social Network sites that recommend new friends, groups or connections based on social networks of user and other groups of people in groups related to the user's groups of interest.

15.3.5 Analytics Based on Incomplete Evolving Data

The RS is to provide mechanisms to develop predictions and recommendations based on incomplete data and refine the predictions and recommendations as more data becomes available. The objective is not to produce wrong recommendations while producing recommendations of increasing value and correctness.

15.3.6 Feedback Mechanism

The RS mechanisms are to identify and collect feedback and use the information to improve future automated recommendations and predictions.

16 References

1) Mahout in Action – Manning Press http://www.amazon.com/Mahout-Action-Sean-Owen/dp/1935182684
2) Apache Hadoop - http://hadoop.apache.org/
   a) Apache MapReduce - http://hadoop.apache.org/docs/r1.0.4/mapred_tutorial.html
   b) Apache Mahout - http://mahout.apache.org/
3) Amazon Web Services (AWS)
   a) Amazon Web Services - http://aws.amazon.com/console/

   i) Amazon Elastic Cloud (EC2) - http://aws.amazon.com/ec2/
   ii) Amazon Cloudfront Streams - http://aws.amazon.com/about-aws/newsletters/2010/01/30/january-2010/
  b) Amazon Web Services (Government) - http://aws.amazon.com/govcloud-us/
4) Eclipse Juno Integrated Development Environment
  a) Eclipse Juno - http://www.eclipse.org/juno/
  b) Eclipse Maven - http://maven.apache.org/eclipse-plugin.html
5) Java
  a) Murach Java - http://www.murach.com/books/javp/index.htm
  b) Java Run Time (JRE/JDK) http://www.oracle.com/technetwork/java/javase/downloads/index.html
6) Linux
  a) RedHat Linux - http://www.redhat.com/products/enterprise-linux/
7) User Interface
  a) User Interface Ozone Widgets - https://github.com/ozoneplatform
8) Machine Learning
  a) Machine Learning Big Data using Map Reduce http://machinelearningbigdata.pbworks.com/w/file/fetch/50030744/Machine%20Learning%20on%20Big%20Data%20-%20ClassIntro.pdf
9) MapReduce
  a) MapReduce Theory/Implementation https://courses.cs.washington.edu/courses/cse490h/08au/lectures/mapred.pdf
10) Hadoop File System
  http://hadoop.apache.org/docs/r2.2.0/api/org/apache/hadoop/fs/FileSystem.html
11) Statistics
  a) Basic Statistics for Behavioral Sciences - MIT OpenCourseWare *Basic Statistics for the Behavioral Sciences* http://ocw.mit.edu/courses/brain-and-cognitive-sciences/9-07-statistical-methods-in-brain-and-cognitive-science-spring-2004/lecture-notes/
  b) *Data Reduction and Error Analysis for Physical Sciences* http://www.amazon.com/Reduction-Error-Analysis-Physical-Sciences/dp/0072472278#reader_0072472278