NASA: Trust-Based Data Service Reuse and Recommendation

Engineering Notebook

Team NASA:

- Neeraj Saini (Project Manager)
- Pujita Rao
- Shuai Wang
- Venkatesh Sriram

Faculty Advisor:

- Jia Zhang
- Martin Griss

Iteration 0: August 26th to September 9th, 2013

09/06/2013 - Minutes of Meeting with Dr. Ramakrishna Nemani

Attendees: Dr. Raman Nemani, Jia Jhang, Venkatesh, Neeraj, Shuai, Pujita, Andy

Notes:

- NASA is bringing all the data together at NEX to make a central database
- Tons of resources at NASA (7th fastest Supercomputer in the world - Pleiades)
- Knowledge based social network desired
- 4-5 TB data every day from over 40 satellites

Key words: Big data, search, recommendation engine

Goal: Simulate an expert; recommend papers, tools and workflow

Expectations:

- Capture knowledge
  - Observe
  - Access their workflows
  - Citation index
- **Recommendation Logic**
  - **Depending on geo-location**
  - **Kind of sensor you are using**
  - **Constrained by computing power and hours**
  - **Disk storage constraints**
  - **Recommend Available tools - tools owner is willing to share**
  - **Based on previous step choices**
  - **Similar to Google scholar (Step up)**
- Results
  - Recommend path in the workflows
  - Recommend algorithms
- Feedback
  - **Thumbs up /thumbs down**
  - Misleading/inconvenient/success
- Challenges
  - Confidentiality - -publications
  - Getting scientist to contribute to training set -- Let us collect data
- Similar work
  - VisTrails - recommendation system currently in use
  - Taverna
  - Kebler

Important questions for the team:

- What features we want to capture?
- What criteria will be used for rudimentary engine?
- Understand last year team's progress so far
- How to let people access domain tools, models, analysis tools hosted at NASA?
- Start with citation, collect thumbs up/thumbs down?

Iteration 1: September 10th to September 30th, 2013

09/16/2013 - NASA internal meeting notes

Attendees: Jia, Neeraj, Pujita, Chris and Shuai (Remote)

Notes:

- Chris will be the python expert for the team and will help with integration
- Setup a demo of VisTrails with Chris if required
- Team to setup a web service which will act as a plugin with vistrail
- We have to constantly collect feedback from Dr. Nemani (Rama), Keep buffer time
- Get last year work up and running and work on connecting to vistrails with Chris help
- Use mysql for DB

Near Future work:

- Once VisTrails lists the results and then iterate to make a call to the web service for ranking
- Chris will initiate a request from VisTrails to our web service and get the results within the tool

Action Items:

- Watch the vistrails videos and tutorials and get acquainted with the tool - whole team
- Implement web service in java, no python
- Contact Shrikant for Last year code -Neeraj
- Prepare statement of work - Pujita
- Setup the environment - Venkatesh, Neeraj & Shuai
- Contact Andy/ Petr for team meeting slot - Neeraj
- Get VisTrails demo video from Petr – Jia

References: (Vistrails links)

- http://vistrails.org/index.php/Main_Page
- http://www.vistrails.org/index.php/Video_Tutorial
- http://www.taverna.org.uk

Progress: Communicated with Shrikant

09/23/2013 – Meeting with Petr

Attendees: Petr, Jia, Chris, Neeraj, Venkatesh, Shuai, Pujita

Notes:

- We will be using a **Relational database – mysql**
- Our solution will be able to carry out visual querying - recommendation based on matching
- One of our goals is to refine the existing system
- For similar work we can see what they are doing at labs
- Get a test case - analyze input and output

- We want to build a proactive recommendation system that is based on **human trust**. We can consider modeling the scenario/conditions scientists have, in order to develop a usable solution
- Using SAP HANA

Action Items:

- Touch base with Claudia to align with their work
- View VisTrails presentation sent by Petr
- Set up pivotal tracker and ASANA

Deliverable: NASA – Practicum Planning Roadmap v1.pdf shared with Petr
References: http://ecocast.adobeconnect.com/p3poaoksscu/ - VisTrails video shared by Petr
Progress: Environment set up for the existing code. Shared Git repository

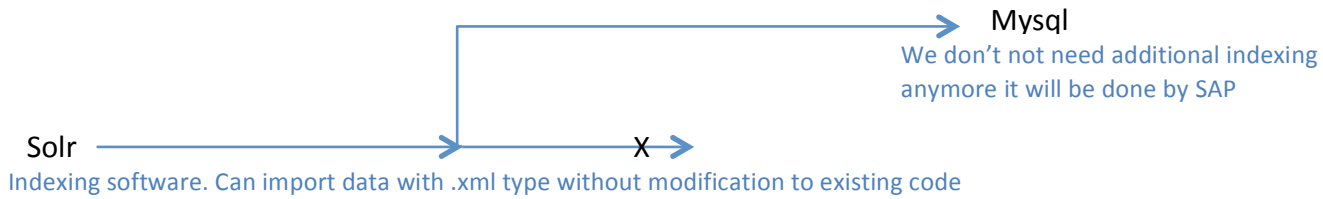


Figure 1: Decision

09/30/2013 – Meeting with Petr

Attendees: Petr, Jia, Chris, Neeraj, Venkatesh, Shuai, Pujita
Notes:

- Best approach to set up a pilot?

  We can get users by involving developers and users of VisTrails. Probably Petr will provide 2-3 from his team and Claudia.
- We **don't need solr because every column will be indexed in SAP**. We don't want more latency for doing redundant work.

Feedback: For road map - Good
Problem: Government is shutting down NASA - Petr might not be available for next two meetings
Progress: ASANA is set up. Have Access to code
Action Items:

- Set up a meeting for walk through of code
- Work on feedback feature – thumbs up & thumbs down
- Access to Database - Set up our own mysql database instead of setting up solr
- Ask Wendy/Ed for license to access Pivotal tracker
- Complete Statement of work and submit to Jia

Figure 2: ER Diagram

- Publication has primary key - publication type id
- DBLP

Iteration2: October 1st to October 14th, 2013

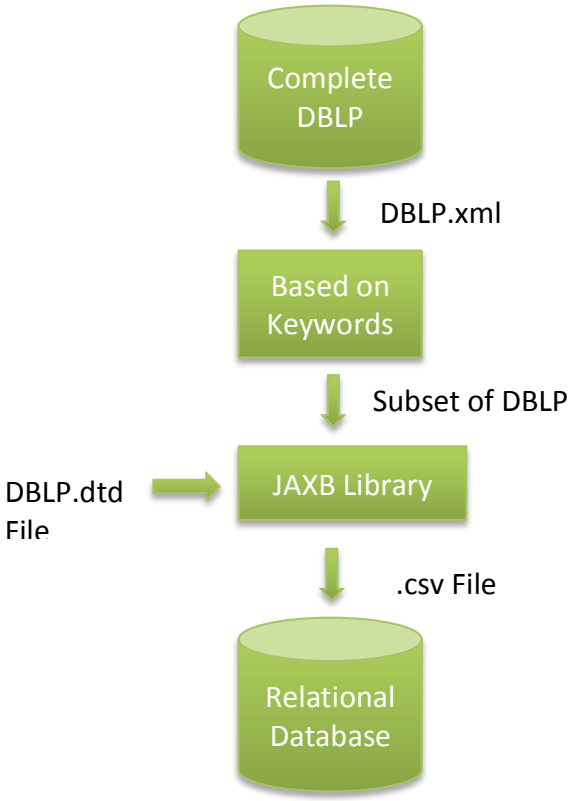


Figure 3: Data Extraction

10/07/2013 – Meeting with Jia

Attendees: Jia, Chris, Neeraj, Venkatesh, Shuai, Pujita

Progress:

- Working code that runs a test case and creates a user node
- Set up Pivotal Tracker, Created User Stories
- Statement of Work Submitted

Problem:

- We do not have access to the **twitter database**, but have relevant code for it.
- For the **DBLP** data that we have, the code was commented out and probably does not work.

Notes:

- DBLP has data for all publications in Computer science field that we need. It is in the DBLP Table
- dblp.dtd is the schema file that specifies the structure of everything
- Find **a java based xml parser** (similar to the one implemented) that takes xml file feed as input along with a dtd feed. It gives the output as classes.
- Mysql to store information
- DBLP data-type gives details of every field
- **Author to author have a link is if they co-author. Link has more weight if they have heavy co-authorship in recent years. Trust of software/models is based on human trust.**
- **Later we can add an additional feature based on what kind of co-authorship.**
- **User-User network is built. A user is a node.**
- **Jung** is the graphic part that had also been used in previous work. It is **used for the implementation of the Visualization part**.

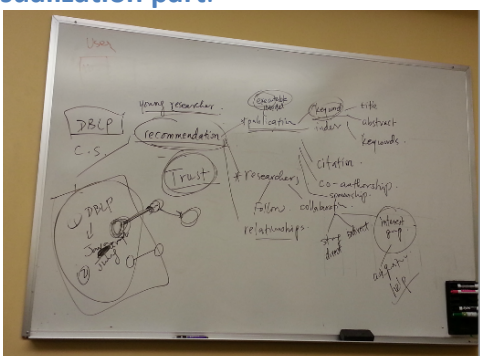


Figure 4: New Approach

New Approach:

- New student/researcher can use in place of Google scholar, for a better recommendation engine
- Recommendation for
  1) **publications** (all code has to be running -- executable model)
     a) keywords (apache has a project for keyword indexing)
        i) title
        ii) abstract

iii) keywords

b) Citation

c) **co-authorship** (first author is more valued than 15th author, but not for mathematical papers)

d) Sponsorship

2) Researchers

a) follow them

b) collaboration

i) strong/direct

ii) indirect

iii) interest group

(1) ask questions

(2) ask help

3) Relationships

- **In addition we complement our search by linkage. When sometimes user has to read another paper to understand a paper. That has been the pattern.**
- **Publication in journal is more reputable than publishing in a less popular place.**
- **Use case: not only for NASA - it can work for others.**
- Can be presented in cloud computing conference.

Action Items:

- Conference call with Srikanth
- Check point on Thursday.
- Find open source code for **User-user network visualization and computation**
- Start clean project and complete DBLP part of coding– Shuai & Venkatesh
- Work on Visualization using Jung – Neeraj & Pujita
- Ask Chris for help if needed for code related to Jung

Suggestions:

- Database Schema should be on git hub

We were to implement the existing working code into VisTrails

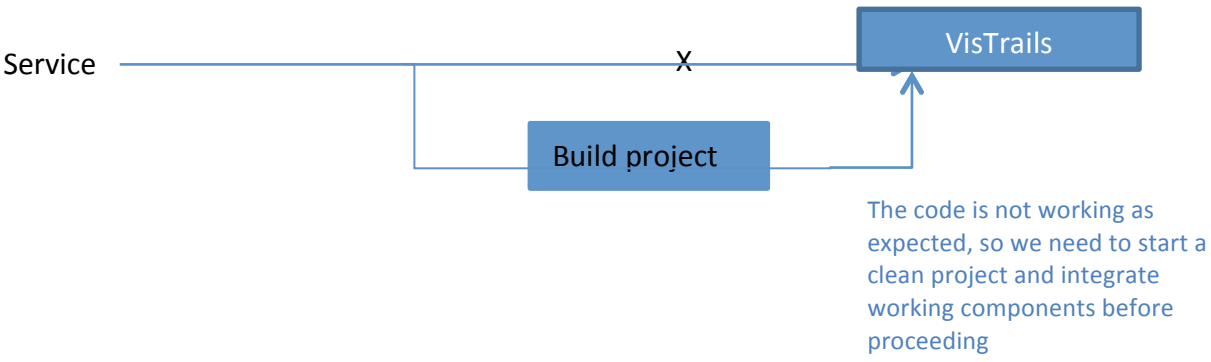


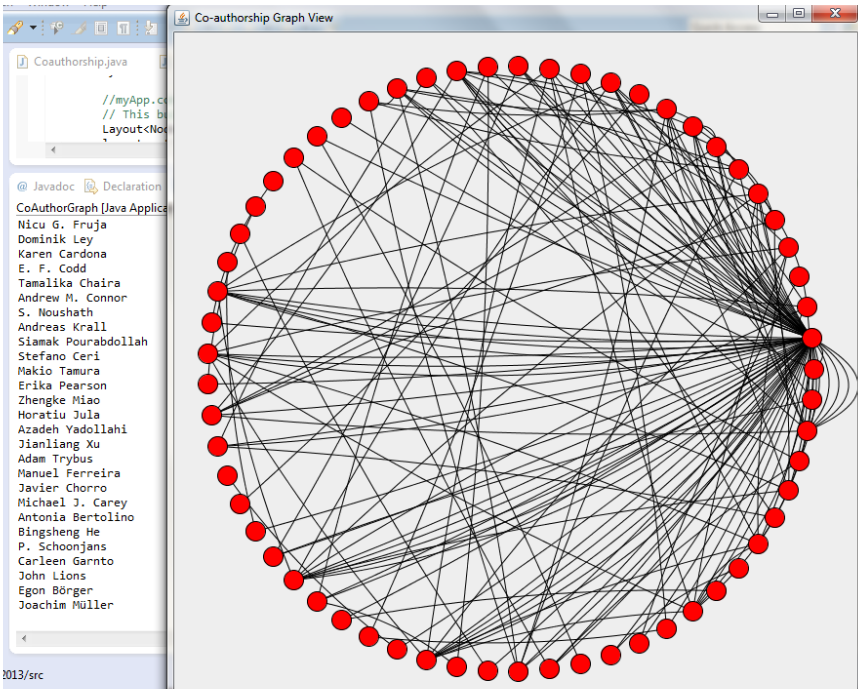
Figure 5: Decision (Strategy to recover from time lost in resuming last year's work)

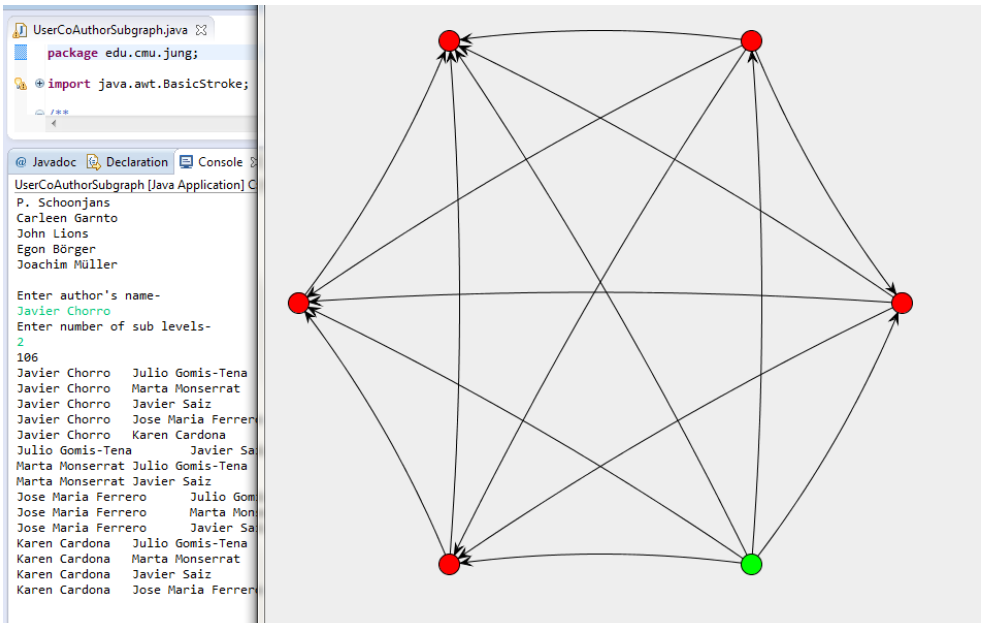
10/14/2013 Adobe Connect Meeting

Attendees: Petr (Client), Jia (China), Claud Wang (External), Chris (Mentor), Shuai (Pittsburgh), Venkatesh, Neeraj & Pujita (CMU-SV)
Notes:

- We can easily change the output from names to DBLP ID
- Last year there was integration with twitter, making use of the communication among Authors. **If we want to use data relevant to scientific researchers we will need real time data. Twitter no longer has open APIs**. We need to figure out more social networks that are being used by the scientific community.
- Chris will use it in **vistrails**. He has the history of workflows and wants to know the author of specific software with trust; combining publications with services.
- Generating synthetic data for scientific workflows
    - We would like to get atleast 2 workflows to get an idea to make a synthetic test set.
    - Chris has found programmer web. It does not give **direct links between workflows**
    - Old **data from Taverna can be used for testing purposes**. We can use that directed graph.
    - Petr can generate 6-12 workflows.
    - Strategy - some of the workflows are not being used but we can do some simple mapping and generate a network for evaluating and testing our tool. Building can be done using VisTrails, since we are not dealing with the execution yet.
- This solution will be helpful for Claud Wang too.

Challenges: Chris uses python for VisTrails. If we use Json it will be easy.
Progress: Working sublevel graph (Demo). With Visualization implemented.





Action item:

- Petr will provide 1-2 examples by next week (An abstraction of the real process - with some complicated and simple samples)
- By next week, we should integrate with VisTrails. Give a demo.
- For this, we will need to build the interface between us and Chris
- plan to come out with API
- Work on setting up **a web service**, data model and data service

Iteration3: October 15th to November 4th, 2013

10/21/2013 Meeting
Attendees: Jia, Venkatesh, Shuai
Progress: Implemented **REST API**
Demo: (with Petr, Jia, Chris, Venkatesh, Shuai)
Action Items: Work with Chris to complete integration with VisTrails

10/28/2013 Meeting
Attendees: Petr, Jia, Chris, Neeraj, Venkatesh, Shuai, Pujita

Progress: Finished our side of work for integration with VisTrails

Notes:

- Abstract, keyword, authors,
- **Harvard** has the entire **database** for earth science

Next steps:

- Convert to sql. We will be using xml parsing
- Work with SAP HANA team. Build up API to use internal HANA

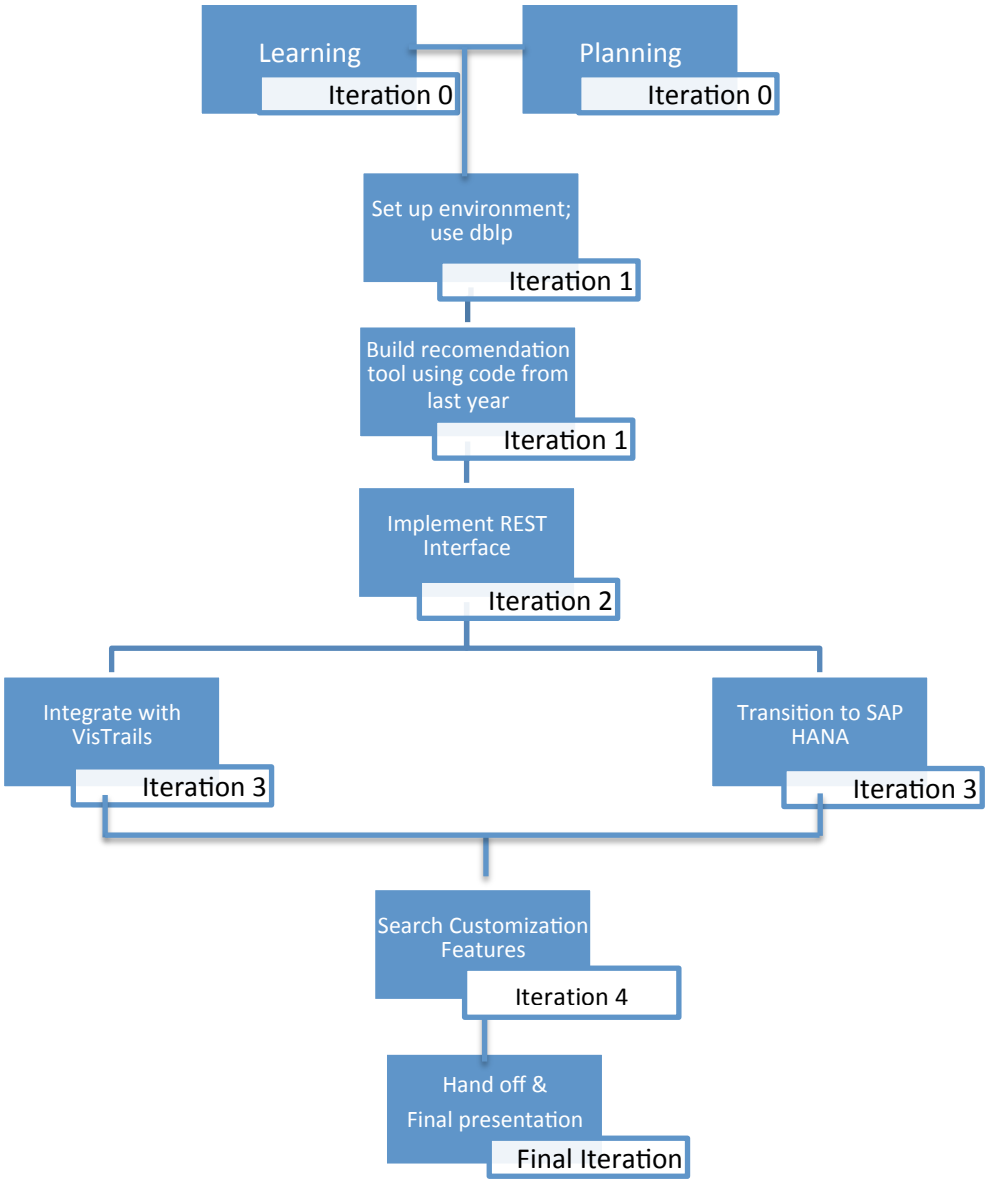


Figure 6: Work Breakdown Structure (WBS)

11/04/2013 Meeting

Attendees: Jia, Petr, Chris, Neeraj, Shuai, Venkatesh & Pujita

Progress:

1. Integration with VisTrails
2. **Completed transition to SAP HANA**

Demo: Integration with VisTrails (By Chris)

- Nodes are randomly placed
- Can the team move the HTML?

Notes:

- We have equals for the beginning. We could use user feedback. Thumbs up thumbs down
- **Every paper can be treated as a workflow, If you cite, it means you are using someone else's workflow**
- Related to Chris' Work:
    - We can generate a test bed using the order of citations. Reference ordering can be based on
        - **order of appearance in paper (not in the reference section)**
        - **Use programmer web,** Jia's work. Paper needs a workflow associated with it. Reviews need to be able to repeat/ evaluate the existing work.
- We can add this layer**. Human trust will move to software trust** this way.
- Apply some machine learning to SAP HANA(Jia to share the paper)

Team internal meeting:

- Chris is performing the search using historical modules together
- The data flows from one module to another
- Analyzing the historical use of modules in NEX
- add human trust as another layer on this usage
- Use the human trust from publications(DBLP)
- More weightage to
    - **reputable and common result from list of recommendations**
    - dpbl has diff publication channels like whitepaper, journal. **More weightage to source that has been popular for many years.**
    - **Relevance to field.**
    - **Similarity to topic**
- **Support/help from collaborators and coauthors, is more likely. So we have more trust on them than competitor**
- publications -> Author (we should provide a graph of co-authorship as API)
- edges(paper) - should be labeled
- User can put more parameters like "Show only co-authorship from last 5 yrs"
- **Small world theory** - social tie is no more than 6 hops for a connection
- **Model: Paper is a workflow, coauthor is an item. Form association rules**

- DBLP can be leveraged using the **DBLP citation Version**
- Scientific collaboration is a social network

Challenges:

1. What should be our next steps? Do we need to do an analysis the twitter database?
    a. Not in these time constraints. We have a much larger base of conclusions we can draw from dblp rather than from Twitter
2. Now that we have it in the SAP HANA database, what machine learning algorithm should we use?
3. Can we import a database with foreign keys using CSV? we have been using the name instead of directing it to another table

. Directing you to a GA students, since Jia's team has been using it a lot

4. Progress on the sample workflows Petr was going to share with us?

Next steps:

- Neeraj & Pujita:
    - How to use the SAP HANA to help with prediction and machine learning
    - Document how to get into the HANA database, library, to repeat for later teams
    - Learn about in memory database usage
    - How to use PAL, the library
    - Bring slides - progress, challenges, next steps
    - Schedule sponsor meeting on 14th and 25th November
- Venkatesh & Shuai:
    - Given the author, find out all coauthors, list
    - I give the starting point of one person, timeframe and topic. Find coauthors.
    - Get algorithm to calculate reputation from Srikanth.
    - Get the graph to find social network based on authorship (work with Chris to make sure it is working at his end)
    - Assign labels(can be paper) on edges

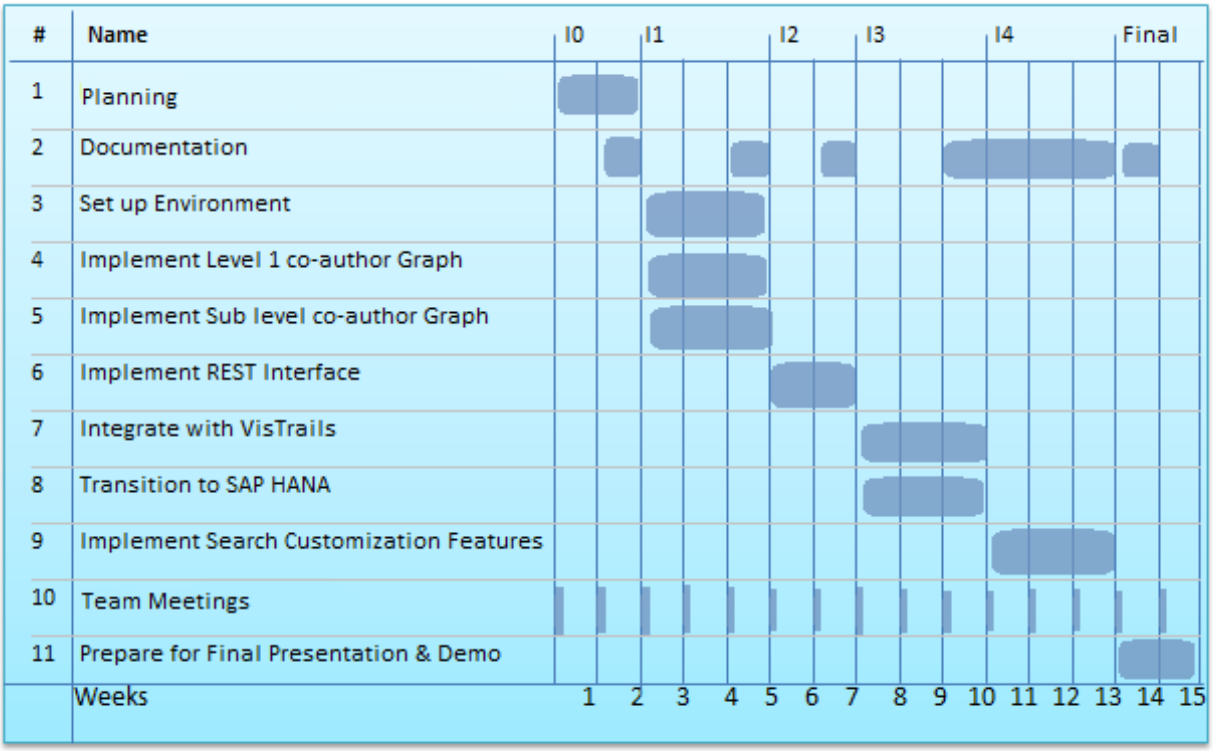


Figure 7: Milestone Chart

Iteration4: November 5th to November 25th, 2013

11-11-2013 Internal Meeting

Progress:

- Now the recommendation is based on Field, criteria, reputation…
- We are using **Naïve Bayes network algorithm** because it is generic to many cases

Notes:

- Architecture course – how to set up **a governance layer for management of ever going project,** documenting, testing layer, management

Challenges:

- There are cases like "C.J. Date" who has more reputation in Automation and less in History
  - We might need **ontology** for such cases
- We have to find out the specific domain for deciding the domain for Reputation.
  - Top journals have keywords. Leverage the title, abstract when keywords are not there.

Action Items:

- Schedule time on Wednesday 5-6 pm

11-13-2013 Internal Meeting

Challenges:

- Naive is not available to us because of **licensing issues** – we are paying for **AWS**, not for HANA
  - Use HANA if we need a categorization package
  - For now we can use anything else
  - Use regression, tree
- DBLP has entries where abstract/keywords are missing, APIS are not available for many, but all have titles
  - Find **similarity between topic and available data {title, abstract, ..}**

Notes:

- Expected future state
  - Build a complete network on all data from DBLP, ahead of time, not dynamically
  - Start with small subset
  - {a1, topic1} should give a subnet/sub graph
  - We are simplifying and not worrying about the direction. Considering everything is non-directional.
  - May **have indirect collaboration. Trust propagation** - if there is a strong link based on trust value between a1 and a2 as well as between a2 and a3, and then you can consider a strong link between a1 and a3.
- How this will be used by NEX: **Combining software trust with author trust**
  - This can be used with C1 ---C2 software components
  - Chris' tool will recommend a software components and give you names of top ranked authors
  - We need to rank them with our algorithms
  - And provide an output as a trust level that Chris can use

Action items:

- Make the a1--- a2 graph as a web service

11-14-2013 Sponsor Meeting

Progress:

- Recommendation tool considers
  - topic
  - time frame
- Gives the reputation and co-author
- Edge between authors (Nodes) has publication

Notes:

- Establish Topic
  - Use **WordWeb – taxonomy** to find similarity between topic and data from DBLP
  - For **similarity of abstract we can use Symantec indexing**
- Because we have **licensing issues with HANA, we can use HANA to store sql, not for machine learning library**
- Database
  - Hana is column based db so depends on how we design db. We have 2 million records
  - Does **DBLP cover the earth science database**?
    - Look for Rama
  - **Google scholar has data but we cannot do queries**
  - Impact factor won't be problem for earth science since we use a combination of all sources
- New Direction: **Calculating willingness to help**
  - People who contribute on stack overflow, etc. are more willing to help. Even if they have not authored
  - Will be useful to software support
  - Study user behavior
  - Business market behavior analysis at NEX to see if recommendations

11-18-2013 Internal Meeting to discuss Machine Learning algorithms

Progress-

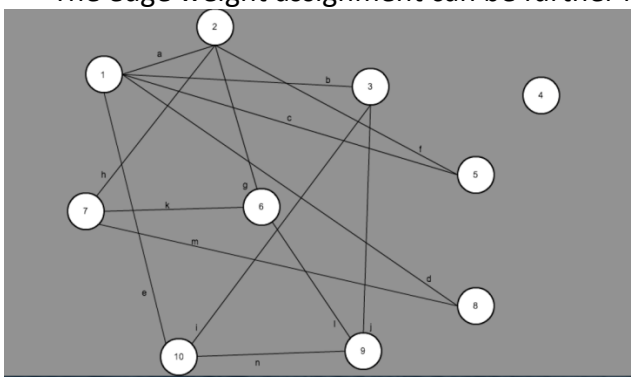
- Formulation and implementation of algorithms for likelihood of helping.
- Algorithms used-
  - **Stage 1: Jaccard Co-efficient**
  - **Stage 2: Naive Bayes**

Notes:

- Jaccard's coefficient measures how likely a neighbor of x is to be a neighbor of y and vice versa

- Naïve Bayes calculates the probability an event can happen if a certain set of independent events occur. It can create a probability model based on a conditional model.
- Modeling our problem
    - Link probability of every single nodes pair (author pair) ------Jaccard
    - Likelihood of help given from another author ------ Naive Bayes
- Give direct analysis based on given graph and calculate the probability, which, according to the model, is essentially the help that can be expected….
- Similarity consideration:
    - The connected nodes (One or Two level, all implemented)
    - The trust level
    - **The journal name**
    - Research domain (ACM Taxonomy)
    - Reputation level
    - **Location**
- **Normalization:** Depends on total number of edges and total Jaccard value.
- Currently information about 10 authors.
- No prior precedence about interpretation and scale of output value. Needs to be interpreted by us.
- We allow two kinds of input currently-
    - One level- Direct co-authorship
    - Two level- Co-authors of co-authors

- Currently assign **Edge weight** = 0.5, to second level connection, and Edge weight = 1 for first level connection, to model the fact that second level connection is less solid than first level connection.
- The edge weight assignment can be further refined.



- Idea from classic machine learning algorithm, codes all created by us. Data all real, apart from the location information (not intact in DBLP).

Feedback:

- Provide clear justification of the choice of machine learning algorithms.
- Remove redundancies from model. The mutually independent events do not look independent.

- **Map problem to generic case in algorithms**

Action Items:

- Refine model, especially how to use trust value and assign weight
- Dig into the analysis of results
- Make sense of our calculation functions (parameters..)

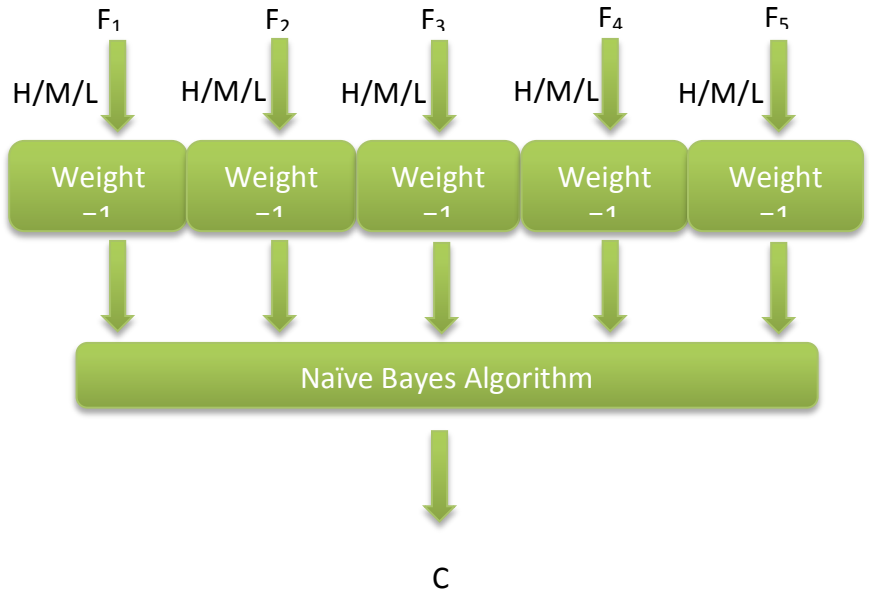


Figure 8: Probability of Co-authorship

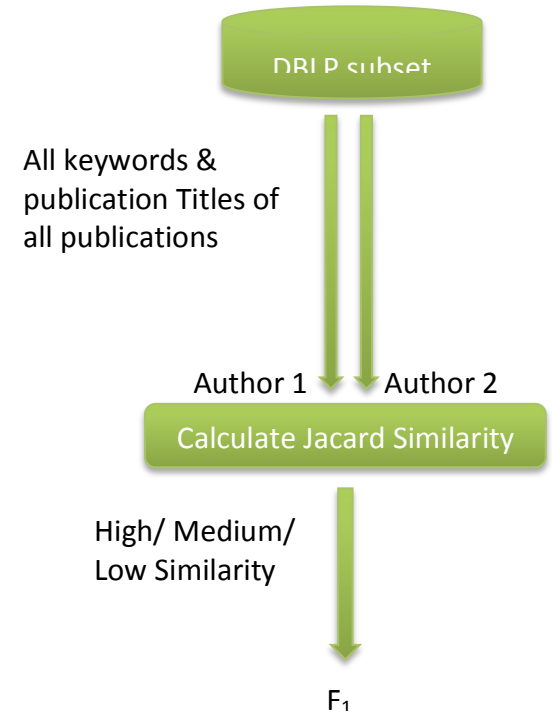


Figure 9: Similarity of research areas

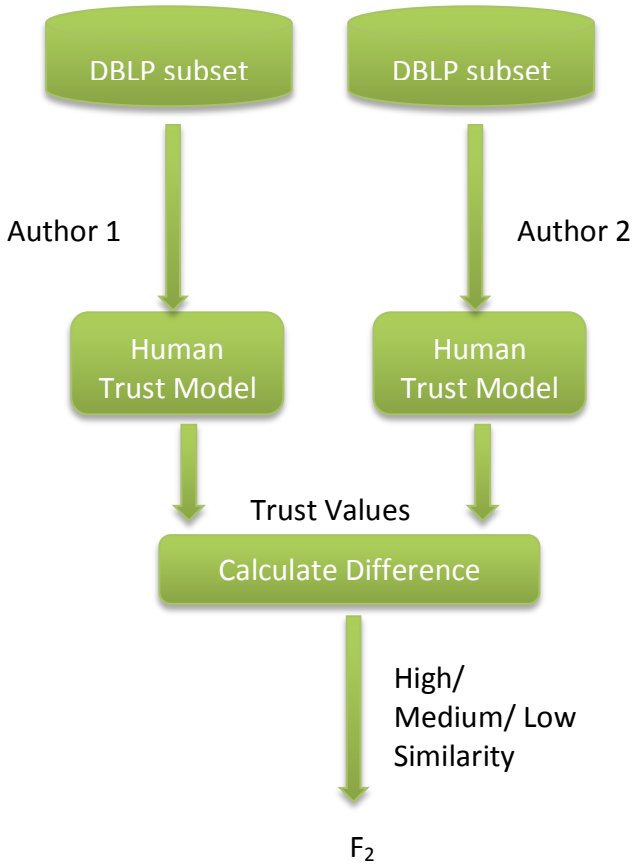
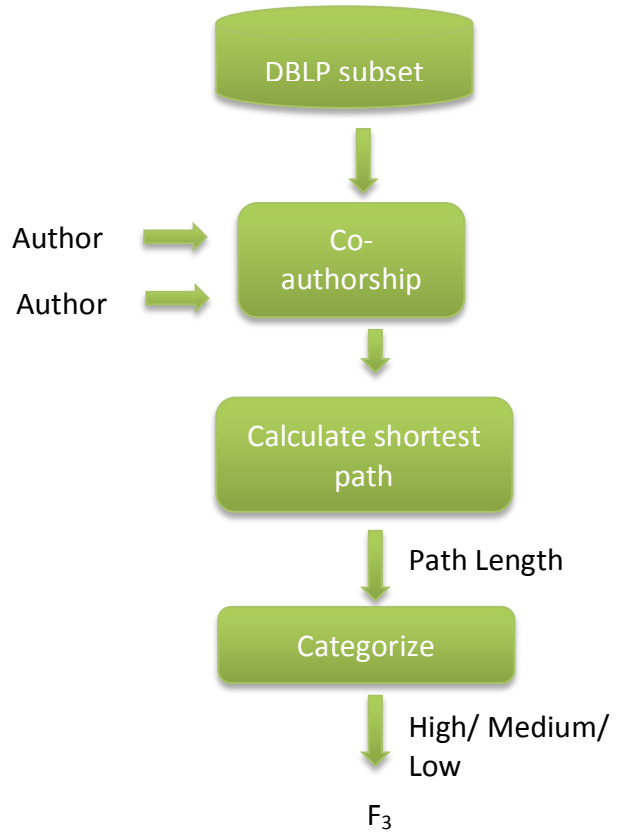


Figure 10: Similarity of author reputation



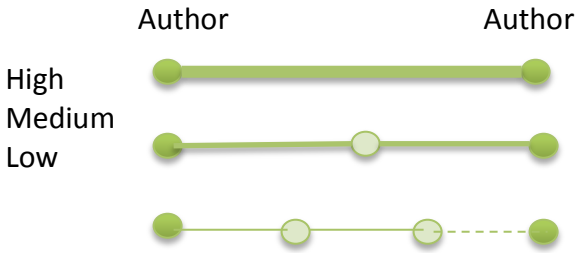


Figure 11: Author connectedness

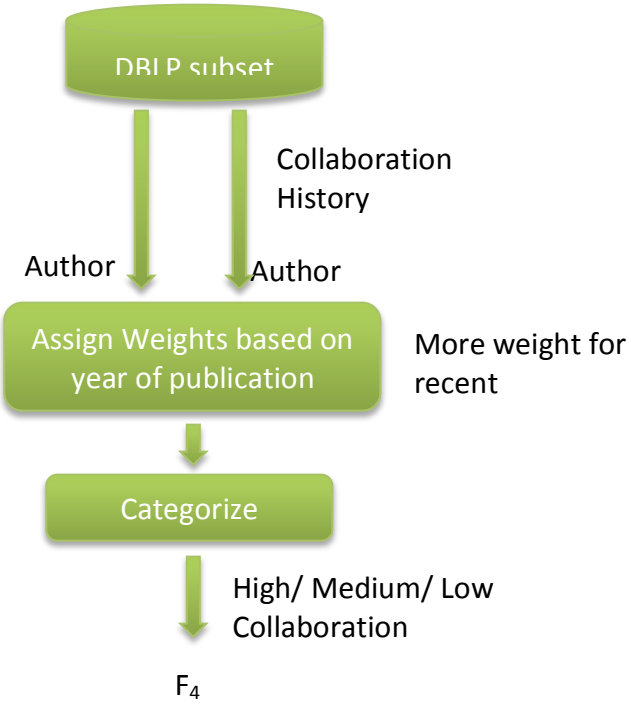


Figure 12: Collaboration history of each author

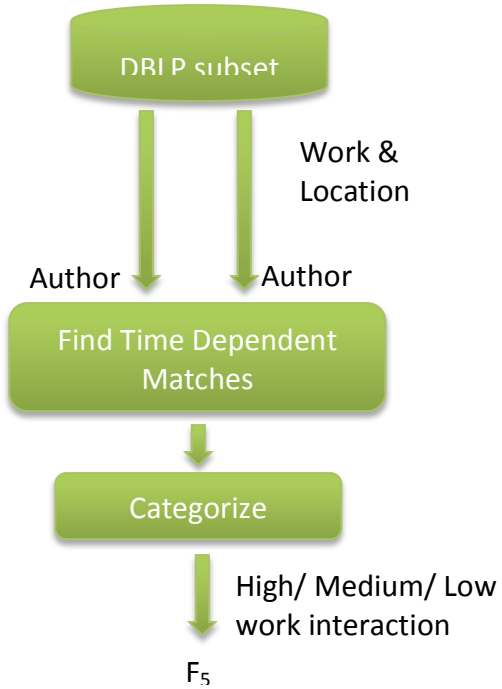


Figure 13: Work interaction strength

11-25-2013 Sponsor meeting

Attendees: Petr, Chris, Venkatesh, Neeraj, Shuai, Pujita

Progress:

Notes:

1) Naïve
    - input = 2 authors, output = probability of co-authorship based on:
        - **Similarity of research**
            - (Jacard for similarity of keywords – titles, ACM taxonomy)
            - High, medium, low similarity
        - **Similarity of reputation** (trust value calculation. Considered absolute difference)
        - **author connectedness** (Low = never coauthored, medium= coauthored once, high = coauthored more than once)
        - **collaboration history of each author**
            - (should be a combined collaboration history, if they have coauthored in general a lot)
        - **work interactions strength** (mainly the location and university)
            - **Maybe use linkedin**. Since very few have an institute.

- Work interaction strength: if we go back in time and compare, we can prove that the connected world has more co-authorships

- Assumption about different reputation might not be true. Because they will be **student mentor publications**

2) Training data:
   - Jacard similarity btw authors
   - Reputation similarity
   - Connectedness
   - Output: if they collaborated

- File format is Training, test, result, features. Example H, M, L, L
- 85% is correct for test set. It has 27 inputs and considers 3 features.
- The Training set is 2000.
- In machine learning, we do training and testing by doing **Precision and recall**. We have just done precision. Use same data for calculation of recall from training data
- Infrastructure is in place. Project is in good status.

Action Items:

- Create a model for front end user interface.
- Clean up the code and comments
- Assign more precise weights- use basian
- Do a 90% 10%. 10 fold, nine fold.
- Schedule Client presentation on Thursday-2:00pm

Final Iteration: November 26th to December 5th, 2013

2-12-2013 Sponsor meeting

Attendees: Petr, Jia, Venkatesh, Neeraj, Shuai & Pujita

Progress:

- Improved data set
- increased test cases
- 4th feature (collaboration history) implemented
- transitioned to java completely- we are not using python anywhere except for machine learning algorithms
- interface created (integration remaining)
- Calculating recall

Notes:

- Had 60 authors, now we are using all the data from dblp related to distributed systems
  - 5000 training data
  - 700 for test
- Usually getting the data for machine learning part is the inherent work. Most of contribution comes from treating the data

Demo: AlgorithmTest file genetrates >Re.txt file
Recall has improved from 85% to 99.86%

Coauthorship.java & twoAuthor.txt gives Result.Text

Suggestion: Can you say which **feature contributed more**? It tells which features are contributing more and which insignificant features can be safely removed

Action Items:

- Write true positive means this. False positive means this. And these are the numbers we already have. Sync up with lingo of domain
- Put information about L,L,M,H in output – L=Low, H= High and M= Medium
- Jia to show us example for Precision and recall
- Goal is transferability than completeness and polished output

Challenges: Incompleteness of existing data

Future Work: Dont have data from linkedIn
Leverage the time dependence

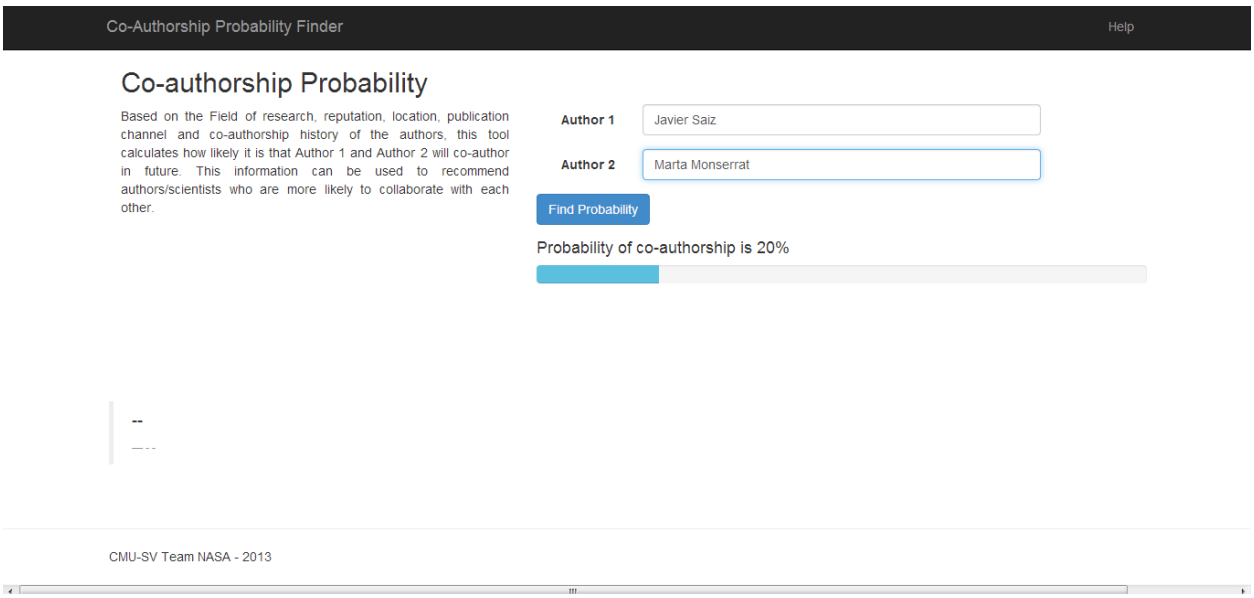


Figure 14: Web UI

Internal Meeting

Progress:

- Precision Recall example given
- First draft of Final presentation

Notes:

- Technical report
- Connectedness and collaboration is different.

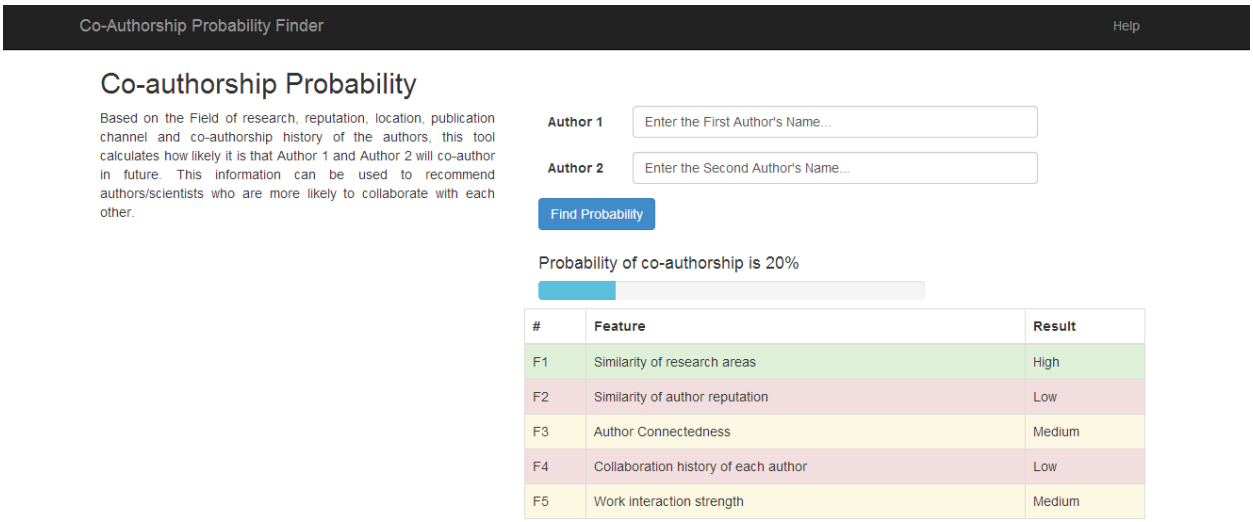


Figure 15: Web UI