Fall 2014

Technical Report

NASA Practicum-Trust Based Recommendation System

Enclosed in this document is the technical report of the Trust Based Recommendation System sponsored by NASA

Point of Contact: Petr Votava
Faculty Advisor: Jia Zhang
Team: Krutika Kamilla, Abhishek Mukhopadhyay,
                      MustafaTasdemir, Jisha Vadake Muthiyil

Table of Contents

1 Introduction

The 'Trust-Based Recommendation System' is a web application that facilitates scientific collaboration and is based on the ongoing research on collaborative systems done in conjunction with NASA. The primary goal of this project is to create an end-to-end solution to NASA Researchers about recommending contributors to research studies that best relate to people's academic references and publications. The end-to-end solution will contain a web-based interface for the users that will create dynamic graphics, which reveal the relevance and trustworthiness of researchers on each other. This visualization will generate reliable inferences between people to suggest the best potential contributors into a projected study. The relation between different authors and the authors' list of publications can be visualized. Meanwhile, the existing trust calculator is in place, it will be enhanced to include some relevant parameters that relate to researchers' online social metadata to increase the reliability of the solution. The trust score is calculated based on the number of papers published by the author and coauthor ship. It is the novelty of our web application as on the basis of that, new researchers in the any field can get recommendations on whose papers to read and with whom they can collaborate with for researching in a particular field.

2 Motivation

Studies show that, per year $1.6 trillion is being spent on Research and development worldwide out of which U.S. alone spends $465 billion. There has been a continuous contraction in budget through the past 50 years. This necessitates efficient utilization of these funds so that it will benefit humanity to the maximum possible extent.

NASA, which is the one of the most important government agencies that fund research and development in the US face the same constraints. This is coupled with scientists working on disparate domains but overlapping problem sets. A lot of redundant research working is happening which can only be avoided if there is a proper platform which helps people find what are the publications in a particular field and who are the experts in that field. There is a need to accelerate research through collaboration. Accelerating collaboration by making it easy to identify collaborators working on the same problems is important. NASA's Earth Exchange (NEX) platform aims to provide a cloud-based platform as a service to accelerate big data analytics and scientific collaboration in Earth science. There is a possibility of extending to or integrating with subject domains too in future.

Collaboration between NASA and CMU-SV started 2 years back in order to come up with a trust based recommendation engine that would facilitate scientific collaboration. There was a need to develop a system, which recommends experts and collaborators based on domain and trust. The previous teams who worked on the trust-based system developed an algorithm to calculate trust based on knowledge reputation and other social factors. Our focus was to develop a prototype, which productized the large amount of research work that had been done in this field. Our aim was to develop a product that would help find existing research in a particular field, find the experts in that field, figure out of these experts who can act as collaborators and perform all of these tasks leveraging the user's trust. From the very beginning of development, our prime focus was accessibility and usability of the product

3   Related work

There are many products available that enable scientific collaboration. This section discusses the main features of a few products currently available.

- **ResearcherId**

ResearcherId is one of the most popular tool used for scientific collaboration. The best feature of ResearcherId is its integration with the Open Researcher and Contributor Id(ORCID). This enables data exchange between the ResearcherId system and ORCID. ResearcherId has got four main interfaces, which facilitates different types of search. There is a map view, which shows the locations conducting research in a particular topic. The biggest disadvantage of ResearcherId is that it is proprietary.

- **Researchgate**

Researchgate is a social network of researchers. It allows users to create profile, upload papers and discuss issues. Users get detailed statistics about their papers such as number of hits, citations etc. The search feature in Researchgate is not very intuitive.

- **PubMed**

PubMed is one of the richest databases of scientific papers. It was first released on 1996 and since then has been a popular search engine. PubMed for handheld devices is one of its best features. The main disadvantage with PubMed is its interface and lack of usability. To use the comprehensive search in PubMed, we need a thorough understanding of

its components. Moreover, they do not have any information that could be used for collaboration and they do not track citations

Apart from the ones discussed here, most of the university research databases are good sources of information. We had the opportunity to review a handful of them and adopt the best features into our product.

4   System design

As every other system requires, our practicum project also needs a system level design for a robust, reliable and maintainable use. This section discusses about the design decisions that are made throughout the whole semester that we have led after previous work done before us.

Quality Attributes

The key points in system design rely upon the attributes that the system must present at all times. We address these attributes according to the overall requirements of the client. Some set of requirements lead to a quality attribute that best identifies the required system attribute. We have listed the following quality attributes according to user requirements.

i)   **Accessibility**

Since the project has been ongoing for the last couple of years, it revealed that the outcome prototype or product was not reachable by the intended audience who are the users of the system. Therefore, the main concern for the client was to make the system accessible via modern browsers at all times. This way, there would not be any installation guideline that the users would struggle about.

ii)   **Usability**

When we achieve the accessibility, the next biggest concern will be how usable the system is. The users will have to perform the functions that the system provides in a browser, which in some cases may create some unusable features considering that the project deals with a substantial amount of data behind. Therefore, optimizing the usability of the system taking possible use cases into consideration will be another aspect of the system design quality attributes.

iii)   **Extensibility**

The system will offer a trust-based publication network out of which users will infer sensible information to learn about specific topics or authors in

computer science field. As one of the previous teams have focused, trust-based calculation is formed with knowledge and social based factors. Each of these may contain different kinds of data sources such as co-authorship for knowledge and Twitter/LinkedIn data for social based factors in order to utilize the trustworthiness in a more reliable way. Hence, being able to add more and different types of data sources holds in our system quality attributes.

System Architecture

i) **Dynamic View**

Since the system focuses on consuming big data and constructing meaningful representations out of it, dynamic flow of the data becomes an important aspect of the architecture. That's why the following schema has been created to show this flow as well as with the underlying physical devices that are required for deployment.
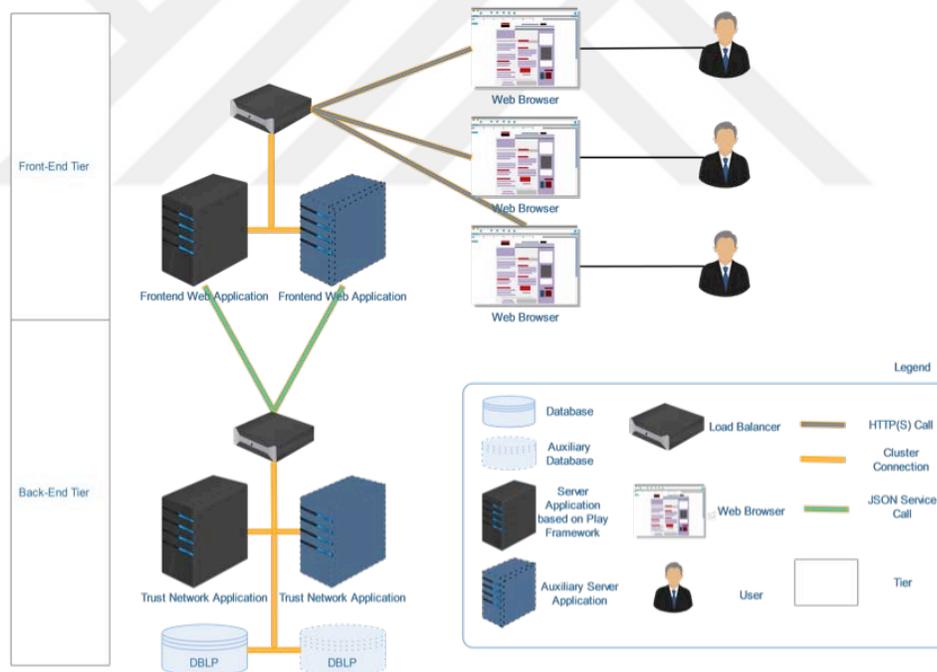

Figure 1: Dynamic View of System Architecture

As seen in Figure 1, the overall system relies on two-tiered system for better maintainability. The back-end tier consists of the processed data to be consumed and the trust modeling application based on that data. The front-end tier consists of the web application to be served to the users. As shown in the figure, the number of servers as per in the tiers can be extended for performance issues. This will require some load

balancer servers to actually distribute the request load to the servers in a reasonable and equal way.

As mentioned above, the two-tier architecture let us enhance modularity of the whole system. The data is served through JSON and the front-tier applications consume them via service calls. Therefore, in case another application need emerges for trust-based publication network, the underlying application and data tiers already exist which help us focus on the third party application design and implementation separately.

The framework choices include Play Framework for all types of application servers either back-end or front-end. The reason behind this preference is that the static structures of both tiers (in terms of packages and classes) help us construct the system faster and more efficiently. Also, the models that we have used for data representations are shared within the tiers. Hence, we thought keeping the same framework for inter-cross applications would be more accurate and appropriate.

ii) **Static View**

Static view refers to models and objects defined for the system. Since the publication network includes different kinds of publications, we have come up with a hierarchical modeling base in order to represent the data in an efficient and easily maintainable way.   The figure below shows the schema of the data representation.
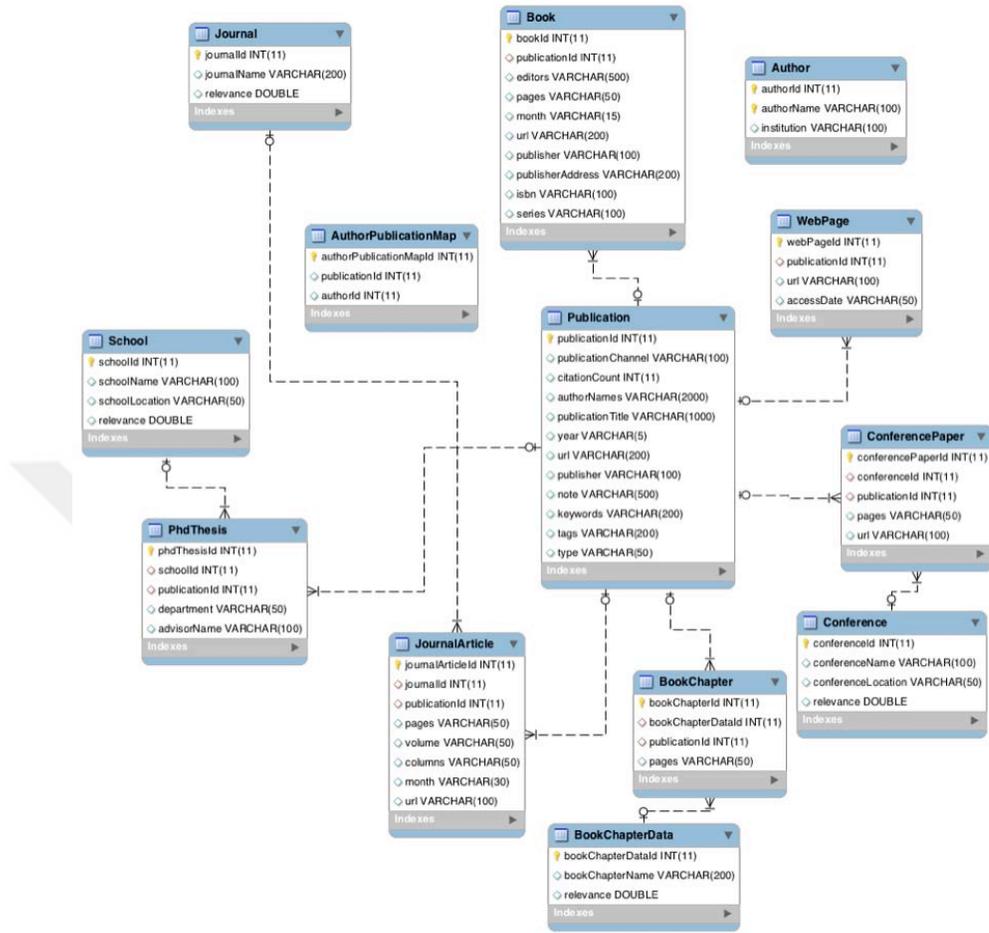
Figure 2: Database Models Schema

How Attributes are met

The previous work on this project had produced desktop-based offline applications that were not accessible enough for the client. Therefore, the solution of a web-based application will fulfill the accessibility requirement as suggested in previous sections. The web-based application will also require a usable interaction for a better utilization of the system and hence the required study for usability optimization has led us to web-based visualization tools that are lightweight and easy to use for possibly projected interactions. This helped us overcome the usability attribute. For extensibility, multi-tiered architecture as well as with normalized data modeling will make the system easily extensible in case of new data source must be integrated or any new business logic must be provided. Thus, the design will provide extensibility to the client for enhancement in the system.

5   System implementation

a)   ETL (Extract & Transform & Load)

Extract & Transform & Load (ETL) process is for preparing the data consumable by the applications developed for this project. To ensure that the data is loaded correctly and efficiently, we have defined the ETL process from scratch. The following sections reveal the details about what our approach is and how we have applied it.

i)   Data Source

Raw data source consists of XML files, which include the details about publications in computer science field and the citations associated with them. According to our estimates, there are around 4.2M publications and nearly twice as many authors within the dataset. Although the total size of the files do not even exceed 2GB, considering that they include just metadata about the publications, it becomes massive enough in data size. The format, that these raw files use are not efficient enough to use directly, especially when considering the data size. Therefore, we have defined an ETL process to prepare the data for the application.

ii)   Data Transformation

As mentioned above, the data format is not consumable directly. Therefore, we need to transform the data for our modeling. In order to do this, we first came up with the models. Since data source had a hierarchical data representation for publications (such as Book, Web page, Conference etc.), we have adopted the same approach. Because, we thought it would be more to the purpose for our extensibility in case we will expand the data set with different types of other data sources especially from social media.

The way that we have transmitted the data from files to our destination database management system consists of two parts. The first part is actually the transformation part in which we are changing the data structure of the raw data source into the same representational form as our data modeling. This way, we can directly load the data and apply performance optimization depending on our need. In detail, we have written a custom XML streaming parser to consume the data source. We have experimented several approaches such as loading all sources content into memory and streaming from the source. We found out that streaming is way more efficient and reliable than memory processing. On the other hand, the implementation became more sophisticated. After streaming the content, the parser exports each individual records that it detects to a CSV format file for corresponding models. So, the data becomes ready for loading phase.

However, one problem we have faced was the fact that the data source was not consistent and for some records not all the attributes were existent. Therefore, we needed to find out these edge cases and exclude them so as to refine the data set.

iii) **Data Load**

After all above processes, data is ready to bulk load into database management system and apply optimization on it for better performance. We have automated all the scripts and put them under a separate main folder in the project. The transformed data can be loaded by running these script files directly.

b) **Data Fetch**

MySQL database is being used in the backend. To fetch the data from the database, prepared statements have been used. Using prepared statements makes it easy to insert parameters into the SQL. Prepared statements pre-compile SQL queries and hence it is more efficient than calling SQL statements. It skips the plan generation step on subsequent calls thereby increasing the efficiency and reusability. Each of the SQL queries have been put into different methods to ensure maximum modularity.

c) **Trust calculation**

Trust score includes two components: a knowledge factor and a social factor. The knowledge factor depends on publications and knowledge reputation. The reputation is decided by the publication channel and citation power. Each component is given different weightage depending on the time of the publication. The weights also vary according to the type of the publication. Each of these weights was modeled using the dblp dataset.

The social factor of the trust takes into account coauthorship. Coauthorship is also given a time scaled weightage such that the papers co-authored in the recent past gets higher weightage compared to the ones published in distant past. The recent papers include the ones published in the past six years with respect to the date of query and the intermediate papers include the ones published in the last 6-12 years.
The final trust score is a combination of the knowledge factor and social factor.
Final trust score =
(CoauthorshipCount*TimeScaledCoauthorship)+KnowledgeFactor

d) **Data Provision layer Implementation**

Controller class function is called to render the graph directly to the view JavaScript. This also calls the transform encoding helper as the topic parameter uses JavaScript's encode Uri component which outputs a different format when compared to Java's URL Encoding/Decoding, which the frontend - backend applications use to transfer parameters. API call of util is used to make the API call to the backend and return the JSON data to the frontend to render it.
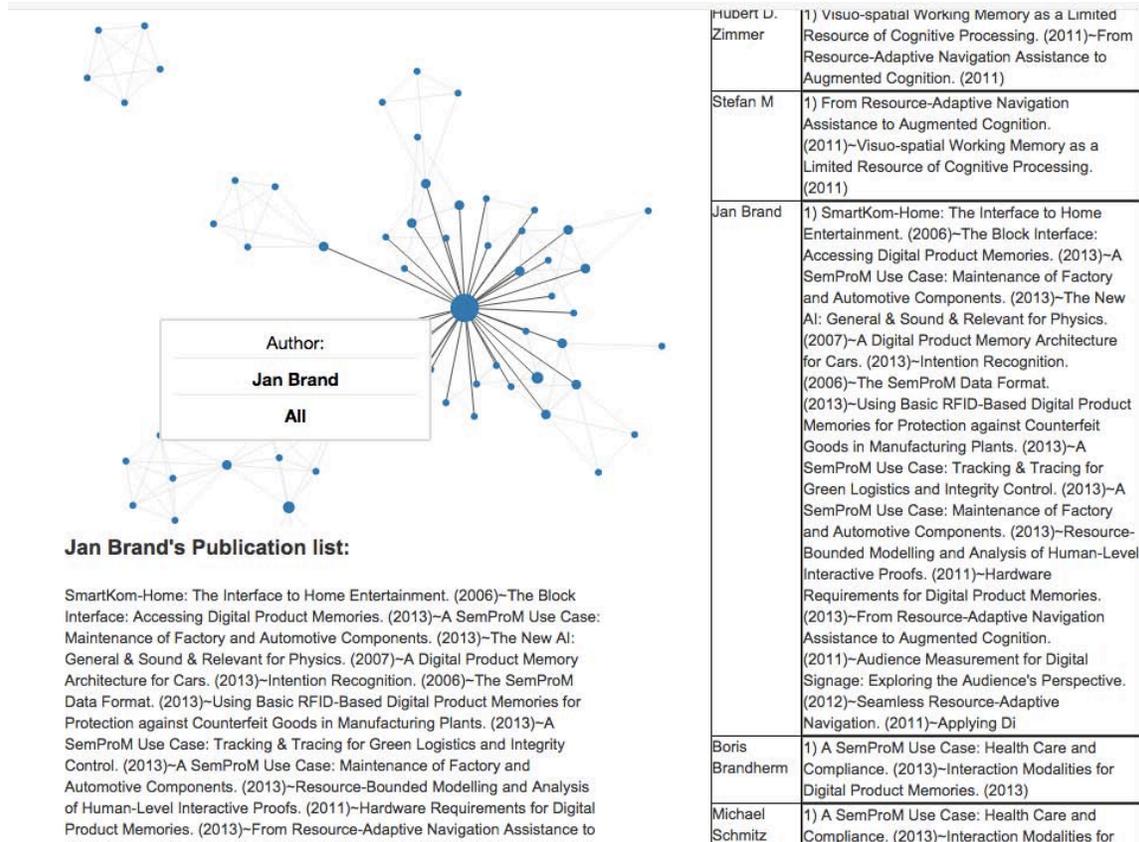
e) **UI Implementation**

Json data which comes from the backend through ajax call is fed as input to the vis.js and the nodes and links objects create the directed graph. Figure below shows a sample node and link json object.

```
{ ⊖
    "nodes":[ ⊖
        { ⊖
            "trustScore":"",
            "topic":"All",
            "name":"E. F. Codd",
            "list":"Interactive Support for Non-
Programmers: The Relational and Network Approaches. (1974)~Interactive Support for Non-
Programmers: The Relational and Network Approaches. (1974)",
            "id":"1",
            "type":"Author",
            "citationCount":2
        },
        { ⊖
            "trustScore":"",
            "topic":"All",
            "name":"C. J. Date",
            "list":"Interactive Support for Non-Programmers: The Relational and Network Approaches. (1974)",
            "id":"3",
            "type":"Author",
            "citationCount":1
        },
```

```json
"links":[
    {
        "source":"353",
        "target":"352",
        "value":1
    },
    {
        "source":"353",
        "target":"354",
        "value":1
    },
    {
        "source":"309",
        "target":"216",
        "value":1
    },
```

There is significance behind the size of each node and the thickness of the links connecting the nodes. Radius of the node is directly proportional to the citation count of the author's publications. The more the citation count, the bigger the radius of the circle of the author node. Thickness of the link between two nodes is proportional to the number of publications that the two authors have coauthored.

| Hubert D. Zimmer	1) Visuo-spatial Working Memory as a Limited Resource of Cognitive Processing. (2011)~From Resource-Adaptive Navigation Assistance to Augmented Cognition. (2011)
Stefan M	1) From Resource-Adaptive Navigation Assistance to Augmented Cognition. (2011)~Visuo-spatial Working Memory as a Limited Resource of Cognitive Processing. (2011)
Jan Brand	1) SmartKom-Home: The Interface to Home Entertainment. (2006)~The Block Interface: Accessing Digital Product Memories. (2013)~A SemProM Use Case: Maintenance of Factory and Automotive Components. (2013)~The New AI: General & Sound & Relevant for Physics. (2007)~A Digital Product Memory Architecture for Cars. (2013)~Intention Recognition. (2006)~The SemProM Data Format. (2013)~Using Basic RFID-Based Digital Product Memories for Protection against Counterfeit Goods in Manufacturing Plants. (2013)~A SemProM Use Case: Tracking & Tracing for Green Logistics and Integrity Control. (2013)~A SemProM Use Case: Maintenance of Factory and Automotive Components. (2013)~Resource-Bounded Modelling and Analysis of Human-Level Interactive Proofs. (2011)~Hardware Requirements for Digital Product Memories. (2013)~From Resource-Adaptive Navigation Assistance to Augmented Cognition. (2011)~Audience Measurement for Digital Signage: Exploring the Audience's Perspective. (2012)~Seamless Resource-Adaptive Navigation. (2011)~Applying Di
Boris Brandherm	1) A SemProM Use Case: Health Care and Compliance. (2013)~Interaction Modalities for Digital Product Memories. (2013)
Michael Schmitz	1) A SemProM Use Case: Health Care and Compliance. (2013)~Interaction Modalities for

6 Experiments and analysis

This section describes how the results of the final web application were tested.

a) DBLP dataset is huge, so there is a need to verify the accuracy of data that the application produces on querying. This verification was done using CMU professor names as input to the tool and checking if the publications and coauthors listed are reliable and trustworthy. For example, when we queried for Professor Jia Zhang, it gave her publications and coauthor network. This is shown in the figure below. This was a very important step after the development work was over.

JIA ZHANG's CoAuthorship Detail:
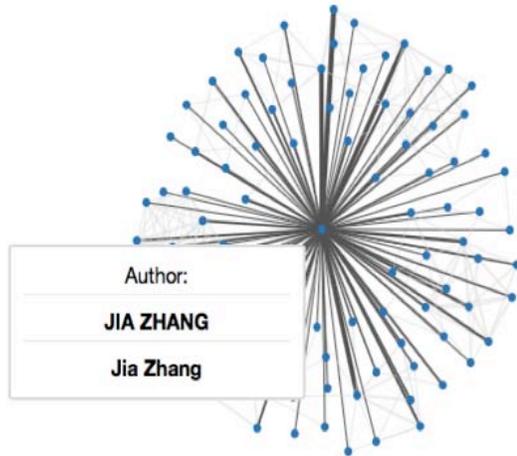
NAME	LIST
NING ZHANG	1) Auction-Based Pricing Model for Web Service Providers.
	2) Assisting Seller Pricing Strategy Selection for Electronic Auctions.
DAVID DE ROURE	1) ServiceMap: Providing Map and GPS Assistance to Service Composition in Bioinformatics.
	2) Providing Map and GPS Assistance to Service Composition in Bioinformatics.
	3) ReputationNet: A Reputation Engine to Enhance ServiceMap by Recommending Trusted Services.
CAROLE A. GOBLE	1) ReputationNet: A Reputation Engine to Enhance ServiceMap by Recommending Trusted Services.
	2) ServiceMap: Providing Map and GPS Assistance to Service Composition in Bioinformatics.
JUN WANG	1) Local Prediction of Complex Time Series Based on Support Vector Machine and Differential Evolution Algorithm.
FEI LIU	1) Start-up analysis for differential ring oscillator with even number of stages.
YUAN REN	1) An Infrastructure Supporting Considerate Sensor Service Provisioning.
	2) Supporting Personizable Virtual Internet of Things.
	3) Sensor Data as a Service - A Federated Platform for Mobile Data-centric Service Development and Sharing.
ZHENG LI	1) Leveraging Incrementally Enriched Domain Knowledge to Enhance Service Categorization.
	2) Leveraging Fragmental Semantic Data to Enhance Services Discovery.

Author:

JIA ZHANG

Jia Zhang

JIA ZHANG's Publication list:

Bridging VisTrails Scientific Workflow Management System to High Performance Computing.

Architecture-Based Development of Web Service Based Applications.

b) Dataset is not complete- After some analysis, it was found that DBLP dataset that is used as the source of information for our tool is not complete, as some attributes in it are missing for some webpage elements (author attribute is missing) as shown in the figure below.

```
<www mdate="2002-01-03" key="www/it/stet-mpeg">
<title>The MPEG Home Page</title>
<url>http://drogo.cselt.stet.it/mpeg/</url>
</www>


<www mdate="2002-01-03" key="www/org/tpc">
<title>Transaction Processing Performance Council</title>
<url>http://www.tpc.org</url>
</www>


<www mdate="2002-01-03" key="www/org/kaffe">
<title>KAFFE</title>
<url>http://www.kaffe.org/</url>
</www>


<www mdate="2009-02-11" key="www/org/oasis-open-cover-xml">
<author>Robin Cover</author>
<title>The SGML/XML Web Page</title>
<url>http://xml.coverpages.org/sgml-xml.html</url>
<year>2006</year>
</www>


<www mdate="2002-01-03" key="www/org/aiim-wfmc">
<title>WfMC Standards: The Workflow Reference Model, Version 1.1.</title>
<year>1995</year>
<url>http://www.aiim.org/wfmc/mainframe.htm</url>
</www>
```

c) Since the whole ETL framework is in place, we can import any useful scientific dataset in future and leverage our framework without spending any additional effort for building the data load process.

7   Conclusions and future work

The trust based recommendation system is a fully functional prototype that has many features useful to the research community. We could learn and experiment with many new and exciting technologies. The system supports topic based and author based searches and calculates a reliable trust score.

The project has the potential to be developed into a full fledged website and we suggest the following steps to be taken next:

a) **Improve the back end:**
   The data is currently stored in a MySQL database. With over four million publications and eight million authors, a MySQL database is not the perfect

choice for the system. As the application scales, the back-end has to be improved. We suggest a high performance database such as HANA, a NoSQL database such as mongo dB or an inverted index. While inserting new records into a database would be easier than an index re-creation, we believe that creating an inverted index would be the ideal choice for the backend.

b) Provision to track the citations:
Citation count plays an import role in deciding the relevance of the paper and in trust calculation. A provision to track citations and update them accordingly would be essential to show accurate information

c) Creating user profile
The system currently does not have a provision to create a user profile. This brings in additional validations, as we need to ensure that the information entered by the user is correct.

d) Improve trust score calculation
The trust score takes into consideration the number of publications, citations and coauthorship information. Different weights have been modeled according to the different times when the papers were published. While considering co-Authorship, we could consider different factors as the possibility of a mentorship and assign weights accordingly. For e.g., a PhD student working under a reputed professor many have published 10 papers with him/her. But that should not make him a more trusted co-author compared to another professor who might have co-authored just 6-7 papers with him. Another interesting factor to be considered would be to integrate LinkedIn profile information. Since LinkedIn is one of the largest professional networks, it could give a fair amount of information about a person's social acceptance.

e) Pagination
Currently we have options to limit the results by a number, which the user can select from a drop down. It would be nice to have a feature to show the next n relevant publications/authors

Appendix:

The following documents have been submitted along with this technical report
1) Readme file: that briefly describes the purpose of the project
2) Code
3) Access information: URL, user name/password
4) Download and installation documents with step-wise descriptions
5) Executive summary
6) Background and motivation
7) Assumptions and considerations
8) Design documents (architectural design documents and various diagrams e.g., UML files)
9) Discussions
10) Presentations (ppt file)
11) Tutorial: step-by-step usage file with screen shots included
12) Future work: to-do list and descriptions

Trust-Based Recommendation System

Sponsor: NASA

Point of Contact: Petr Votava
Faculty Advisor: Jia Zhang
Team: Krutika Kamilla, Abhishek Mukhopadhyay,
                    MustafaTasdemir, Jisha Vadake Muthiyil
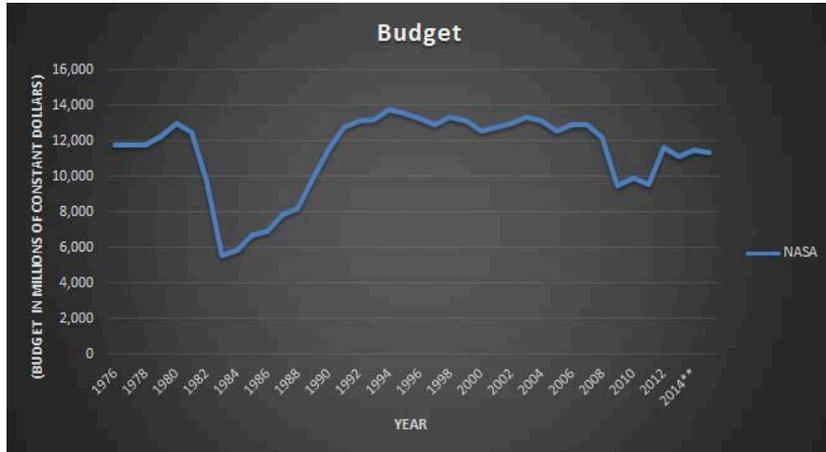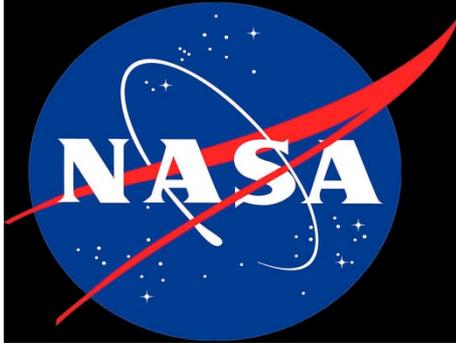
INI Project Practicum, Fall 2014

Outline

- Motivation
- Project Background
- Related work
- System design
- System implementation
- Demo
- Experiments/analysis
- Conclusions and future work

Motivation

R&D as Percent of the Federal Budget:
FY 1962-2015, in outlays

Legend: Nondefense R&D / Total Budget — Total R&D / Total Budget

Source: *Budget of the U.S. Government FY 2015* Historical Tables. FY 2014 is the request. © 2014 AAAS
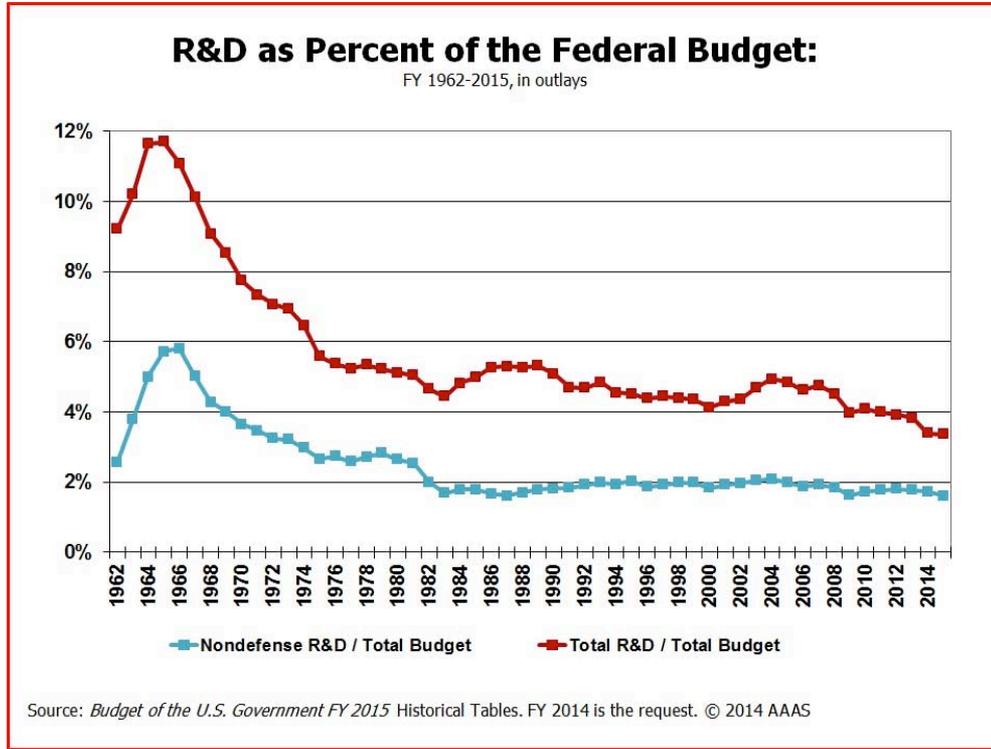
- R&D spending worldwide - $1.6 trillion, U.S. - $465 billion.

- Continuous contraction in budget through the past 50 years

- Necessitates efficient utilization of these funds.

2

Motivation

- NASA faces the same constraints. Coupled with scientists working on disparate domains but overlapping problem sets.

- Need to accelerate research through collaboration. Accelerate collaboration by making it easy to identifying collaborators working on the same problems.

3

Outline

- Motivation
- Project Background
- Related work
- System design
- System implementation
- Demo
- Experiments/analysis
- Conclusions and future work

4

Project Background

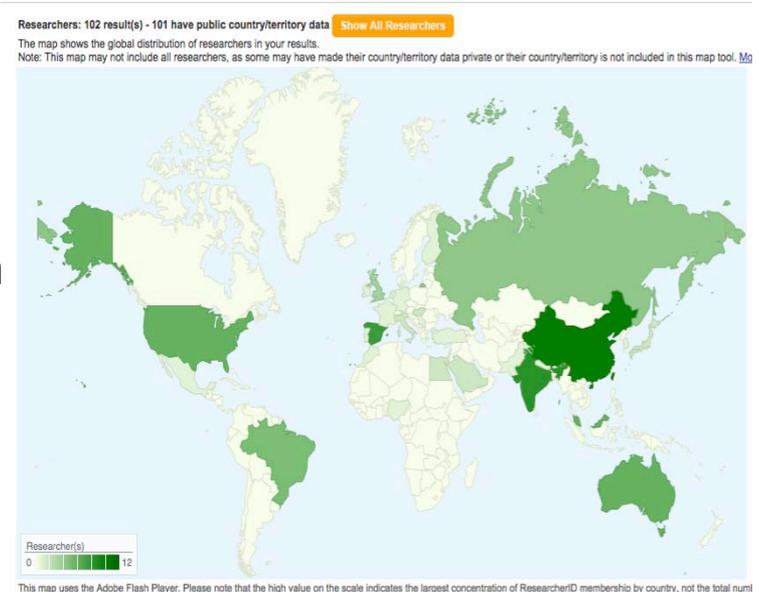- NASA's Earth Exchange (NEX) platform aims to provide a cloud-based platform as a service to accelerate big data analytics and scientific collaboration in Earth science.

- Possibility of extending to or integrating with subject domains

Project Background

- Collaboration between NASA and CMU-SV started 2 years back.

- Need of a system which recommends experts and collaborators based on domain and trust.

- Previous teams have worked on machine learning the trust score and trying to predict future collaborators.

- Need to make a working product prototype.

Outline

- Motivation
- Introduction
- Related work
- System design
- System implementation
- Demo
- Experiments/analysis
- Conclusions and future work

7

Existing products

- Many platforms exist for scientific collaboration

- Global resources and university databases

- Top ones are:
  - **ResearcherID**
  - **ResearchGate**
  - **PubMed**

Existing products

ResearcherID(Thomson Reuters)

- One of the best tool available
- Users are identified with ORCID(

    Open Researcher and Contributor ID)
- Four different search interfaces
- Option to map the geographical location
- Interactive map

    Commercial and Proprietary

Researchers: 102 result(s) - 101 have public country/territory data Show All Researchers
The map shows the global distribution of researchers in your results.
Note: This map may not include all researchers, as some may have made their country/territory data private or their country/territory is not included in this map tool. Mo

Researcher(s)
0 ▢▢▢▢▢▢ 12

This map uses the Adobe Flash Player. Please note that the high value on the scale indicates the largest concentration of ResearcherID membership by country, not the total num

9

Existing products

ResearchGate:

- Helps researchers collaborate
- Can search for researchers, publications and publish papers
- Researchers need to create a profile and upload their works
- Stats about your paper

  Search is not intuitive
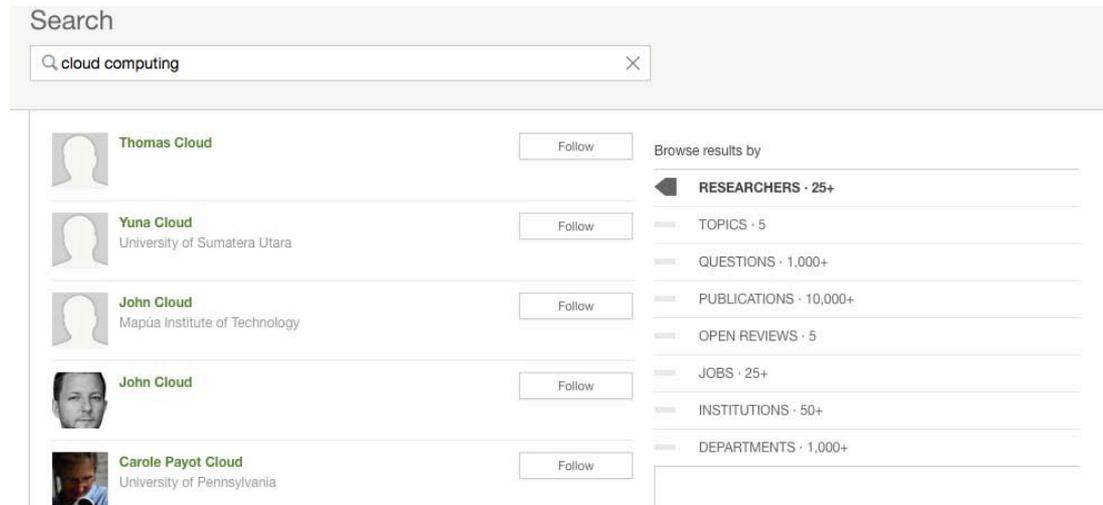
Existing products

PubMed

- Simple search and comprehensive search
- PubMed for handheld devices
- Relatively older database

Using PubMed	PubMed Tools
PubMed Quick Start Guide	PubMed Mobile
Full Text Articles	Single Citation Matcher
PubMed FAQs	Batch Citation Matcher
PubMed Tutorials	Clinical Queries
New and Noteworthy	Topic-Specific Queries

No information for collaboration

PubMed does not track citations

The GAP

- No good visualization
- No method to calculate trustworthiness
- Most of the products are discussion forums or networking sites

Anuj Tiwari asked a question in Data Mining:

Where can I get sample disaster data set to understand data mining technique and possible results ?
I want to learn data mining in context to disaster management. Where can I get sample datasets of disasters and the possible results ?

5 answers added | Follow | Answer |

Angel Santhanam asked a question in Data Mining:

How can an affinity analysis be done ? Is there any software to perform affinity analysis ?
I wish to do affinity analysis for my article based on food retailing. So please suggest a method to perform affinity analysis.

New *Be the first to answer* | Follow | Answer |

Hassan Abedi asked a question in Data Mining:

Efficient clustering algorithms for a similarity matrix?
I've got a huge similarity matrix(more precisely it's about 30000x30000 in size, and it's not a sparse matrix) now I wanted to see that what is the most efficient way I could cluster its data [m
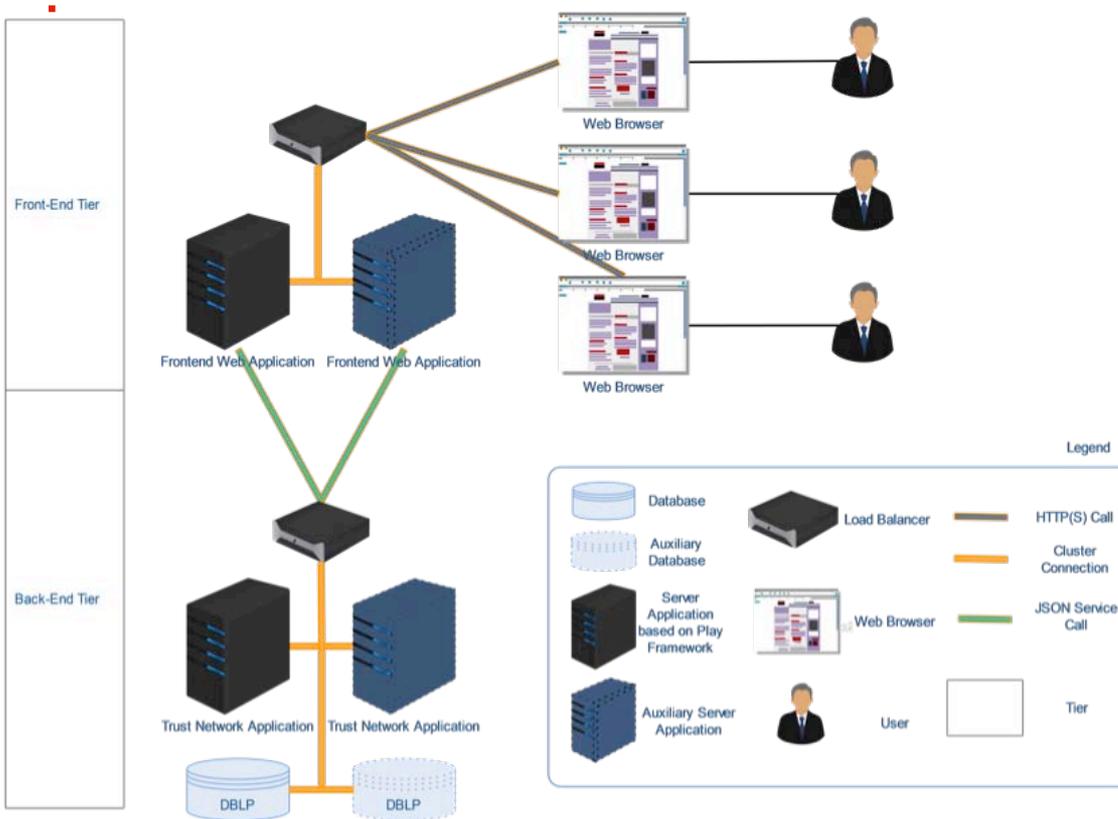
12

Outline

- Motivation
- Introduction
- Related work
- System design
- System implementation
- Demo
- Experiments/analysis
- Conclusions and future work

System Design - Quality Attributes

- Accessibility
  - Available via browser

- Usability
  - Easy and intuitive to use

- Extensibility
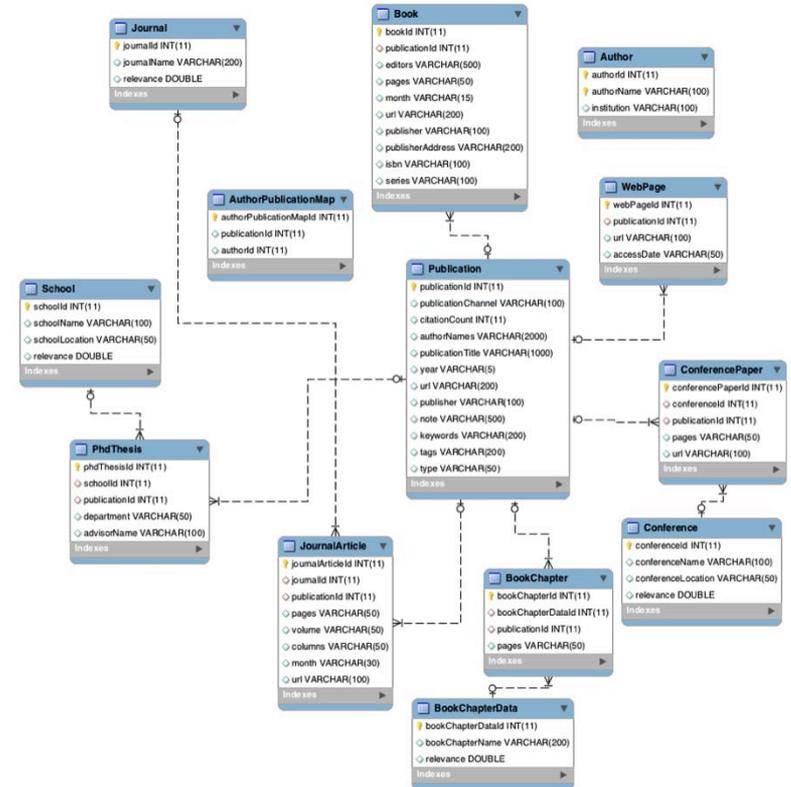 - Flexible enough to integrate new modules

System Design - Architecture

- 2-tiered architecture

- Ready for service-based API

- Horizontally Scalable

- Maintainable

15

System Design - Modeling

- Relational dataset

- Data normalization to increase efficiency

- Convenient for new data source association such as online social data

Outline

- Motivation
- Introduction
- Related work
- System design
- System implementation
- Demo
- Experiments/analysis
- Conclusions and future work

17

ETL - I

- Big Data
  - Data size ~2GB, about 4.5M publications and nearly twice as many authors
- Load and insert into database
  - Load all data into memory
  - Iteratively send to database for insertion
  - Straightforward and intuitive
- But,
  - Very inefficient due to JVM and system memory size
  - So, this inherited approach was abandoned

ETL - II

- Load DBLP and citations dataset
  - Stream data source (XML) due to huge size
  - Use of a custom parser due to file complexity
  - More complex but robust and efficient approach

- Export to files instead of directly loading
  - Iterative insertion is inefficient due to big data set
  - Using csv files to bulk load is a lot faster

Data Fetch

- MySQL database is used in the back end

- Prepared statement object is used for sending objects to the database
  - prepared statements pre-compile SQLs

- Each SQL is put into different re-usable methods

Trust Score Calculation

- Trust score consists of two components: a knowledge factor and a social factor
- Knowledge factor is built upon publication records and knowledge reputation
- Three important deciding factors of the Knowledge factor:Publication Channel (C), Publication Time(T) and citation power
  Contribution of each these factors depend on the weights(α, β, γ) assigned

Trust Score Calculation cont.

- Social factor takes into coauthorship
- Coauthorship count and time scaled coauthorship and calculated
- Time scaled coauthorship weights depend on the time of coauthorship

Final trust score = (CoauthorshipCount*TimeScaledCoauthorship) +KnowledgeFactor

Web Application - Backend

- Play framework(Java) used to serve up the data to the frontend.

- Connects to the database of choice. MySQL is in use now but multiple databases can be used if needed.

- Wraps the data in business logic required to serve up the correct Nodes and Links(edges) of the graph being queried.

- Serves up data in json format in RESTful manner, making it feasible to run this as a stand alone web service.

22

Web Application - Frontend

- Play framework(Java) used to request and receive data from the backend.

- Uses a combination of Java and Scala to render the web pages.

- Extensive use of jQuery and Ajax to handle client side events and callbacks for more data.

- Uses D3.js to render the graphs and related animations.

23

Outline

- Motivation
- Introduction
- Related work
- System design
- System implementation
- Demo
- Experiments/analysis
- Conclusions and future work

24

Demo

Below is the youtube link to a demo video
We demonstrate a few main use cases of our application in the video

https://www.youtube.com/watch?v=qPMCTP_8eo8&feature=youtu.be

Outline

- Motivation
- Introduction
- Related work
- System design
- System implementation
- Demo
- Experiments/analysis
- Conclusions and future work

26

Experiment and Analysis

- Large Dataset- need to verify the accuracy of data(Verified the search result by querying for CMU professors and checking if all their publication is listed for a topic)

- Data is not complete- some attributes were missing

- Since framework is in place, we can import any useful dataset in future and leverage the framework.

Challenges Faced and Steps Taken

- Loading DBLP dataset-exporting dblp into csv and then loading it
- Inefficient Trust calculation mechanism- improvised how the trust algorithm consumes data
- Good Visualization technique-experimented different methods
- Understanding the code base(no proper documentation)-spent more time
- Performance bottleneck-not our focus, but still we tried to add indexes, partitions in tables.

28

Outline

- Motivation
- Introduction
- Related work
- System design
- System implementation
- Demo
- Experiments/analysis
- Conclusions and future work

Conclusion and Future Work

- Final product has a reliable trust calculation technique.
- It has a robust ETL process in place for future data load
- Architecture of the project is strong enough for building new features on top of it and adding new scientific exploration dataset

- Use inverted index in the back end
- Provision to track citations
- Option to create user profile

Acknowledgement

We wish to acknowledge the contribution of Prof. Jia Zhang, for her guidance and continuous support.

We thank Petr Votava for his encouragement, feedback and support.

Thank You.

References

- http://www.battelle.org/docs/tpp/2014_global_rd_funding_forecast.pdf
- http://www.aaas.org/page/historical-trends-federal-rd
- http://www.aaas.org/sites/default/files/Budget_0.jpg
- http://www.aaas.org/sites/default/files/Agencies_0.xlsx

Questions?

Executive Summary

The 'Trust-Based Recommendation System' is a web application that facilitates scientific collaboration. The project is based on the ongoing research on collaborative systems done in conjunction with NASA and led by Prof.Jia Zhang.

The application allows users to search for publications and authors. The relation between different authors and the authors' list of publications can be visualized. There is an option to view the trust score of each author in a particular field. The trust score is calculated based on the number of papers published by the author and coauthorship. The product has got a map-view which shows the geographic regions where most of the research happens in a particular field.

The dataset used for the project is the dblp[1] dataset. The data has been stored in a MySQL database and is fetched through prepared statements. The dataload is part of the ETL module which includes a custom made parser to parse the dblp xml and load the data into the tables. The application has been developed using the Play[2] framework which consumes data in json format. The server sends the data in RESTful manner which makes it possible to separate the front end and back end and run it as a stand alone application. Play uses a combination of java and scala to render the web pages. D3.js has been used for the visualizations. Google earth API has been used for integrating map into the application. The trust score consists of two components: the knowledge factor and the social factor. The knowledge factor depends on publication channel, publication time and number of citations. The social factor depends on the coauthorship. The final trust score is a weighted sum of these factors.

The project goal was to develop an open source product for scientific exploration. We had the opportunity to learn many new technologies during the course of the development. We see the potential to develop this product into a full fledged website that would benefit the research community.

1. http://www.informatik.uni-trier.de/~ley/db/
2. https://www.playframework.com/

Project Summary

The 'Trust-Based Recommendation System' is a web application that facilitates scientific collaboration. The project is based on the ongoing research on collaborative systems done in conjunction with NASA and led by Prof.Jia Zhang.

The application allows users to search for publications and authors. The relation between different authors and the authors' list of publications can be visualized. There is an option to view the trust score of each author in a particular field. The trust score is calculated based on the number of papers published by the author and coauthorship. The product has got a map-view, which shows the geographic regions where most of the research happens in a particular field. We have used the dblp[1] dataset for the project MySQL database to store the data. The application has been developed using Play[2] framework. Play uses a combination of java and scala to render the web pages. D3.js has been used for the visualizations. Google earth API has been used for integrating map into the application. The trust score consists of two components: the knowledge factor and the social factor. The knowledge factor depends on publication channel, publication time and number of citations. The social factor depends on the coauthorship. The final trust score is a weighted sum of these factors.

The team members are, Mustafa Tasdemir, Jisha Vadake Muthiyil, Krutika Kamilla and Abhishek Mukhopadhyay. All four team members have software development background with more than two years' of work experience and are pursuing masters in Information Technology. Mustafa, Jisha and Abhishek belong to the software management track and Krutika belong to the security track. Mustafa interned as a software engineer with Quixey last summer. Jisha worked with Adobe systems as a data operations intern. Krutika interned as a security analyst with Visa and Abhishek interned with Polaris Wireless.

1. http://www.informatik.uni-trier.de/~ley/db/
2. https://www.playframework.com/

Access Information & Credentials for Source and Applications

The document describes how to access the source files of NASA Trust-Based Publication Network project and use it. It includes the credentials which are valid as of now.

1. **Source Code - GitHub**

   The source code can be found here:
   https://github.com/cmusv-sc/Practicum2014-NASA.git

   Kindly contact Professor Jia Zhang(jia.zhang@sv.cmu.edu)   or   Chris Lee(chris.lee@west.cmu.edu)   in order to get access to the repository.

2. **Application**

   The project consists of separate back-end and front-end applications running simultaneously on the Play framework. The only distinction between the two in term of accessibility is the port number defined when running each application. After you deploy any of the backend or frontend application, you can it access via the link:

   http://localhost:[port number]/

   Neither the front-end nor back-end applications are password protected. To access the database, the password of the db server is required. Backend application has application.conf file under conf folder in which you can find the database user and password. Current settings are as in the following:
   user: root
   password: nasa

3. **ETL**

   For ETL, dblp.xml file is needed since it contains all data source required. You can find it here:
   http://dblp.uni-trier.de/xml/
   You can run project named 'etl' to import data. Details are described in technical document.

Bu doküman oluşabilecek yanılsamaları ortadan kaldırmak amacıyla yazılmıştır. İlişik olan belgeler Carnegie Mellon Üniversitesi'nde 2 yıllık tam bir yüksek lisans eğitiminin sonunda Silikon Vadisi içerisindeki firmalar ve kurumlar ile işbirliği halinde yapılan proje bazlı tezlerin bilgilerini içermektedir. Bu projelerin hazırlık ve sunum aşamaları tıpkı bir tez gibidir ve anlaşılan kurum tarafından onay dahilinde ilgili bölümden mezuniyet verilmektedir. Durumla alakalı detaylı bilgi için tez danışmanı Jia Zhang (jia.zhang@sv.cmu.edu) ile iletişime geçebilirsiniz.

Mustafa Taşdemir

Assumptions and Considerations

While a conscious effort is made to make the solution as complete as possible, there are certain assumptions and considerations that have been made. The reasons for making such assumptions stem from various constraints such as time, resource availability and current technology. The following are the assumptions and considerations made and the future scope in which these can be removed.

The primary source of data for the system is the DBLP dataset. It has been assumed that the data contained within it is valid and verified. However, while all the publications can thought of as valid publications, there exists ambiguity about authors with similar names and initials. The xml schema is such that there are optional attributes which are otherwise extremely important. This some entries to miss out on information which would be extremely useful to certain use cases. Unique Ids assigned to authors, institutions and publication channels would also have removed ambiguity and made the end result more complete.

As pointed out in the previous case, with regards to missing attributes, there is a lack of overall data as well. The dataset worked with is assumed to be a complete list of all publications in computer science. While it is very significant but misses out on many publication channels due to unavailability of data from a few important publication channels like ACM. This normally would be a minor problem, however, as the trust score is being calculated, missing data could affect this final output.

Background and Motivation

Studies show that, per year $1.6 trillion is being spent on Research and development worldwide out of which U.S. alone spends $465 billion. There has been a continuous contraction in budget through the past 50 years. This necessitates efficient utilization of these funds so that it will benefit humanity to the maximum possible extent.

NASA, which is the one of the most important government agencies that fund research and development in the US face the same constraints. This is coupled with scientists working on disparate domains but overlapping problem sets. A lot of redundant research working is happening which can only be avoided if there is a proper platform which helps people find what are the publications in a particular field and who are the experts in that field. There is a need to accelerate research through collaboration. Accelerating collaboration by making it easy to identify collaborators working on the same problems is important. NASA's Earth Exchange (NEX) platform aims to provide a cloud-based platform as a service to accelerate big data analytics and scientific collaboration in Earth science. There is a possibility of extending to or integrating with subject domains too in future.

Collaboration between NASA and CMU-SV started 2 years back in order to come up with a trust based recommendation engine that would facilitate scientific collaboration. There was a need to develop a system, which recommends experts and collaborators based on domain and trust. The previous teams who worked on the trust based system developed an algorithm to calculate trust based on knowledge reputation and other social factors. Our focus was to develop a prototype which productized the large amount of research work that had been done in this field. Our aim was to develop a product that would help find existing research in a particular field, find the experts in that field, figure out of these experts who can act as collaborators and perform all of these tasks leveraging the user's trust. From the very beginning of development, our prime focus was accessibility and usability of the product

Design Document

This document explains the design decisions that were made throughout the development process. We touch on a few algorithms and decisions that were made by the previous research teams that affect our product directly.

Quality Attributes

Most of the design decisions are influenced by the quality attributes that the system should possess at any time during the development. We address these attributes according to the overall requirements of the client. We have listed the quality attributes that were given prime focus according to our sponsor's requirements.

1. **Accessibility**

   The previous teams working on the trust based recommendation system had a research focus. The prototypes they built were not accessible to the intended audience which in turn resulted in lack of proper feedback. Therefore, one of our main focuses was to make the system accessible via modern browsers at all times. This way, the product could be accessed by any user and we could integrate continuous feedback into our system

2. **Usability**

   When we achieve the accessibility, the next biggest concern will be how usable the system is. Usability has got two main aspects; the ease of usage and the performance of the system. Our focus was to improve user experience by it was a natural choice to use a graphical user interface.

3. **Extensibility**

   The systems provides users with information about publications and researchers in a particular field. It also calculates a trust score based on

knowledge and social factors. To calculate the most accurate trust score and to porvide users with the most accurate information about the research happening in any field, it is important to be able to extend the system to add varied data sources. The whole product has been developed in such a way that multiple data sources can be integrated into the system at any point of time.

System Architecture

Dynamic View

Since the system focuses on consuming big data and constructing meaningful representations out of it, dynamic flow of the data becomes an important aspect of the architecture. The following schema diagram shows this flow as well as the underlying physical devices that are required for deployment.
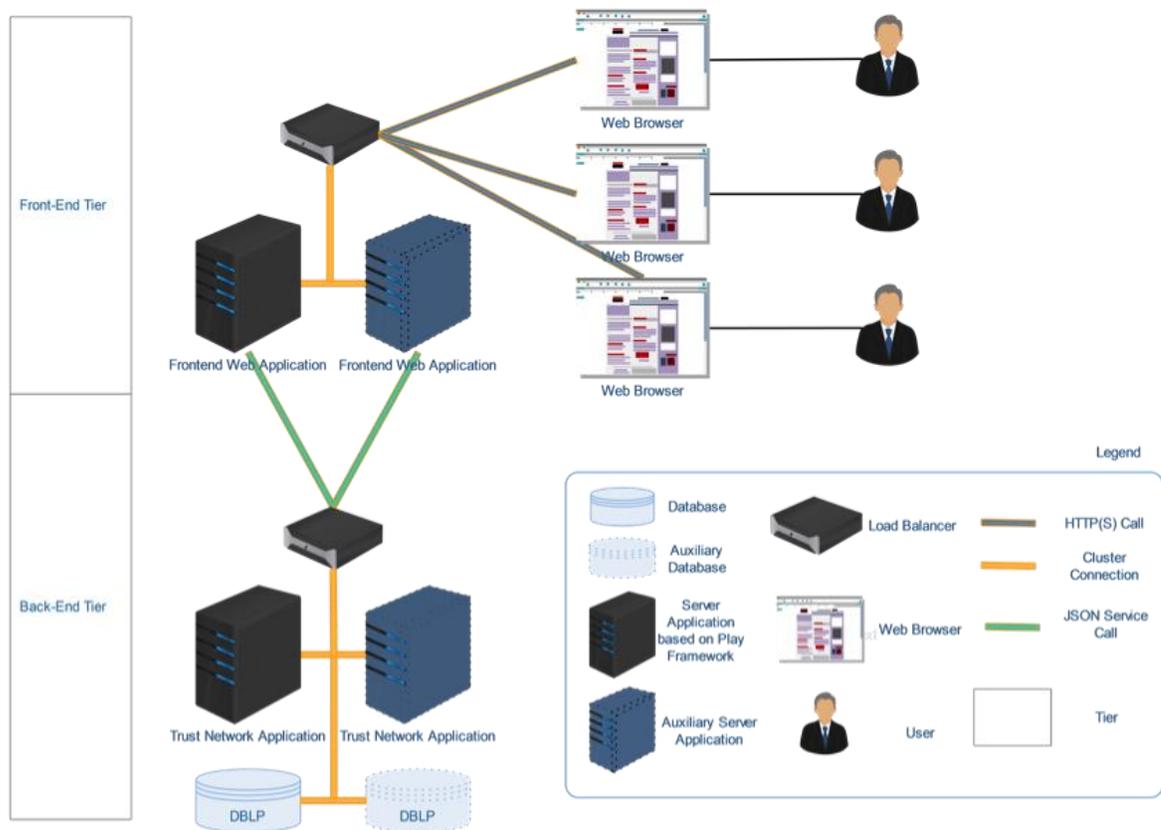
Figure 1: Dynamic View of the System

As seen in Figure 1, the overall system follows a two-tiered architecture for better maintainability. The back-end tier consists of the processed data to be consumed and the trust modeling application based on that data. The front-end tier consists of the web application to be served to the users. As shown in the figure, the number of servers can be extended according to performance requirements and scale. This will require some load balancer servers to distribute the request load to the servers in a reasonable way.

As mentioned above, the two-tier architecture let us enhance modularity of the whole system. The data is served as JSON and the front-end tier applications consume them via service calls. Therefore, the back end system is completely independant of the front-end system and can be used to serve another application in need of a similar back-end engine

Play framework has been used for both the back-end and front-end applications. Play is built on top of the asynchronous framework, Netty. Play is an open source framework developed in Java and Scala. The static nature of the back-end and front-end tiers improves the efficiency of the overall system. Another motivation to use the same framework for back-end and front end was the ability to share the models that we have used for data representations

Static View

Static view refers to models and objects defined for the system. Since the publication network includes different types of publications, we have come up with a hierarchical structure to represent the data in an efficient and easily maintainable way. The schema diagram below shows the hierarchical structure and relation between all the tables.
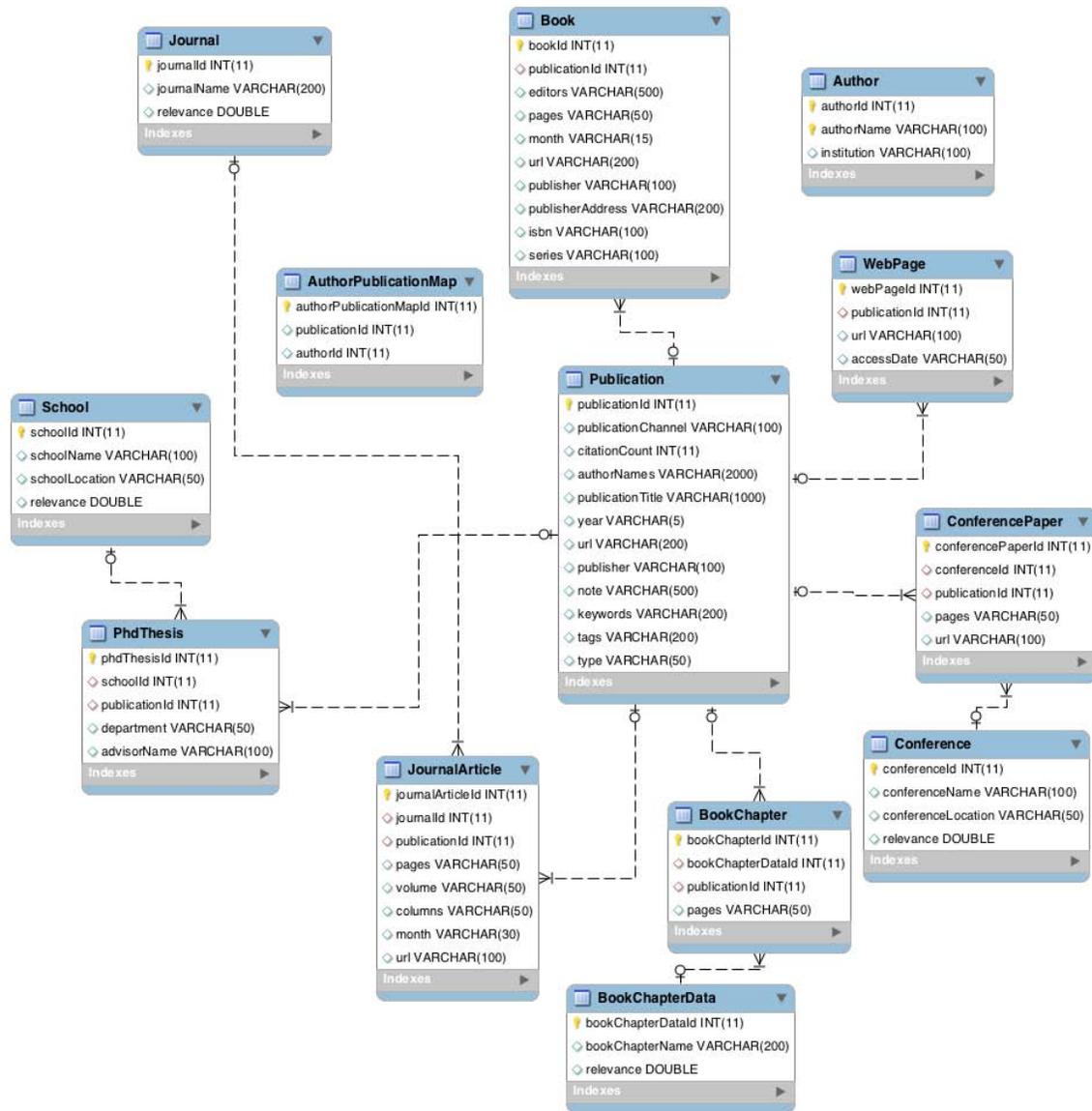
Figure 2: Database Models Schema

How Attributes are Met

The previous team who worked on this project had developed a desktop-based offline application that was not accessible to the client. Therefore, the solution of a web-based application will fulfill the accessibility requirement as suggested in previous sections. The web-based application that provides

graphical helps achieve the goal of better usability. Multi-tiered architecture as well as normalized data modeling makes the system easily extensible.

Discussions & Hints

This document provides overall discussions made within the team in order to make design decisions and choose technologies to be used in the project. The project involves big data operations, web application development and data visualization techniques. Also, the requirements include accessibility and usability. Therefore, all the decisions are to be made carefully in order to provide a maintainable system.

The first major discussion was about the choice of database.We had done extensive research on using any indexing services such as lucene to serve the back-end. Since the data was highly structured, we decided to use a SQL database. We chose MySQL as the database. Though MySQL database do not have any restrictions on the amount of data it can handle, we ran into some serious performance issues during the development period. Several optimizations techniques were employed. Since scalability was not the primary focus of this project, we did not invest a lot of time in performance improvement. We suggest the next team to focus on improving the back end which would significantly improve the overall quality of the system

The second major focus was to chose the technology for visualizations. We started development using JUNG framework. JUNG was already being used in the project and could be integrated easily with java. However, we realized that applets were needed to support JUNG for an accessible system. Because, applets are not a good choice for a usable system especially when it is about massive visualization. After consulting and confirming with our advisor, we came up with D3 javascript library to render our diagrams on the UI.

The last but not least major discussion was about the framework that we need for the application. We had chosen JSF initially in order to utilize applets in the system. However, after additional research and help from our advisor (also thanks to Hao, her PhD student), we decided to use the Play framework. Data can be served in a RESTful manner to the application. This let us isolate the frontend application from the backend.

Download and Installation Steps

1. Download Play framework- https://www.playframework.com/download and activator unix executable file and activator-launch-1.2.10.jar
2. Paste activator unix executable file and activator-launch-1.2.10.jar in /usr/local/bin folder
3. Clone the project from git and run "activator eclipse".
4. Run the existing project from its directory by going to command prompt:
   "activator run 9009" where 9009 is the port number the application is running at.

Tutorial: step-by-step

A. ETL(Data Load steps):
   1. Navigate to "etl" project under the main folder of the source files
   2. Make sure that you have 'a folder named "csv" under project "etl" and have "citations.txt" file in it.
   3. Specify your path to dblp.xml in Main class of the project
   4. Run the project (or Main class)
   5. You will have csv files generated under csv folder
   6. Use "init_db_final.sql" under dbscripts folder under the main folder and run it in MySql. This script will initialize the database and tables.
   7. Use "load_all_data_and_index.sql" script to load the csv files into MySql. Beware that you need to change the paths specified in this script as they are in your system.
   8. The ETL process should finish in under 10 minutes in total.

B. Running Nasa/backend:
   1. Open the command prompt and enter the following commands:
      > activator eclipse
      > activator "run 9010" where 9010 is the port number

C. Running Nasa/frontend:
   1. Open the command prompt and enter the following commands:
      > activator eclipse
      > activator "run 9009" where 9009 is the port number

D. Using the Web application:
   1. Open the web application at http://einstein.sv.cmu.edu:9005
      You should see the screen as shown below. Click on "Collaboration Network" button which will navigate to the next page.
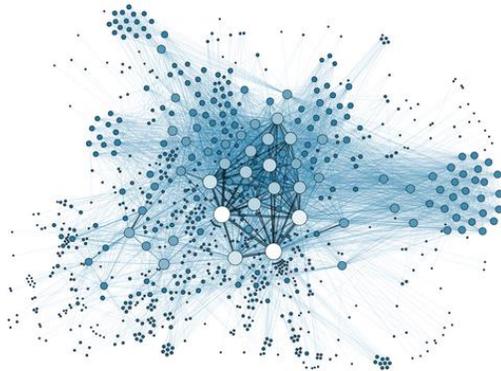
Research Community Service Platform

A data driven service to recommend papers and network collaborators for the research community

Get Started »

Generate Scientific Collaboration Network

Collaboration Network »

© Carnegie Mellon University. Silicon Valley. 2014

2. Clicking on "Collaboration Network" will navigate to page below.

Scientific Collaboration Network

Researchers have been collaborating to make scientific breakthroughs. In this project, we aim to visualize and analyze researchers' past collaboration history and base on that, help to recommend potential collaborators.(Below is a sample co-author graph for cloud computing)

Search Panel

Type of Graph   .
Author Network   Author and Publication Network   Graph Node:

Map View
Map Institutes
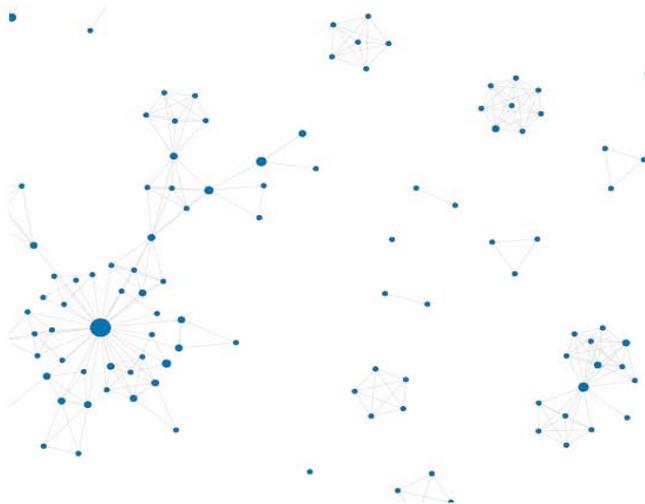
3. Clicking on "Search Panel" tab will show all the search options available as shown
     below. User can search either based on topic or author. If topic is selected, then the
user has options to enter either single topic/multiple topics separated by comma. User can also
sort by either publication count/ citation count for a topic. To see only the top 10/30/50/100

authors in that particular topic, the user has a Limit to option too.

Scientific Collaboration Network

Researchers have been collaborating to make scientific breakthroughs. In this project, we aim to visualize and analyze researchers' past collaboration history and base on that, help to recommend potential collaborators.(Below is a sample co-author graph for cloud computing)

Search Panel

- ● TOPIC
- ○ AUTHOR

Enter Topics(Multiple topics separated by comma): [   ]

Sort by: [Publication Count   ▾]

Limit to: [Less than 10   ▾]

☐ Add Trust Score

[Search]

Type of Graph .

[Author Network]   [Author and Publication Network]   Graph Node: [   ]   **Map View** [Map Institutes]

4. After selecting all these options graphs are generated- for "author-author" network and "author-publication" network.(Both the graphs are generated and rendered as and when the user clicks on author-author/author-publication network).

- ● TOPIC
- ○ AUTHOR
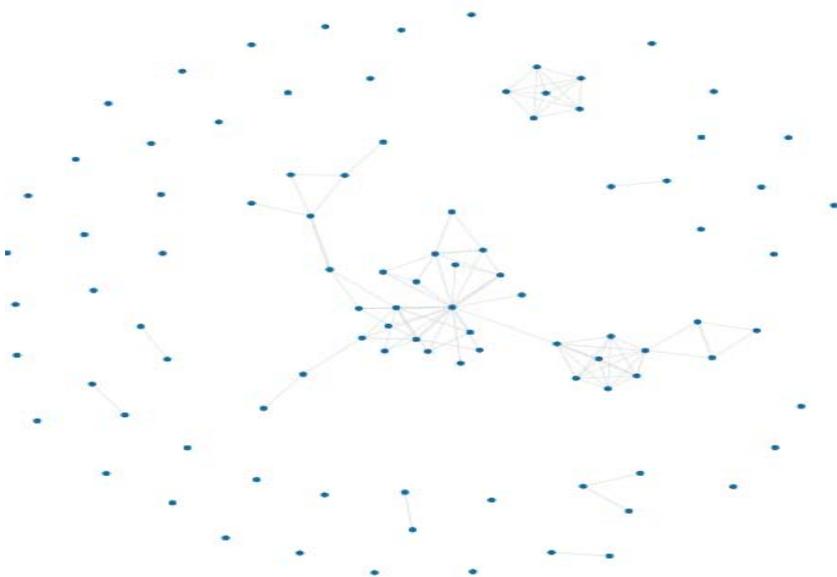
Enter Topics(Multiple topics separated by comma): [mobile]

Sort by: [Citation Count   ▾]

Limit to: [Less than 100   ▾]

☐ Add Trust Score

[Search]

Type of Graph .

[Author Network]   [Author and Publication Network]   Graph Node: [   ]   **Map View** [Map Institutes]

5. Even after the author network is generated still loading image is visible as author-publication network is being generated in the background but the user can interact with generated network.
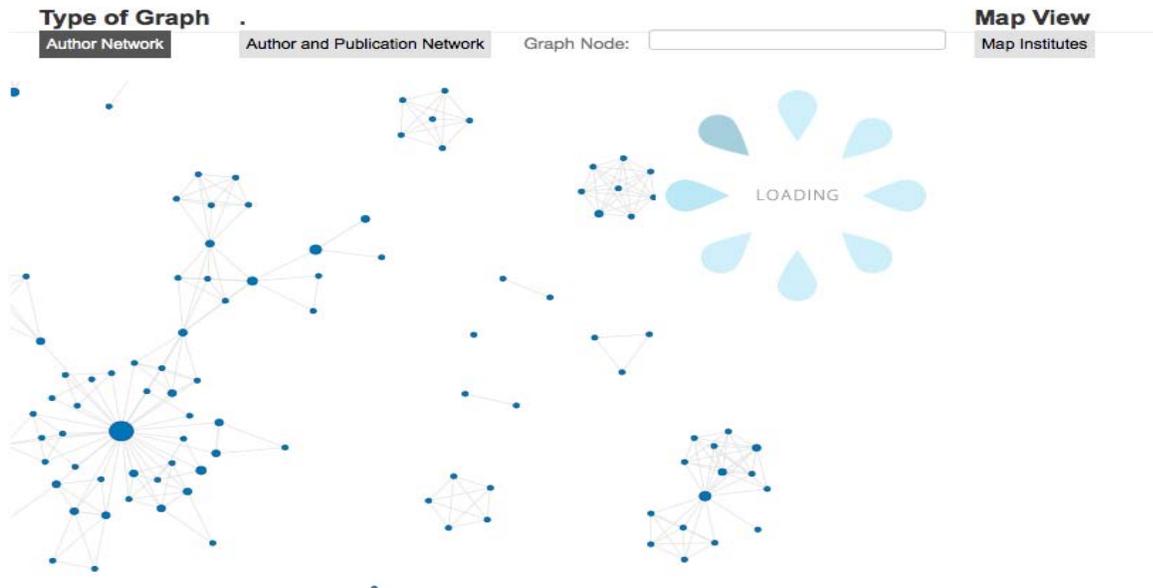
6. User can interact with the graph using mouse as follows-
    a. Hovering on the nodes, will display all the coauthors of that particular author towards the right in a tabular format as shown below.
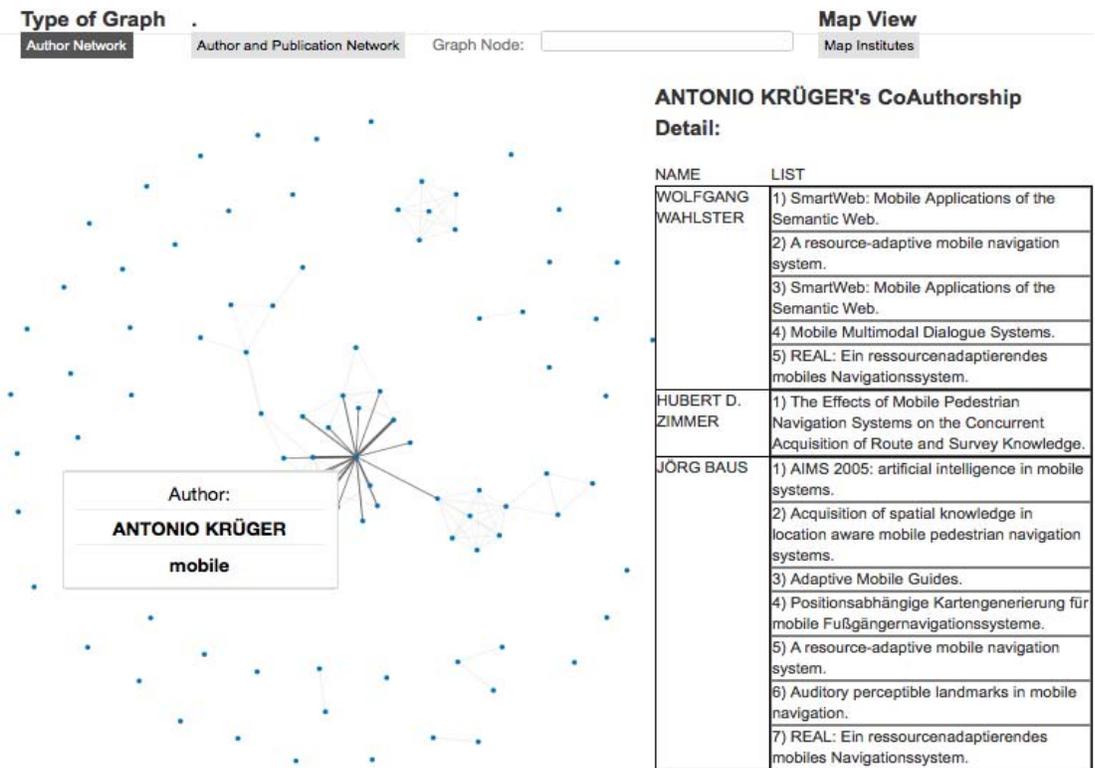
b. On clicking on the node itself, author's own publications are displayed below the graph as shown below.
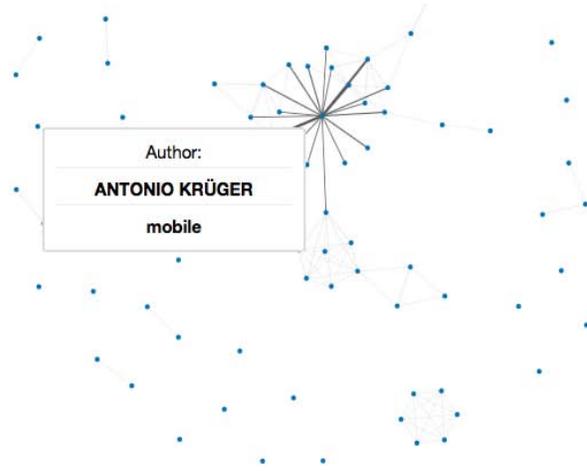
ANTONIO KRÜGER's Publication list:

Marauders light: replacing the wand with a mobile camera projector unit.

Designing gestures for mobile 3D gaming.

Scrutable User Models and Personalised Item Recommendation in Mobile Lifestyle Applications.

COMPASS2008: Multimodal, Multilingual and Crosslingual Interaction for Mobile Tourist Guide Applications.

NAME	LIST
	4) Mobile Multimodal Dialogue Systems.
	5) REAL: Ein ressourcenadaptierendes mobiles Navigationssystem.
HUBERT D. ZIMMER	1) The Effects of Mobile Pedestrian Navigation Systems on the Concurrent Acquisition of Route and Survey Knowledge.
JÖRG BAUS	1) AIMS 2005: artificial intelligence in mobile systems.
	2) Acquisition of spatial knowledge in location aware mobile pedestrian navigation systems.
	3) Adaptive Mobile Guides.
	4) Positionsabhängige Kartengenerierung für mobile Fußgängernavigationssysteme.
	5) A resource-adaptive mobile navigation system.
	6) Auditory perceptible landmarks in mobile navigation.
	7) REAL: Ein ressourcenadaptierendes mobiles Navigationssystem.
JÖRG MÜLLER	1) Experiences of Supporting Local and Remote Mobile Phone Interaction in Situated Public Display Deployments.
	2) PalmSpace: continuous around-device gestures vs. multitouch for 3D rotation tasks on mobile devices.
	3) Body, movement, gesture & tactility in interaction with mobile devices.
	4) Exploring factors that influence the combined use of mobile devices and public displays for pedestrian navigation.
KEITH CHEVERST	1) Exploring Mobile Phone Interaction with Situated Displays.

c. User can see where the graph is dense as it is zoomed out and then he can zoom-in to particular sections of the graph by mouse wheeling. The user can even drag and drop the graph to move the graph and look at other sections of the graph when zoomed-in

d. User can search for a particular author name in the graph too by entering the name of the author or just initials in the graph node textbox and the nodes in the graph containing those letters in the name of the author node will be highlighted as shown below(all nodes with "ai" in the name are highlighted as orange in the graph).

Type of Graph .

Author Network | Author and Publication Network | Graph Node: | ai

RAINER MALA

NAME	LIST
ROBERT PORZEL	1) Smar Mobile
HORST BISCHOF	1) Multi guarant
	2) Mobil Illumina
FAISAL TAHER	1) Form support mobile
	2) Explo navigati mobile a
	3) Photo of public mobile
	4) Explo Display Formati
RAINER STEFFEN	1) A sys transmis applicat
RAINER MALAKA	1) pitchl Explorin
	2) Lister Represe Commu
	3) Artific
	4) Explo folksono

Author:

RAINER MALAKA

mobile

e. The other kind of network is displayed by clicking on "Author and Publication network" on top and the author nodes are dark blue in color while the publication network is light blue in color. User can interact with the graph in the similar way as author-author graph. Similarly the table on the right side displays details about the authors who have coauthored that particular publication on hovering on a publication node.(first figure below shows the author node and the second figure below shows a publication node)
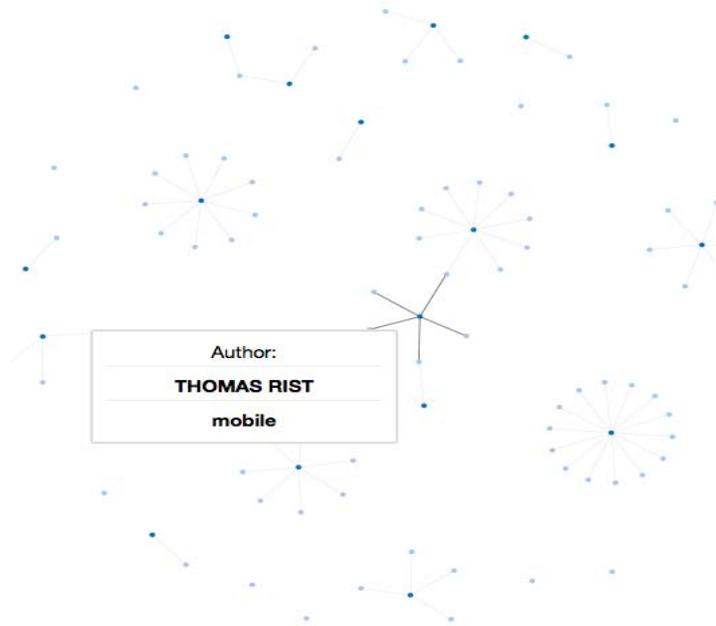
THOMAS RIST's CoAuthorship Detail:

NAME	LIST
Aligning information browsing and exploration methods with a spatial navigation aid for mobile city visitors.	1) THOMAS RIST 2) STEPHAN BALDES 3) PATRICK BRANDMEIER
Getting the Mobile Users in: Three Systems that Support Collaboration in an Environment with Heterogeneous Communication Devices.	1) THOMAS RIST 2) ELISABETH ANDRÉ 3) GERD HERZOG 4) PATRICK BRANDMEIER
Using Mobile Communication Devices to Access Virtual Meeting Spaces.	1) THOMAS RIST
Customizing Graphics for Tiny Displays of Mobile Devices.	1) THOMAS RIST 2) PATRICK BRANDMEIER
Visualisierungstechniken für mobile Geräte mit kleinen Bildschirmen.	1) THOMAS RIST 2) PATRICK BRANDMEIER
ELISABETH ANDRÉ	1) Direct, bodily or mobile interaction?: comparing interaction techniques for personalized public displays. 2) Mobile augmented reality and...

Author:

THOMAS RIST

mobile

Tool-Supported User-Centred Prototyping of Mobile Applications.'s CoAuthorship Detail:

NAME	LIST
ELISABETH ANDRÉ	1) Direct, bodily or mobile interaction?: comparing interaction techniques for personalized public displays.
	2) Mobile augmented reality and adaptive art: a game-based motivation for energy saving.
	3) Role Assignment Via Physical Mobile Interaction Techniques in Mobile Multi-user Applications for Children.
	4) Getting the Mobile Users in: Three Systems that Support Collaboration in an Environment with Heterogeneous Communication Devices.
	5) User-Centred Development of Mobile Interfaces to a Pervasive Computing Environment.
	6) EVAL - an evaluation component for mobile interfaces.
	7) Tool-Supported User-Centred Prototyping of Mobile Applications.
	8) Social Mobile Interaction using Tangible User Interfaces and Mobile Phones.
	9) Using the mobile application EDDY for gathering user information in the requirement analysis.

Publication:

Tool-Supported User-Centred Prototyping of Mobile Applications.

mobile

Tool-Supported User-Centred Prototyping of Mobile Applications.'s Publication list:

Type of Graph .

Author Network | Author and Publication Network | Graph Node: [   ]

Map View

Map Institutes

Publication:

Role Assignment Via Physical Mobile Interaction Techniques in Mobile Multi-user Applications for Children.

mobile

Role Assignment Via Physical Mobile Interaction Techniques in Mobile Multi-user Applications for Children.'s CoAuthorship Detail:

NAME	LIST
ELISABETH ANDRÉ	1) Direct, bodily or mobile interaction?: comparing interaction techniques for personalized public displays.
	2) Mobile augmented reality and adaptive art: a game-based motivation for energy saving.
	3) Role Assignment Via Physical Mobile Interaction Techniques in Mobile Multi-user Applications for Children.
	4) Getting the Mobile Users in: Three Systems that Support Collaboration in an Environment with Heterogeneous Communication Devices.
	5) User-Centred Development of Mobile Interfaces to a Pervasive Computing Environment.
	6) EVAL - an evaluation component for mobile interfaces.
	7) Tool-Supported User-Centred Prototyping of Mobile Applications.
	8) Social Mobile Interaction using Tangible User Interfaces and Mobile Phones.
	9) Using the mobile application EDDY for gathering user information in the requirement analysis.

Role Assignment Via Physical Mobile Interaction Techniques in Mobile Multi-user Applications for

7. User can search for multiple topics too as shown below(eg. mobile,big data below).

○ TOPIC
○ AUTHOR

Enter Topics(Multiple topics separated by comma): [mobile,big data]

Sort by: [Citation Count ▼]

Limit to: [Less than 100 ▼]

[Search]

Type of Graph .

Author Network | Graph Node: [   ]

Map View

Map Institutes

Author:

FRANCISCO FACCHINEI

big data

FRANCISCO FACCHINEI's CoAuthorship Detail:

NAME	LIST
GESUALDO SCUTARI	1) Flexible parallel algorithms for big data optimization.
SIMONE SAGRATELLA	1) Flexible parallel algorithms for big data optimization.

8. User can see the trust scores of different authors on a particular topic search by clicking on the checkbox which says "Add trust score". This score depends on the author's publications and his coauthors.

ANTONIO KRÜGER's CoAuthorship Detail:

NAME	LIST
CHRISTOPH STAHL	1) Tool-support for mobile and pervasive application development - issues and challenges.
	2) Positionsabhängige Kartengenerierung für mobile Fußgängernavigationssysteme.
	3) Robust speech interaction in a mobile environment through the use of multiple and different media input types.
BOB KUMMERFELD	1) Safe Execution of Dynamically Loaded Code on Mobile Phones.
	2) Scrutable User Models and Personalised Item Recommendation in Mobile Lifestyle Applications.
	3) PersonisJ: Mobile, Client-Side User Modelling.
JÖRG BAUS	1) AIMS 2005: artificial intelligence in mobile systems.
	2) Acquisition of spatial knowledge in location aware mobile pedestrian navigation systems.
	3) Adaptive Mobile Guides.

9. User can search for a particular author to see his network of coauthors by selecting author search option. This also gives the similar kind of sort and limit options.
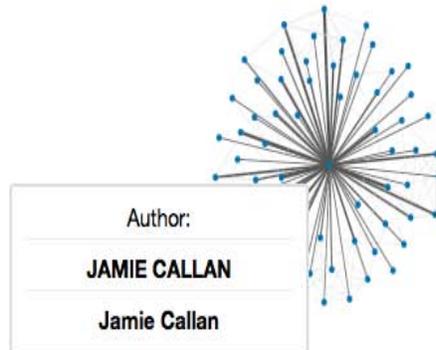
10. Below graph shows Jamie Callan's publications and his coauthors.

Author:

JAMIE CALLAN

Jamie Callan

JAMIE CALLAN's Publication list:

Pairwise Document Classification for Relevance Feedback.

Reducing Storage Costs for Federated Search of Text Databases.

Ontology generation for large email collections.

NAME	LIST
JIE LU	1) Combining Multiple Resources, Evidences and Criteria for Genomic Information Retrieval.
	2) Full-text federated search of text-based digital libraries in peer-to-peer networks.
	3) Federated Search of Text-Based Digital Libraries in Hierarchical Peer-to-Peer Networks.
	4) Content-Based Peer-to-Peer Network Overlay for Full-Text Federated Search.
	5) Reducing Storage Costs for Federated Search of Text Databases.
	6) Merging retrieval results in hierarchical peer-to-peer networks.
	7) Federated Search of Text-Based Digital Libraries in Hierarchical Peer-to-Peer Networks.
REMA PADMAN	1) Incremental hierarchical clustering of text documents.
PINAR DONMEZ	1) Pairwise Document Classification for Relevance Feedback.
JAIME G. CARBONELL	1) Stuctured Queries for Legal Search.
	2) Retrieval and feedback models for blog feed search.
	3) The impact of history length on personalized search.
	4) Document Representation and Query Expansion Models for Blog Recommendation.
	5) Retrieval and Feedback Models for Blog Distillation.
	6) Pairwise Document Classification for Relevance Feedback.
	7) Document and Query Expansion Models for Blog Distillation.
EDUARD H. HOVY	1) SGER Collaborative: A Testbed for eRulemaking Data

11. User can also check author-publication graph of a given author by clicking on "Author and Publication Network" button on top of the graph.

Type of Graph .

Author Network | Author and Publication Network | Graph Node: | Map View | Map Institutes

JAIME G. CARBONELL's CoAuthorship Detail:

Author:
JAIME G. CARBONELL
Jamie Callan

NAME	LIST
PINAR DONMEZ	1) Pairwise Document Classification for Relevance Feedback.
Pairwise Document Classification for Relevance Feedback.	1) PINAR DONMEZ 2) JAIME G. CARBONELL 3) JAMIE CALLAN 4) JONATHAN L. ELSAS
Stuctured Queries for Legal Search.	1) JAIME G. CARBONELL 2) JAMIE CALLAN 3) YANGBO ZHU 4) LE ZHAO
Retrieval and feedback models for blog feed search.	1) JAIME G. CARBONELL 2) JAMIE CALLAN 3) JAIME ARGUELLO 4) JONATHAN L. ELSAS
The impact of history length on personalized search.	1) JAIME G. CARBONELL 2) JAMIE CALLAN 3) YANGBO ZHU
Document Representation and Query	1) JAIME G. CARBONELL 2) JAMIE CALLAN 3) JAIME ARGUELLO

12. Clicking on the "Map View" button takes you to the following page.

Carnegie Mellon University Silicon Valley

Topic machine learning | Search

Data SIO, NOAA, U.S. Navy, NGA, GEBCO
© 2014 Google
US Dept of State Geographer
© 2009 GeoBasis–DE/BKG

Google earth
Terms of Use
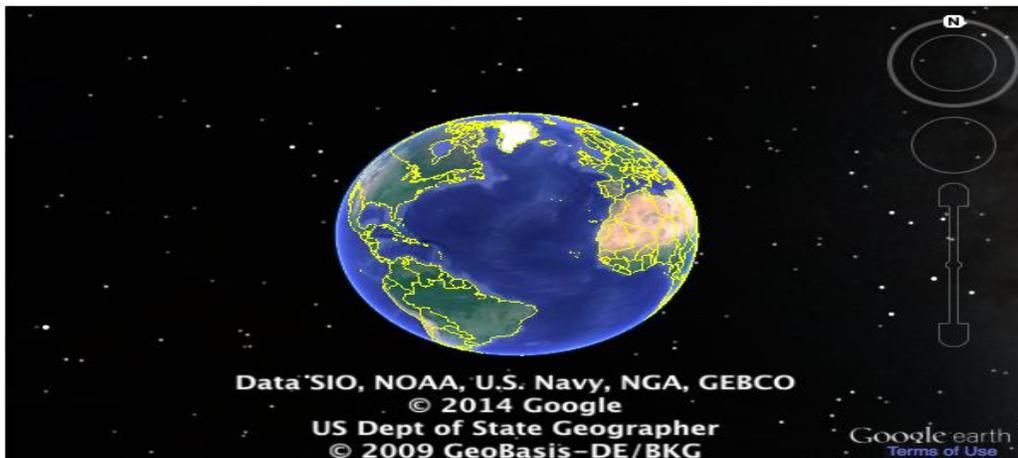
This search is for schools that produce most number of PhD thesis in the world on a particular topic. on searching for "machine learning" the following schools are highlighted.