Scholar’s Search Engine
A scholar’s search and visualisation System

Pradeep Vairamani
Ankita Jain
Ashish Kumar
Shubham Agrawal
Ashita Prasad

Introduction

The Scholar’s Search Engine is an alternative to existing academic search engines, with better organization and visualization for similar scholar’s based on a wide range of features such as field of study, papers, collaborations etc. It is targeted towards the research community to help them explore researchers in an easy and meaningful manner by expanding their network. We hope that this project can lead to a more open and collaborative research environment, where people can find researchers with similar interests easily.

Problem definition

Currently researchers use text based academic search engines which are not particularly intuitive or easy to use. One of the limitations of current practice is that search results only show papers, without showing other researchers working on similar topics. We intend to make it easy for users to find similar researchers with an easy to use bubble based visualisation.

Survey

[1] This paper describes the Microsoft academic service which is the basis for most of the data that we will be using for this project. We learned from the recommendation model in this paper to implement our similarity algorithm.
[2, 3] It provides an idea to build a framework for developing similarity based methods. This helps us in our project as we need to run different kinds of clustering based on similarity, such as affiliation and field of study on our data.
[4, 6] We are dealing with a big dataset and are trying to provide quick query responses, so our analytics need to work fast. This paper tells us how to speed it up by keeping the data in memory for most of the operations.

[9, 14] These papers propose ways to subspace clustering approach for recommender systems.

[10] This paper used geometry and shading approaches in order to visualize hierarchical data, and a cushion treemap method which could be used to manage large data sets. We don’t intend to use new visualization methods in our project, these ideas help us think about possible ways in which we can think about hierarchies of data.

[12] This paper proposed a personalized academic research recommender system by defining text similarity between two research papers. We will be applying a similar approach to find similarity between keywords in research papers to show most relevant search results.

Proposed Method

Intuition : Our approach combines traditional searching with visualisation. We would provide similarity among authors based on their fields of studies, collaborations, affiliation etc. Such an approach will be successful because there is a dearth of such an intuitive search, which will help researchers find people working on similar fields, and potential collaborators.

Description of approaches : Content based approach has been adopted for the design and implementation of our recommendation system. This approach uses contents describing the items such as papers, journals and authors and the user input query for an author. The high level architecture of our content based recommender system has the following components:

1. **Content Analyzer**: This component is responsible for converting the raw data to a structured information representation. This is one of the pre-processing step which is needed to convert the original information space to vector space model (VSM) by extracting features.
2. **Filtering Component**: This component utilizes user input and suggest most relevant items by matching the user input against the items to be recommended by computing the similarity metrics.

The idea is to provide top 10 similar authors for the author queried by the user based on their field of study attribute. Since, the Microsoft Academic Graph dataset has sparse set of around 50K field of studies, we tried to group these field of studies using K-means based on their significance across various journals. Further, contribution of each author to these grouped field of studies has been computed. The contribution of each author is represented as vector of weights. Finally, similarity between queried author and all other authors will be computed to give the top ten similar authors.

Item Representation as Vector Space Model: For reducing field of studies, we find out the weights of each field of study in various journals. TF-IDF is used as follows:

$$TF(f,j) = \frac{\#papers\ associated\ to\ fos\ f\ presented\ in\ journal\ j}{Max\ \#papers\ associated\ with\ any\ fos\ in\ journal\ j}$$

$$IDF(f,j) = \log \frac{\#\ journals}{\#\ journals\ in\ which\ atleast\ one\ paper\ associated\ to\ fos\ f\ appears}$$

$$w(f,j) = \frac{TF(f,j) * IDF(f,j)}{\sqrt{\sum_{f=1}^{F} (TF(f,j) * IDF(f,j))^2}}$$

With this, each field of study is represented as a vector of weights. On these vectors, we performed K-means to group field of studies into K clusters. After reducing field of studies, we represent each author with the normalized weight vectors using TF-IDF showing their degree of association to the particular field of study.

$$TF(a,f') = \frac{\#papers\ presented\ by\ author\ a\ in\ fos\ f}{Max\ \#papers\ presented\ by\ any\ author\ in\ fos\ f}$$

$$IDF(a,f') = \log \frac{\#fos}{\#fos\ in\ which\ atleast\ one\ paper\ is\ of\ author\ a}$$

$$w(a,f) = \frac{TF(a,f') * IDF(a,f')}{\sqrt{\sum_{a=1}^{A} (TF(a,f') * IDF(a,f'))^2}}$$

Similarity Measures: Once, we have vector representation of all authors, we will compute the cosine similarity among queried author and other authors as follows:

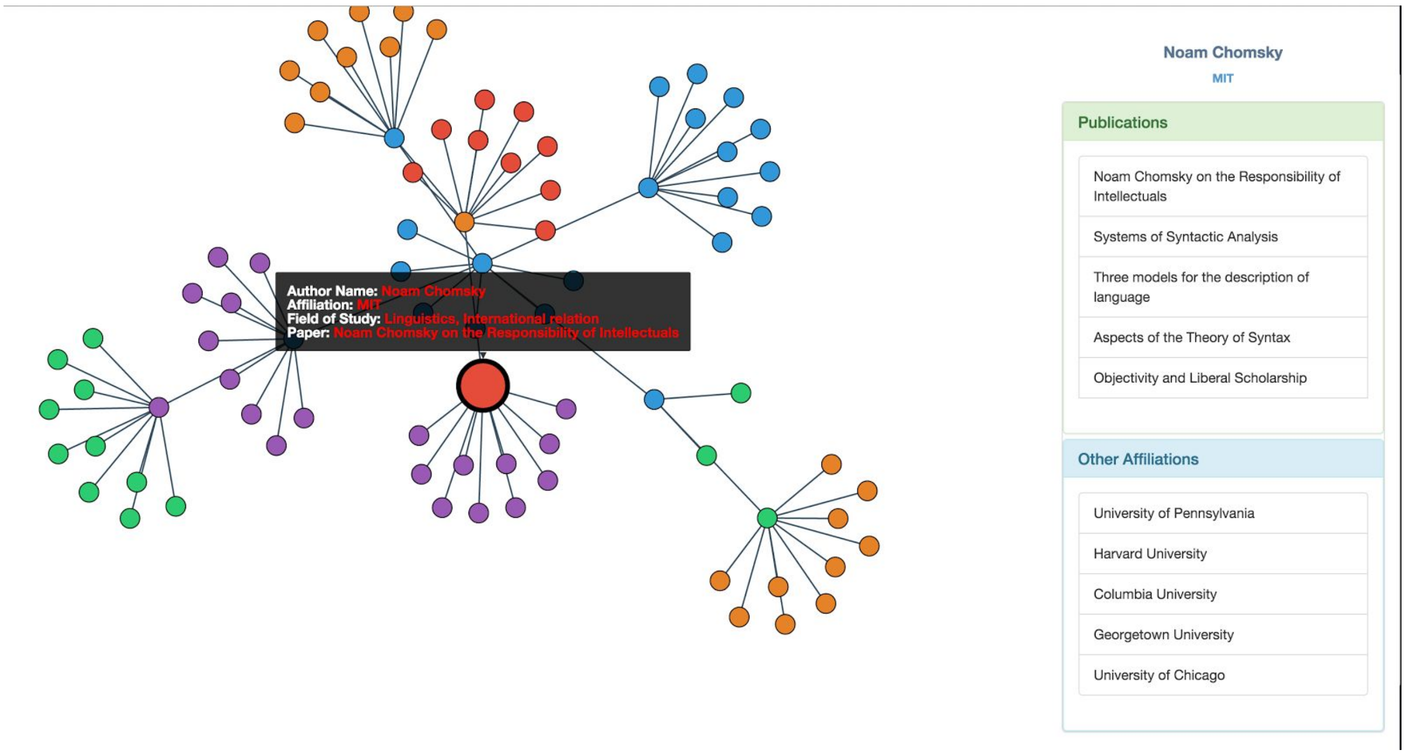
$$sim(a_{i,} a_j) = \frac{\sum_K w(k,i) * w(k,j)}{\sqrt{\sum_K w(k,i)^2} * \sqrt{\sum_K w(k,j)^2}}$$

Algorithm:

```
a. START
b. Get the author Aj queried by user.
c. Determine the normalized weight of each field of study (F) using TF-IDF in
   journal J. This represents the importance of a field of study F in the journal J.
d. Run K-means on field of studies and group them into K clusters. This will provide
   reduced K field of study representation.
e. Represent author Aj as a vector of normalized weight that measures the
   association of author (Aj) in particular reduced field of study.
f. FOR i=1 to N
           i) Represent author Ai as a vector of weights.
           ii) Compute similarity of author Ai to author Aj using cosine similarity.
           iii)Store similarity value Sim_Values(Aj,i)=Sim(Aj,Ai)
   NEXT
g. Sort Sim_Values(Aj) in descending order with respect to similarity value.
h. THRESHOLD of Sim_Values to DISPLAY top 10 similar authors.
e. END
```

Frontend Framework: We have implemented the two main components that will form the major portion of the frontend. We have the bubble based interaction for displaying the similarity among authors. We also have the sidebar, which displays the necessary information about the individual authors.

Screenshot of the Progress so far



Author Name: Noam Chomsky
Affiliation: MIT
Field of Study: Linguistics, International relation
Paper: Noam Chomsky on the Responsibility of Intellectuals

Noam Chomsky

MIT

Publications

- Noam Chomsky on the Responsibility of Intellectuals
- Systems of Syntactic Analysis
- Three models for the description of language
- Aspects of the Theory of Syntax
- Objectivity and Liberal Scholarship

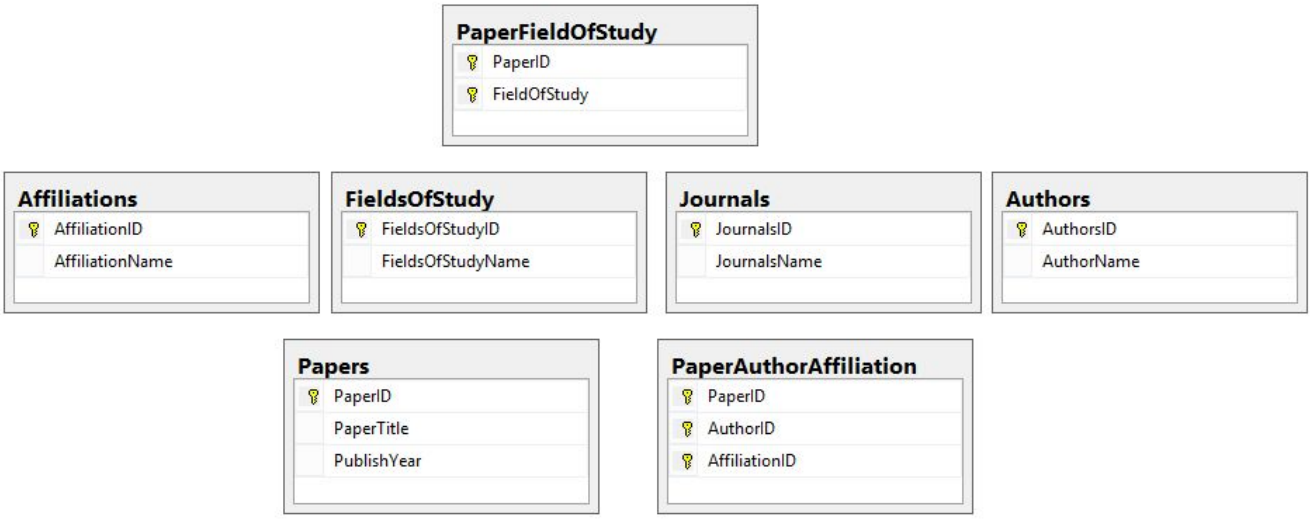
Other Affiliations

- University of Pennsylvania
- Harvard University
- Columbia University
- Georgetown University
- University of Chicago

Backend Framework: We have implemented the framework for the UI to talk to the data. There is a python server using the bottle library which will handle the API calls. The server can also handle Cross-origin resource sharing. The 3 APIs that will be used are :

1) /similar/<author_id> - Which returns a json with 10 similar authors with some metadata about each author.
2) /details/<author_id> - This API returns the details about the individual author. This information is used to populate the sidebar.
3) /search/<author_name> - This API is used to populate the first node based on a query in the search bar.

Data Model:



PaperFieldOfStudy
🔑 PaperID
🔑 FieldOfStudy

Affiliations
🔑 AffiliationID
AffiliationName

FieldsOfStudy
🔑 FieldsOfStudyID
FieldsOfStudyName

Journals
🔑 JournalsID
JournalsName

Authors
🔑 AuthorsID
AuthorName

Papers
🔑 PaperID
PaperTitle
PublishYear

PaperAuthorAffiliation
🔑 PaperID
🔑 AuthorID
🔑 AffiliationID

<u>Experiments/Evaluations</u>: Using R, AWS, AWK, SSMS, PIG, with the data at [5]. For learning, we used the data of Papers(PaperID, PublishYear, JournalID), PaperAuthorAffiliations(PaperID, AuthorID) and PaperKeywords(PaperID, FieldID). As the uncompressed data was more than 50GB, AWK was used to segment the mentioned columns. We used R for exploratory data analysis (e.g. number of different field of studies, papers published per year) for arriving at the data filtering and learning algorithms decisions. We used PIG on AWS to aggregate number of papers published per author. Interestingly, more than 90M authors(among ~125M) have published only 1 paper. For our prototype, we first plan to take a subset of authors who have published more than 10 papers and then lower the threshold to 5 and eventually to 1. The goal is to generate similarity matrix for the subset of authors and evaluate that before moving to a larger set. Papers data have JournalID mapped to ~45M PaperIDs, we intend to use this data combined with PaperKeywords(PaperID, FieldID) data to group Field of studies. Hypothesis for this grouping is, papers published in a journal tend to be in a related field of studies. So, if different field of studies have similar contribution across journals, they are very much related.

<u>List of Innovations</u>

1) Our approach combines traditional searching with visualisation. Most of the existing academic search engines are purely text based.
2) We have used K-means to cluster fields of study based on their significance across various journals.

3) We will be using TF-IDF to compute degree of association among field-of-studies through journals and among authors through associated field of studies.

Conclusion: [Placeholder]

Distribution of Team Member Effort: All team members contributed similar amount of efforts.

APPENDIX

Plan of activities

Frontend:

1) We need to add relevant information about the authors in Tooltips.
2) Work on the colors for the bubbles (The recent author search should be made more prominent).
3) Integrate the frontend with real data from the backend.

Backend:

1) Generate feature vectors for Field of Study(keywords for papers) representing their contributions through papers in different journals.
2) Implement clustering algorithm to group field of studies using above data
3) Filter data for a subset of authors having published more than 10 papers
4) Generate feature vectors for authors representing their contributions through papers in different group of field of studies
5) Implement algorithm to learn similarity among authors using above features data.

Time estimates:

Revised Plan: The activities that are in blue have already been completed. The activities in Orange are semi-complete.

	Week 1	Week 2-5	Week 6-9	Final week
Pradeep, Ashish	Brainstorm over topic, prepare project proposal, presentation	Backend Framework for frontend-backend communication	Work on visualization, Final integration of the backend and frontend	Prepare report, Give final touches to visualization

Ankita, Shubham, Ashita	Brainstorm over topic, prepare project proposal, presentation	Preprocess/ group data through awk/PIG scripts and RStudio. Devise algorithm to classify field of studies and find author similarities.	Work on creating feature vectors,find similarity and rank the result, final integration of the backend and frontend	Evaluate performance,prepare final report.

Original

	Week 1	Week 2-5	Week 6-9	Final week
Pradeep	Brainstorm over topic, prepare project proposal, presentation	Backend Framework for frontend-backend communication	Work on visualization, Final integration of the backend and frontend	Prepare report, Give final touches to visualization
Ankita	Brainstorm over topic, prepare project proposal, presentation	Work on for field of study classifier.	Work on visualization, Final integration of the backend and frontend	Prepare report
Shubham	Brainstorm over topic, prepare project proposal, presentation	Work on for field of study classifier.	Machine learning technique to find similarity and rank the result	Prepare report
Ashish	Brainstorm over topic, prepare project	Machine learning	Preprocessing grouping/techn	Give final touches to visualization,

	proposal, presentation	technique to find similarity and rank the result	iques for backend	Prepare report
Ashita	Brainstorm over topic, prepare project proposal, presentation	Backend Framework, work on the UI	Preprocessing grouping/techniques for backend	Prepare report

References:

[1] Arnab Sinha, Zhihong Shen, Yang Song, Hao Ma, Darrin Eide, and Kuansan Wang, “An Overview of Microsoft Academic Service (MAS) and Applications, WWW – World Wide Web Consortium (W3C)”, 18 May 2015.

[2] Włodzisław Duch, “Similarity-based methods: a general framework for classification, approximation and association”, Available: http://www.fizyka.umk.pl/publications/kmk/00cc-kn.pdf

[3] “Similarity based Decision Models”, Available: http://www.econ.core.hu/file/download/korosi/2010/kovacs10.pdf

[4] “Using In-Memory Analytics to Quickly Crunch Big Data”, Available: http://www.computer.org/csdl/mags/co/2012/10/mco2012100016.pdf

[5] https://academicgraph.blob.core.windows.net/graph-2015-08-20/index.html

[6] Nitin Agarwal, Ehtesham Haque, Huan Liu, and Lance Parsons Arizona State University, Tempe AZ 85281, USA - “Research Paper Recommender Systems: A Subspace Clustering Approach”

[7]Joeran Beel, Bela Gipp, Stefan Langer, and Corinna Breitinger, “Research Paper Recommender Systems: A Literature Survey” International Journal on Digital Libraries, Springer Berlin Heidelberg, 26 Jul 2015, Available: http://docear.org/papers/research_paper_recommender_system_evaluation--a_quantitative_literature_survey.pdf

[8] Available: http://files.grouplens.org/papers/techlens-cscw2002.pdf

[9] Megha Jain, “ALGORITHM FOR RESEARCH PAPER RECOMMENDATION SYSTEM”,International Journal of Information Technology and Knowledge Management

[10] Jarke J. van Wijk, Frank van Ham, Huub van de Wetering , “Rendering Hierarchical Data”, Available: http://www.win.tue.nl/~vanwijk/rhd.pdf

[11] Balabanovic, M., Shoham, “Fab: Content-based, Collaborative Recommendation. Communications of the ACM”

[12] Joonseok Lee, Kisung Lee, Jennifer G. Kim, “Personalized Academic Research Paper Recommendation System”

[13] Jie Tang, Duo Zhang, Limin Yao, ”Social Network Extraction of Academic Researchers”

[14] “An index to quantify an individual’s scientific research output”, Available: http://arxiv.org/pdf/physics/0508025.pdf

[15] Jöran Beel Bela Gipp, “Google Scholar‘s Ranking Algorithm: An Introductory Overview.” In Birger Larsen and Jacqueline Leta, editors, Proceedings of the 12th International Conference on Scientometrics and Informetrics (ISSI’09), volume 1, pages 230–241, Rio de Janeiro (Brazil), July 2009. International Society for Scientometrics and Informetrics. ISSN 2175-1935, Available: http://www.sciplore.org/publications/2009-Google_Scholar%27s_Ranking_Algorithm_--_An_Introductory_Overview_--_preprint.pdf