The Microsoft Academic Search Dataset and KDD Cup 2013

Senjuti Basu Roy
Institute of Technology
University of Washington
Tacoma, WA 98402, USA
senjutib@uw.edu

Martine De Cock
Dept. of Appl. Math., CS and
Statistics, Ghent University
9000 Gent, Belgium
martine.decock@ugent.be

Vani Mandava
Microsoft Research
Redmond, WA 98052, USA
vanim@microsoft.com

Swapna Savanna
Institute of Technology
University of Washington
Tacoma, WA 98402, USA
ssavvana@uw.edu

Brian Dalessandro
Media6Degrees
New York
NY 10003, USA
briand@m6d.com

Claudia Perlich
Media6Degrees
New York
NY 10003, USA
claudia@m6d.com

William Cukierski
Kaggle
Millington
NJ 07946, USA
will.cukierski@kaggle.com

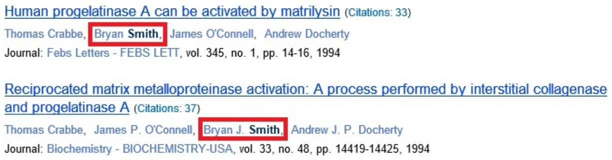
Ben Hamner
Kaggle
San Francisco
CA 94107, USA
ben.hamner@kaggle.com

ABSTRACT

KDD Cup 2013 challenged participants to tackle the problem of author name ambiguity in a digital library of scientific publications. The competition consisted of two tracks, which were based on large-scale datasets from a snapshot of Microsoft Academic Search, taken in January 2013 and including 250K authors and 2.5M papers. Participants were asked to determine which papers in an author profile are truly written by a given author (track 1), as well as to identify duplicate author profiles (track 2). Track 1 and track 2 were launched respectively on April 18 and April 20, 2013, with a common final submission deadline on June 12, 2013. For track 1 a training dataset with correct labels was diclosed at the start of the competition. This track was the most popular one, attracting submissions of 561 different teams. Track 2, which was formulated as an unsupervised learning task, received submissions from 241 participants. This paper presents details about the problem definitions, the datasets, the evaluation metrics and the results.

1. INTRODUCTION

The ability to search literature and collect/aggregate metrics around publications is a central tool for modern research. Both academic and industry researchers across hundreds of scientific disciplines, from astronomy to zoology, increasingly rely on search to understand what has been published and by whom. Microsoft Academic Search[1] is an open platform that provides a variety of metrics and experiences for the research community, in addition to literature search. It currently covers more than 50 million publications and over 19 million authors across a variety of domains, with updates added each week. One of the main challenges of providing this service is caused by author-name ambiguity. On one hand, there are many authors who publish under several variations of their own name, as illustrated in Figure 1. On the other hand, different authors might share a similar or even the same name. This makes correctly distinguishing between different authors and associating papers correctly with their true authors difficult tasks.



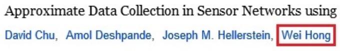
Human progelatinase A can be activated by matrilysin (Citations: 33)
Thomas Crabbe, Bryan **Smith**, James O'Connell, Andrew Docherty
Journal: Febs Letters - FEBS LETT, vol. 345, no. 1, pp. 14-16, 1994

Reciprocated matrix metalloproteinase activation: A process performed by interstitial collagenase and progelatinase A (Citations: 37)
Thomas Crabbe, James P. O'Connell, Bryan J. **Smith**, Andrew J. P. Docherty
Journal: Biochemistry - BIOCHEMISTRY-USA, vol. 33, no. 48, pp. 14419-14425, 1994

Figure 1: Example of author name variations: the names "Bryan J Smith" and "Bryan Smith" refer to the same person

These problems are further complicated by the fact that, depending on the source of the paper, the affiliation of the authors is often missing in the publications dataset of Microsoft Academic Search. Figure 2 shows an example of such a paper, co-authored by an author named "Wei Hong". As evident in Figure 3, there are several authors in the system with the name "Wei Hong" as it appears in the paper. The problem is in determining which of the "Wei Hong" authors in the authors dataset is the correct author of the paper.

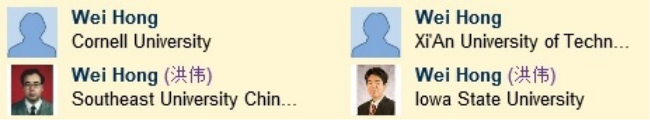
[1] `http://academic.research.microsoft.com/`



Approximate Data Collection in Sensor Networks using Probabilistic Models

David Chu, Amol Deshpande, Joseph M. Hellerstein, Wei Hong

Figure 2: Example of a paper that is co-authored by an author with an ambiguous name and missing affiliation



Wei Hong
Cornell University

Wei Hong
Xi'An University of Techn...

Wei Hong (洪伟)
Southeast University Chin...

Wei Hong (洪伟)
Iowa State University

Figure 3: Multiple authors sharing the name "Wei Hong"

Ambiguity in author names sometimes causes a paper to be assigned to the wrong author in Microsoft Academic Search, which leads to noisy author profiles. In addition, sometimes the system is too conservative, in the sense that it keeps two or even more separate author profiles for one and the same author. In this case, each of these separate profiles contains a subset of the papers of the author, and the system does not join the profiles because it does not believe that is has enough evidence to assume that they are from the same author.

In addition to automatically indexing millions of scientific publications, Microsoft Academic Search provides an interface in which researchers can manage their author profiles. Among other things, researchers can add and delete publications, and request author profiles to be merged. All these manual activities are logged and provide a rich ground truth dataset for the tasks of KDD Cup 2013. Concretely, KDD Cup 2013 challenged participants to determine which papers in an author profile are truly written by a given author (track 1) and to identify which author profiles in a given dataset should be merged because they represent the same author (track 2).

In Section 2 we describe how the ground truth data was used to create a training, a validation and a test set for track 1, and a validation and a test set for track 2. Contest participants were also given an additional large background dataset, containing information about 250K authors and 2.5M papers, that they could potentially leverage to design their solutions for the challenges of both track 1 and track 2 (see Section 2.3). In Section 3 we provide a description of the tasks and the evaluation metrics for both tracks. Contest conduct and results are presented in Section 4.

2. THE MICROSOFT ACADEMIC SEARCH DATASET

The data used in the competition is based on a snapshot of Microsoft Academic Search, taken in January 2013 and including 250K authors and 2.5M papers. The scores on the leaderboards are calculated based on ground truth data supplied by authors who manually corrected the information in their profile on the website of Microsoft Academic Search. Roughly speaking the competition's challenges consist in replicating the corrections requested by authors in terms of deleting wrongly assigned papers (track 1) and merging author profiles (track 2). Since by the start of the competition the manually suggested corrections had already been incorporated on the website of Microsoft Academic Search, crawling these updated author profiles would have been an easy and trivial way to obtain perfect solutions to the challenges of the competition. Participants were therefore told that they could not use any external data page for the purposes of the competition. The reason why no external data was allowed, and the way in which the ground truth data were obtained, were not communicated to participants until after the competition.

The ground truth data for track 1 is split in a training, a validation and a test dataset. The validation dataset was used for the scores on the public leaderboard throughout the competition, while the test dataset was used to determine the final ranking of the contestants. The correct labels for the training dataset were given at the start of the competition; those for the validation dataset were revealed two weeks before the end of the competition. Track 2 on the other hand is structured as an unsupervised learning problem, in the sense that no correct answers were provided while the competition was ongoing. However, for both tracks, a substantial background dataset was released to help participants design their solutions. Next, we describe each of these datasets in detail.

2.1 Ground Truth Data for Track 1

The ground truth data for track 1 contains 7479 unique author profiles and 424384 papers that are split in a training, a validation, and a test dataset. The ground truth is obtained from the user edits at the Academic Search website, where an assignment of a paper to an author is known to be incorrect if an author deleted the paper from the profile, or correct if an author confirmed it. Some authors who edit their profile do not explicitly confirm or delete every paper. We assume that these "untouched" papers in edited profiles correspond to correct assignments.

The training, the validation and the test dataset were created by random sampling from the ground truth dataset and contain respectively 3739, 1496 and 2244 authorIds. Every authorId comes with a set of assigned paperIds. On average, an author has been assigned 62 papers; however, some of these assignments are incorrect. On average, 33 papers are correct assignments and 29 papers are incorrect assignments. As an example, the distribution of the confirmed and the deleted papers in the training dataset are presented in Figure 4 and 5.

The training and validation datasets were provided at the start of the competition. However, only for the training dataset it was revealed at that time which papers are correct assignments and which ones are not. For the authors and papers in the validation dataset this information was kept hidden from contest participants and used to score their solutions on the public leaderboard using MAP (see Section 3). Two weeks before the end of the competition, the labels for the validation dataset were revealed as well, to give participants the opportunity to optionally retrain their model on the combined training validation sets. The test dataset, obviously without labels, was released one week before the end of the competition and used to score solutions to obtain

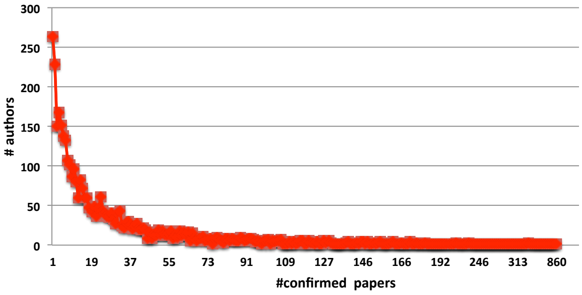


Figure 4: Distribution of confirmed papers in training set

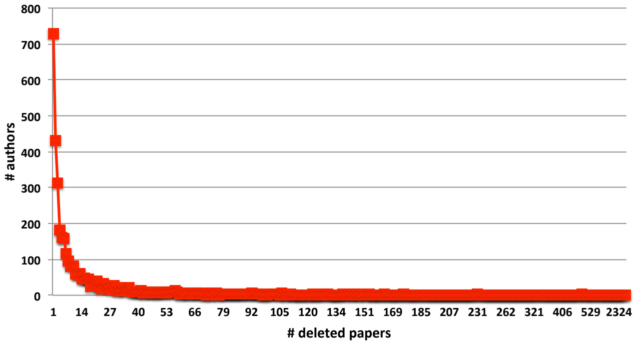


Figure 5: Distribution of deleted papers in training set

the final ranking and determine the winner.

2.2 Ground Truth Data for Track 2

Track 2 is structured as a "cold start" problem, meaning that there were no training labels provided. As in track 1, the ground truth for scoring of solutions for track 2 is obtained from user edits at the website of Microsoft Academic Search, where users can request to merge author profiles because they belong to the same author. The data for track 2 contains 201542 authorIds, out of which 20680 unique authors have requested to merge their profile with at least one other author profile in the dataset, leading to a total of 33648 merge requests. On average, the dataset contains 2 duplicate profiles per author.

A fixed, randomly selected 20% of the dataset was used to provide leaderboard feedback while the competition was running. The error on the remaining 80% was not shown to participants during the competition and was used to determine the final ranking. The ground truth was expected to have a number of false negatives, since a duplicated author who has not requested to merge their profile would not be labeled as a duplicate in the ground truth. This was an unavoidable source of noise that affected all participants equally.

2.3 Background Dataset

In addition to the data above, a background dataset was provided which the contestants could potentially leverage to design their solutions for the challenges of both track 1 and track 2. This raw dataset primarily describes the co-authorship network, where the associated metadata is presented using three different tables: an author table, a paper table, and a paper-author table. The schemas of these tables (datasets) are given in Figure 6.

The *Author dataset* contains profile information about 250K authors, such as author name and affiliation. The same author can appear more than once in this dataset, for instance because he/she publishes under different versions of his/her name. The *Paper dataset* contains data about 2.5M papers, such as paper title, conference/journal information, and keywords. The same paper may have been obtained through different data sources and hence have multiple copies in the dataset. A given paper could fall either in the conference or in the journal category. The third and final dataset is the *Paper-Author dataset* with (PaperId, AuthorId) pairs. The Paper-Author dataset is noisy, containing possibly incorrect paper-author assignments that are due to author-name ambiguity and variations of author names.

All the AuthorIds and PaperIds from the train, validation and test sets of track 1 and track 2 also appear in the background dataset.

Two forms of data leakage appeared in this data. In the Paper-Author dataset, some paper-author pairs appeared multiple times. When a paper-author pair appeared more than once, the author was disproportionally likely to have written the corresponding paper. Many competition participants identified this as their most important feature in Track 1. Additionally, in Track 1 training, validation, and test datasets were provided containing a subset of paper-author pairs where the author had explicitly confirmed or denied his authorship of the paper. There were a small number of duplicated pairs here as well, where the author was disproportionally likely to have written the corresponding paper.

3. TASK DESCRIPTION

3.1 Track 1

For track 1 of the competition, participants were given a set $\mathcal{A}$ of authors and, for each author $a \in \mathcal{A}$, a set P_a of papers that might or might not have been written by that author. The challenge was to split P_a up into two disjoint subsets, namely the set Y_a of papers authored by a, and the set N_a of papers not authored by a. The problem was somewhat complicated by the fact that the size of the set P_a was not the same for all authors, and neither was the proportion of the size of Y_a to that of N_a.

Concretely, participants were asked to rank the papers for each author, so that the "yes" instances (the papers from the set Y_a) come before the "no" instances (the papers from the set N_a). To evaluate the solutions we used *Mean Average Precision (MAP)*, a well known measure from information retrieval that factors in precision at all recall levels (see e.g. [1]). To give the definition of MAP in our setting, let R_a denote the ranked list given for author a, and, for a given paper $p \in Y_a$, let R_a^p denote the set of papers in R_a from the first paper until paper p is reached. Then

$$\text{MAP}(\mathcal{A}) = \frac{1}{|\mathcal{A}|} \sum_{a \in \mathcal{A}} \text{AP}(R_a)$$

Author dataset	Id	Id of the author
	Name	Author Name
	Affiliation	Organization name with which the author is affiliated
Paper dataset	Id	Id of the paper
	Title	Title of the paper
	Year	Year of publication of the paper
	ConferenceId	Id of the conference in which the paper is published
	JournalId	Id of the journal in which the paper is published
	Keywords	Keywords describing the paper
Paper-Author dataset	PaperId	Id of the paper
	AuthorId	Id of the author
	Name	Name of the author (as it appears on the paper)
	Affiliation	Author affiliation (as it appears on the paper)

Figure 6: Background dataset

with

$$\mathrm{AP}(R_a) = \frac{1}{|Y_a|} \sum_{p \in Y_a} \frac{|R_a^p \cap Y_a|}{|R_a^p|}$$

In other words, to determine the average precision AP(a) for author a, one calculates the precision at each point when an actual "yes" instance occurs in the ranking, and then takes the average of those values.

Example 1. *Let $P_a = \{p_1, p_2, p_3, p_4, p_5\}$ with $Y_a = \{p_3, p_5\}$, i.e. out of the 5 papers only p_3 and p_5 have been written by author a. The average precision of ranking*

$$R_a = (p_3, p_1, p_4, p_5, p_2)$$

is given by $\mathrm{AP}(R_a) = (1/1 + 2/4)/2 = 0.75$. Note that the relative ordering of the "no" instances in the ranking does not have an effect on the score. Neither does the relative ordering of the "yes" instances. For instance, the average precision of ranking

$$R'_a = (p_5, p_2, p_1, p_3, p_4)$$

is also $(1/1 + 2/4)/2 = 0.75$.

3.2 Track 2

The goal of track 2 was to identify which authors in the dataset are duplicates. To this end, for every author a in the dataset, participants were asked to provide the set of authors from the dataset that are, in reality, the same as author a. Every author counted as his/her own duplicate, and every duplicate had to be listed under each of its respective ids. For example, if a participant's system suspected that authors a, b, and c are the same, it should list $(a, \{a, b, c\}), (b, \{b, a, c\}), (c, \{c, a, b\})$.

The solutions for this task were evaluated using the *Mean F1 score.* The F1 score, commonly used in information retrieval (see e.g. [1]), measures accuracy using the statistics precision p and recall r. Precision is the ratio of the number of true positives (tp) to all predicted positives (tp + fp). Recall is the ratio of the number of true positives to all actual positives (tp + fn). The F1 score is given by:

$$\mathrm{F1} = \frac{2p \cdot r}{p + r}$$

where

$$p = \frac{tp}{tp + fp} \quad \text{and} \quad r = \frac{tp}{tp + fn}$$

Example 2. *Assume that the dataset contains only 7 authors $\{a, b, c, d, e, f, g\}$. In reality author a is the same as b and d, but the system believes that a is the same as b and c. I.e. the system predicts $(a, \{a, b, c\})$ while it should have predicted $(a, \{a, b, d\})$. In this case, the set of true positives is $\{a, b\}$, the set of false positives is $\{c\}$ and the set of false negatives is $\{d\}$. Hence $\mathrm{F1} = 2/3$.*

The F1 metric weights recall and precision equally, and a good retrieval algorithm will maximize both precision and recall simultaneously. Thus, moderately good performance on both will be favored over extremely good performance on one and poor performance on the other.

Since the majority of authors are not duplicates, the F1 score tends to be close to 1 for this task. It is therefore instructive to compare the F1 score of a solution to the benchmark representing the "null" prediction (each author is his own duplicate). Small differences in the absolute magnitude of the F1 score can therefore represent meaningful improvements in model performance, despite an inclination to assume lesser decimal places are insignificant.

4. CONTEST RESULTS

Track 1 and track 2 were respectively launched on Apr 18 and Apr 20, 2013. Both tracks had a final submission deadline on Jun 12. Figure 7 shows an overview of the number of daily submissions during the competitions. The number of allowed submissions per team was limited to 5 per day for track 1 and 2 per day for track 2. For track 1 a total of 9558 submissions were made by 561 different teams from 43 countries (determined by IP address). For track 2 there were 2309 submissions by 241 teams from 40 countries.

Figure 8 shows the best score on the Track 1 validation set over the course of the competition and Figure 9 shows the best score on the Track 2 test set over the course of the competition. The test set for Track 1 was only released at the end of the competition, so longitudinal performance isn't available for that set.

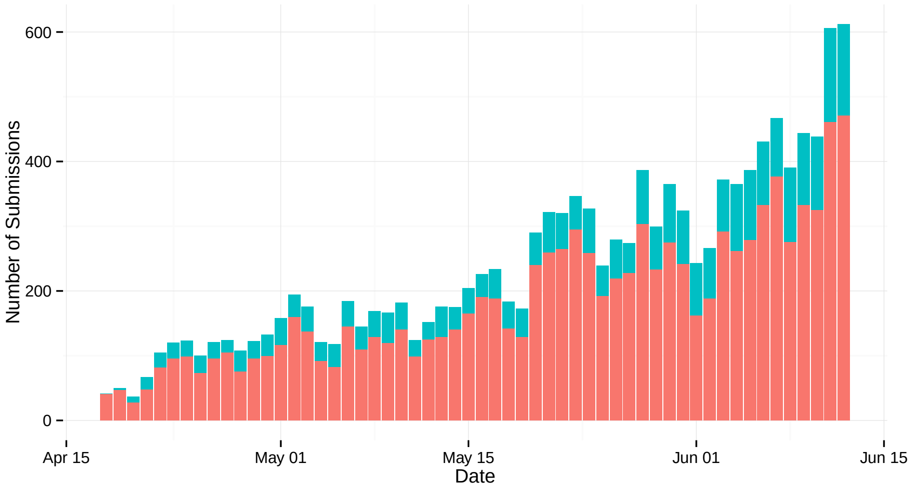


Figure 7: Number of submissions per day (track 1 in orange; track 2 in blue)

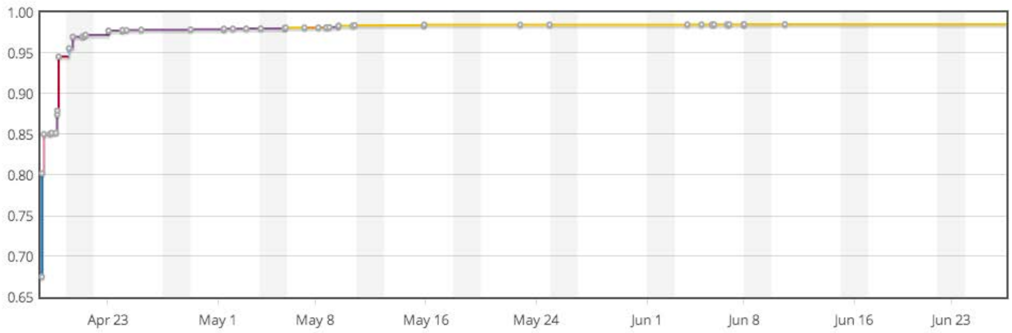


Figure 8: Track 1 best public score over competition duration (color change denotes a new leader, dots denote a new best score). A private score not available because track 1 had a second test set release.

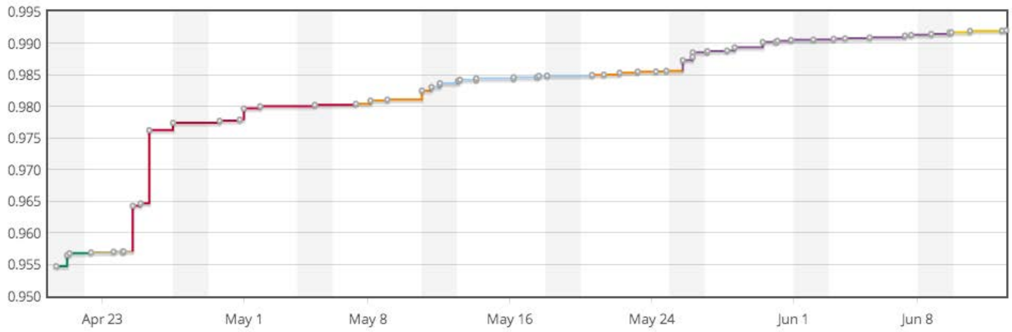


Figure 9: Track 2 best private score over competition duration (color change denotes a new leader, dots denote a new best score)

At the end of the competition, the best solution for Track 1 had a 88.9% performance improvement relative to the baseline, and the best solution for Track 2 had a 82.1% performance improvement relative to the baseline (calculated with regards to the maximum possible improvement).

5. REFERENCES

[1] P. R. Christopher D. Manning and H. Schütze. *Introduction to Information Retrieval*. Cambridge University Press, 2008.