The Microsoft Academic Search Challenges at KDD Cup 2013

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

Senjuti Basu Roy, Swapna Savvana
Institute of Technology
University of Washington
Tacoma, WA 98402, USA
Email: {senjutib,ssavvana}@uw.edu

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

Brian Dalessandro, Claudia Perlich
Media6Degrees
New York, NY 10003, USA
Email: {briand, claudia}@m6d.com

William Cukierski, Ben Hamner
Kaggle
Millington NJ 07946, USA
Email: {will.cukierski, ben.hamner}@kaggle.com

***Abstract*—Microsoft Academic Search is a free search engine specific to scholarly material. It currently covers more than 50 million publications and over 19 million authors across a variety of domains. One of the main challenges in correctly indexing this material is author name ambiguity and the resulting noise in author profiles. KDD Cup 2013 invited participants to tackle this problem in 2 ways: (1) by automatically determining which papers in an author profile are truly written by a given author, and (2) by identifying which author profiles need to be merged because they belong to the same author. This paper presents a brief account of the contest and the lessons learned.**

***Keywords*-Microsoft Academic Search; author name disambiguation**

I. Introduction

Microsoft Academic Search[1] is a free search engine specific to scholarly material. It is developed by Microsoft Research as a service to easily find information about academic content, authors and institutions. It currently covers more than 50 million publications and over 19 million authors. The platform allows search by keyword as well as by scholarly attributes such as author, conference, journal, organization, year and DOI. This makes it easy to find e.g. the highest impact papers published in ICDM a decade ago, using either the query "year=2003 conf:(icdm)" or the graphical user interface. It is also possible to search within scientific domains and subdomains, and to retrieve and compare information about specific organizations (see Figure 1 and 2). In addition, Microsoft Academic Search data is exposed via an API to allow non-commercial research entities to build compelling tools and experiences on top of the rich data.

An important prerequisite for a high quality search experience and trustworthy computation of publication metrics is the correct identification of authors and the assignment of papers to their right authors. Author name ambiguity poses a major challenge in this context. On one hand, there are many authors who publish under several variations of their own name, and on the other hand different authors might share a similar or even the same name. An example of a typical problem is deciding whether *Bryan J. Smith* who co-authored a paper on *Reciprocated matrix metalloproteinase activation* in 1994 is the same as *Bryan Smith* who co-authored a paper on *Human progelatinase A* in that same year. Another example is deciding whether a paper authored by *Wei Hong* was written by *Wei Hong* from *Cornell University* or by *Wei Hong* from *Iowa State University* or — possibly even worse — whether this is one and the same author who has a double affiliation or changed affiliation during his career. This problem is further complicated by the fact that, depending on the source of the paper, the affiliation data of the authors is often missing in the publications dataset.

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.

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). Both tracks 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. The solutions of participants were evaluated based on ground truth data supplied by authors who manually corrected the information in their profile on the website of Microsoft Academic Search. In Section II and III we respectively provide a description of the tasks and the datasets. More information can be found in [1]. Contest

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

978-1-4799-1293-3/13/$31.00 ©2013 IEEE

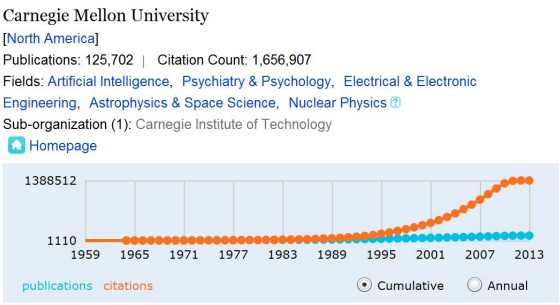

Figure 1. Use of Microsoft Academic Search to retrieve publication metrics of Carnegie Mellon University

conduct and results are presented in Section IV.

II. Problem Description

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. 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). The solutions were evaluated using *Mean Average Precision (MAP)*, a well known measure from information retrieval that factors in precision at all recall levels (see e.g. [4]). For example 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 $(p_3, p_1, p_4, p_5, p_2)$ is given by $(1/1 + 2/4)/2 = 0.75$. The average of these average precisions over all author profiles from the test dataset for track 1 was calculated and displayed as the score on the competition's final leaderboard.

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, measures accuracy using the statistics precision p and recall r (see e.g. [4]). For example, assume that the dataset contains only 7 authors[2] $\{a, b, c, d, e, f, g\}$. 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.

[2]The actual dataset used in the competition contained over 200000 authors; see Section III.

III. Datasets

The data used in the competition was 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 were 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 consisted in replicating the corrections requested by authors in terms of deleting wrongly assigned papers (track 1) and merging author profiles (track 2).

W.r.t. the ground truth data for track 1, as assignment of a paper to an author is known to be correct if an author confirmed it, and incorrect if an author deleted the paper from the profile. Some authors who edited their profile did not explicitly confirm or delete every paper. These "untouched" papers in edited profiles were assumed to correspond to correct assignments. In this way ground truth data for 7479 unique author profiles was collected, involving a total of 424384 papers. These author profiles were were split in a training, a validation, and a test dataset by random sampling from the ground truth dataset. They 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.

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 II). 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 the final ranking and determine the winner.

Track 2 was 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 was obtained from user edits at the website of Microsoft

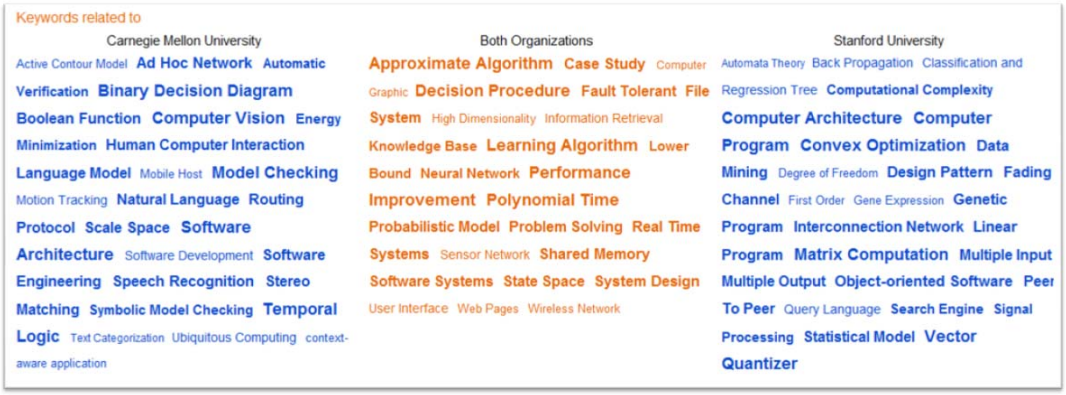

Figure 2. Use of Microsoft Academic Search to compare publication keywords of Carnegie Mellon University and Stanford University

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.

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 with information about 250K authors, a paper table with data about 2.5M papers, and a paper-author table 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.

More details can be found in [1] and the data is available for download at the competition's webpages for the 1st track[3] and the 2nd track[4].

[3]http://www.kaggle.com/c/kdd-cup-2013-author-paper-identification-challenge

[4]http://www.kaggle.com/c/kdd-cup-2013-author-disambiguation

IV. 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. 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. Track 2 was in a sense more difficult than track 1 because no training data was provided. This might explain the higher popularity of track 1, although it could be attributed as well to the fact that track 1 was launched a few days earlier and thus attracted more attention.

Figure 3 shows the best score on the Track 1 validation set over the course of the competition and Figure 4 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. 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).

Detailed descriptions of the strongest approaches are available at the website[5] of KDD Cup 2013. The key to success for track 1 was in extensive feature engineering combined with the use of known state of the art supervised learning algorithms, while for track 2 participants resorted more to custom made algorithms designed for the problem at hand. The winning solution for Track 1 [3] was designed with exhaustive feature engineering (97 features) extracted from the provided dataset, followed by employing an ensemble of supervised learning algorithms (i.e., classifiers), as the ranking task was designed as a binary classification problem.

[5]http://www.kdd.org/kddcup2013

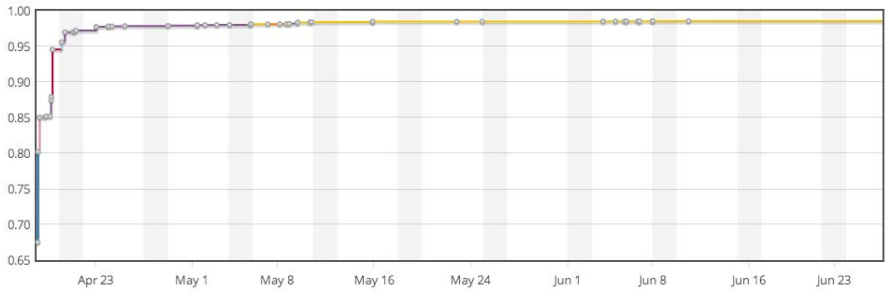

Figure 3. Track 1 best public score over competition duration (color change denotes a new leader, dots denote a new best score). A private score not available because track 1 had a second test set release.

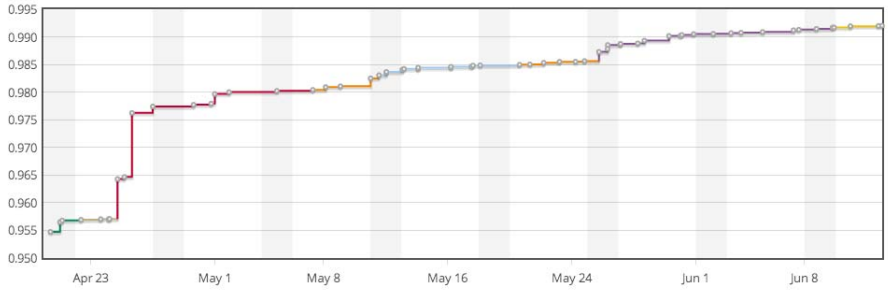

Figure 4. Track 2 best private score over competition duration (color change denotes a new leader, dots denote a new best score)

Gradient Boosting Trees and Decision Trees are used in the ensemble design. After training the ensemble of classifiers, the validation dataset is used to calculate MAP. To boost the MAP score even further, the team had designed two post-processing heuristics to further improve the ranking.

On the other hand, the winning solution for Track 2 [2] was proposed by designing an effective name matching framework, and considering two different implementations inside. The team was divided into two sub-groups and two different implementations were developed. The overall framework was designed in six-steps: 1) identifying a name Chinese/Non-Chinese, 2) cleaning, 3) selection, 4) identification, 5) splitting, and 6) linking.The authors underscored that by handling Chinese and non-Chinese names separately, an improvement in the quality (F1 Score) of the overall solution was observed. Post-processing strategies were designed to further improve the F1 Score.

The popularity of this year's KDD Cup, with its focus on one of the world's largest research databases, illustrates that there is a wide interest in the research community for solving fundamental problems underlying scholarly big data applications. The quality of the top 10 solutions designed by the contest's participants was very high in terms of MAP and F1-score. Closer inspection however revealed that none of these approaches directly scales sufficiently well for use on the entire Microsoft Academic Search author and publication data, out of which the datasets provided for the contest were only a relatively small sample (although already perceived as big by participants!). The development of author name disambiguation approaches that achieve a good trade-off between accuracy and scalability continues to be an interesting research problem.

REFERENCES

[1] S. Basu Roy, M. De Cock, V. Mandava, S. Savvana, B. Dalessandro, C. Perlich, W. Cukierski, B. Hamner, *The Microsoft Academic Search Dataset and KDD Cup 2013*, in: Proceedings of KDD Cup 2013 Workshop at KDD2013.

[2] W.-S. Chin, Y.-C. Juan, Y. Zhuang, F. Wu, H.-Y. Tung, T. Yu, J.-P. Wang, C.-X. Chang, C.-P. Yang, W.-C. Chang, K.-H. Huang, T.-M. Kuo, S.-W. Lin, Y.-S. Lin, Y.-C. Lu, Y.-C. Su, C.-K. Wei, T.-C. Yin, C.-L. Li, T.-W. Lin, C.-H. Tsai, S.-D. Lin, H.-T. Lin, C.-J. Lin. *Effective String Processing and Matching for Author Disambiguation*, in: Proceedings of KDD Cup 2013 Workshop at KDD2013.

[3] C.-L. Li, Y.-C. Su, T.-W. Lin, C.-H. Tsai, W.-C. Chang, K.-H. Huang, T.-M. Kuo, S.-W. Lin, Y.-S. Lin, Y.-C. Lu, C.-P. Yang, C.-X. Chang, W.-S. Chin, Y.-C. Juan, H.-Y. Tung, J.-P. Wang, C.-K. Wei, F. Wu, T.-C. Yin, T. Yu, Y. Zhuang, S.-D. Lin, H.-T. Lin, C.-J. Lin, *Combination of Feature Engineering and Ranking Models for Paper-Author Identication in KDD Cup 2013*, in: Proceedings of KDD Cup 2013 Workshop at KDD2013.

[4] C. D. Manning, P. Raghavan, H. Schütze, *Introduction to Information Retrieval*, Cambridge University Press, 2008.