Predicting the future relevance of research institutions

Vlad Sandulescu

Mihai Chiru

2016.08.14

Phase 1

Start simple

#1: **build a dataset**

#2: explore the dataset

Dataset

- only used MAG
- for all the authors of the full research papers between 2011-2015, get all their papers starting with 2000
- for all these papers, get all related info (references, keywords, fields of study, etc.)

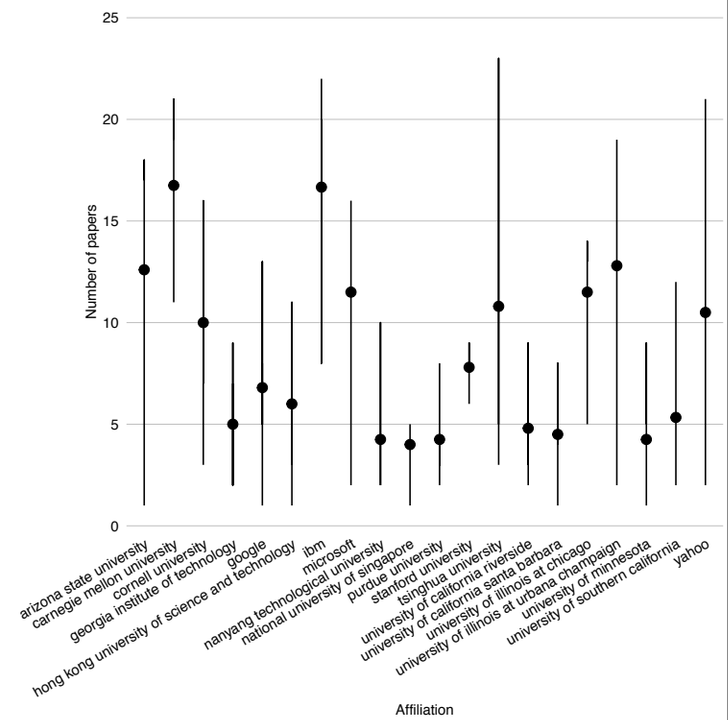


Figure: Full range and mean value of the number of accepted full research papers for top 20 affiliations at KDD between 2011 and 2015

Dataset: read slide

Phase 1

Start simple

#1: build a dataset

#2: **explore the dataset**

Dataset

- only used MAG
- for all the authors of the full research papers between 2011-2015, get all their papers starting with 2000
- for all these papers, get all related info (references, keywords, fields of study, etc.)

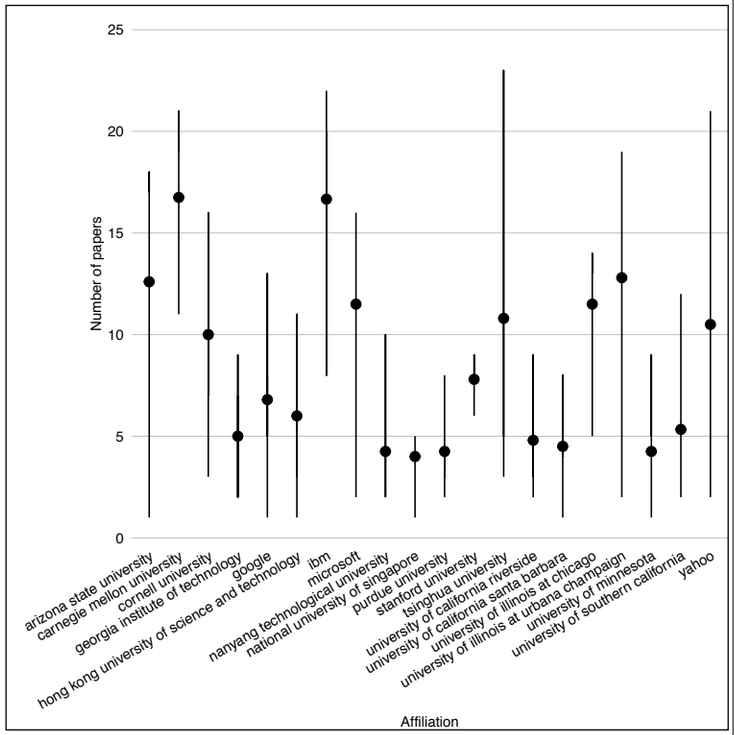


Figure: Full range and mean value of the number of accepted full research papers for top 20 affiliations at KDD between 2011 and 2015

One of the first analysis we carry out is to check for any obvious trends for the accepted papers made by top affiliations to each conference. We consider a top affiliation one which had a large number of accepted papers at a conference in the last five years. The assumption is that for large conferences at least, the top 20 places each year will be taken by more prolific affiliations, likely to have participated in the past to the conference. It is unlikely that affiliations which have few sporadic research papers accepted in the last years are going to be present in the top 20 places. The plot shows the number of full research papers accepted at the KDD conference between 2011 and 2015 for the top 20 affiliations. The length of each line maps the range of the number of accepted papers for the affiliation and the mean number of papers is marked by the larger dot on each line. The plot shows the mean number of papers across all years could be a good predictor to how an affiliation will score in the future.

Phase 1

Start simple

#1: build a dataset

#2: explore the dataset

#3: set up a baseline

Baseline model:
Count the number of full research papers an affiliation published at each conference in 2011-2015
Compute the probability of a full research paper is published by an affiliation

Conference	2015	2014	2013
SIGIR	0.95	0.94	0.89
SIGMOD	0.87	0.94	0.82
SIGCOMM	0.93	0.95	0.77

Table: NDCG@20 results for the probabilities model in phase 1 for 2013, 2014 and 2015

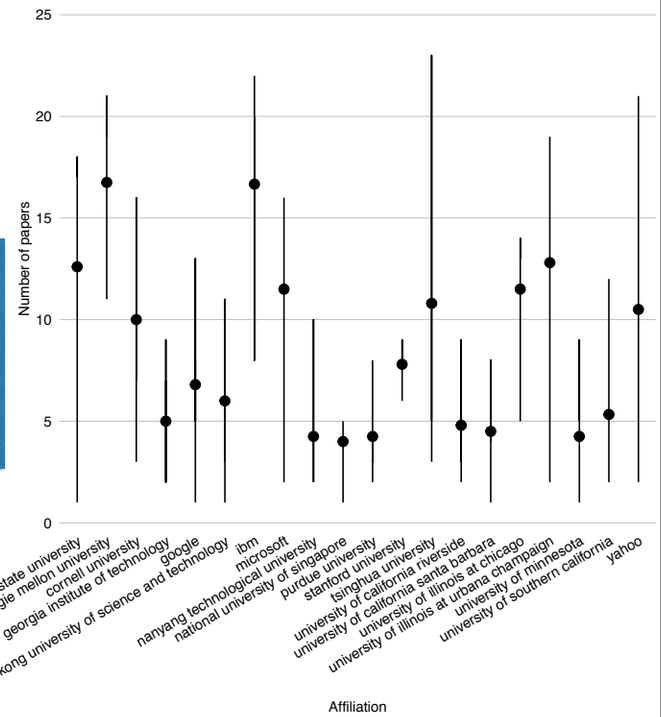


Figure: Full range and mean value of the number of accepted full research papers for top 20 affiliations at KDD between 2011 and 2015

In this first phase, we compute the probability of a full research paper is published by a given affiliation, based on the number of all their accepted full research papers between 2011-2015. The affiliations are ranked according to these probabilities and this represents our submission for the first phase of the competition. You can see in the table that predictions made using probabilities computed on more years (e.g. compute the probabilities over the entire range 2011-2014) are generally more accurate. They also follow an ascending trend for SIGIR when predicting the next year (e.g. predict 2015 rankings), than using 2011-2013 to compute the probabilities and predicting 2014 rankings. The predictions for 2013 using only the past couple of years are the worst across all the conferences.

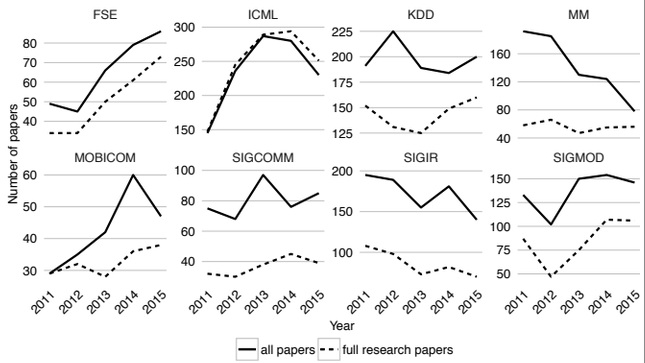
Phase 2

Start improving the baseline

#1: **predict relevance directly**

#2: explore the dataset even more

#3: try GBDT & Mixed models



- The evaluation metric directly uses the relevance

Figure: Number of all the papers vs the full research papers for all the conferences in the competition for 2011-2015

Our initial models use the number of accepted papers per year per affiliation as main predictor, but this predictor misses the fractional contributions of authors to the final affiliation rank. The relevance score, as calculated in the competition rules, distributes the paper score to each author and thus, each related affiliation receives only a fraction of the score. Starting with the second phase of the competition, we decide to use the relevance score as the prediction target in our models. So we train our models to predict the relevance directly.

Phase 2

Start improving the baseline

#1: predict relevance directly

#2: **explore the dataset even more**

#3: **try GBDT & Mixed models**

- The evaluation metric directly uses the relevance

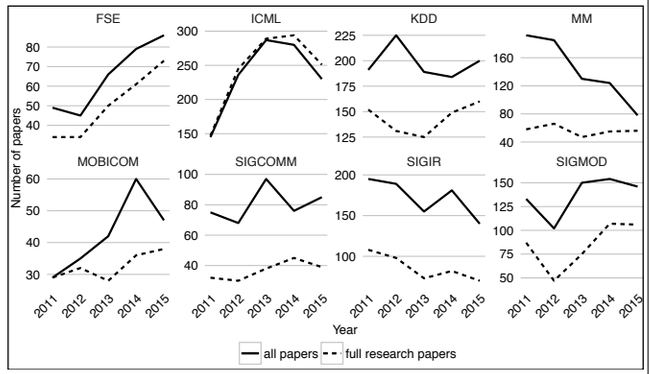


Figure: Number of all the papers vs the full research papers for all the conferences in the competition for 2011-2015

We check if the number of full research papers per conference per year is correlated with the total number of papers at the conference for that same year. Besides research papers, there are usually other types of contributions also present in the MAG such as journal papers, workshop papers or posters. We plot the number of papers on the Y axis across all the years on the X axis. The solid line represents the number of all the papers, while the dotted line is the number of full research papers only. The figure shows the two curves look similar for most the conferences and there is correlation between the number of full research papers and all the papers accepted at each conference. We aim to use two types of models, GBDT and mixed models. The first is more predictive, while the latter is more interpretable.

Phase 2

Start improving the baseline

#1: predict relevance directly

#2: explore the dataset even more

#3: try GBDT & Mixed models

#4: **expand the dataset**

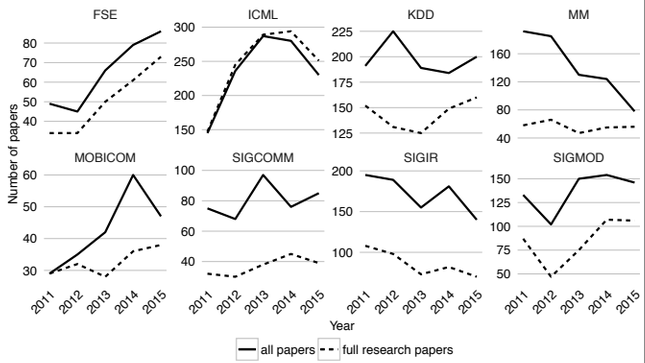


Figure: Number of all the papers vs the full research papers for all the conferences in the competition for 2011-2015

- The evaluation metric directly uses the relevance

	# samples
Full research papers	3,677
Phase1: probabilities	1,296
Phase 2: full research papers	**8,605**
Phase 2: all papers	**10,900**

Table: Dataset evolution between phases

Conference x Affiliation x Year x [features]

We use this correlation to extend our dataset and consider all the papers for our predictions. We build the dataset as combinations of (conferences x affiliations x years). The large number of combinations helps grow the dataset from several hundreds samples to several thousands.

Phase 2

Start improving the baseline

#1: predict relevance directly

#2: explore the dataset even more

#3: try GBDT & Mixed models

#4: expand the dataset

#5: engineer features

- The evaluation metric directly uses the relevance

	# samples
Full research papers	3,677
Phase1: probabilities	1,296
Phase 2: full research papers	**8,605**
Phase 2: all papers	**10,900**

Table: Dataset evolution between phases

Features

- Stats of all previous relevance scores (std, sum, mean, median, min, max)
- Previous relevance scores in windows from previous year up to 4 years ago
- Weighted moving-average of previous relevance scores in windows from previous year up to 4 years ago
- Stats of AIF metrics (std, sum, mean, median, min, max)

GBDT model:
Using all papers consistently improved our predictions across all conferences
Predict the relevance of each affiliation in 2016 using all the papers from 2011-2015

We already know the mean historical relevance has a strong impact on how relevant an affiliation will be at a particular conference in the next year. Besides this, other features such as weighted trends, which are derived from past relevance values, as well as authors publishing trends also have an impact on the future relevance. First we compute simple statistics to measure different properties of past years' relevance scores. Then we try to capture the relevance trend across past years. Finally, another set of features is built using the author impact factor (AIF) measure. For a given author, the AIF is computed as the number of citations the author receives in a year for his publications from previous 5 years. Our final submission for the second phase are the prediction made by a single GBDT model using all the papers from 2011-2015 and the mentioned features.

Phase 3

Improve the model further

#1: **find related conferences**

- Authors submit papers to similar conferences
- Jaccard similarity using authors & keywords
- sim = (#common authors) / (#all authors)

By authors	By keywords
ICDM	CIKM
CIKM	ICDM
WWW	WWW
AAAI	SIGIR
ICML	SIGMOD
SDM	ICML
PAKDD	AAAI
ICDE	NIPS

Table: Conferences related to KDD

Most researchers publish their work at different conferences. However they specialize in a specific area and so the conferences they publish at have to be similar at least in a few respects. We use authors and keywords from the papers in MAG to cluster similar conferences together. It is a straightforward way to grow the dataset even more. The assumption is the information from related conferences will enforce the patterns discovered by the models, because prolific affiliations are prolific across all conferences they submit to, not just at one of them. We compute the Jaccard similarity for both authors and keywords for any pair of conferences in the MAG. From this, we can determine which conferences are for example most similar to KDD in terms of common authors and common papers' keywords. The table shows the most related conferences to KDD.

Phase 3

Improve the model further

#1: find related conferences

#2: expand the dataset even more

- Expand the dataset with papers starting with the year 2000

	# samples
Full research papers	3,677
Phase1: probabilities	1,296
Phase 2: full research papers	8,605
Phase 2: all papers	10,900
Phase 3: FSE + 5 related conferences	**25,136**
Phase 3: MOBICOM + 5 related conferences	**21,872**
Phase 3: MM + 10 related conferences	**92,762**

Table: Dataset evolution between phases

As we shown previously, the number of full research papers is correlated with the total number of papers an affiliation has at a conference. Although the full research papers are not explicitly marked in the MAG before 2011, we assume this also applies to papers before this year. So we extend our dataset using papers from year 2000 and onward. The number of observations we use to train our models in this final phase reaches for some conferences almost 100K.

Phase 3

Improve the model further

#1: find related conferences

#2: expand the dataset even more

#3: refine the engineered features

Features

- Stats of all previous relevance scores (std, sum, mean, median, min, max)
- Previous relevance scores computed in windows from previous year up to 4 years ago
- Stats of previous relevance scores (std, sum, mean, median, min, max) computed in **windows** from previous year up to 4 years ago

In this phase, we keep most of the features we used in the previous competition phase, but we try to refine them. Besides the simple statistics which measure an affiliation's past relevance at a particular conference across all the years, we create four **window** versions of each of them for the last year up to four years back.

Phase 3

Improve the model further

#1: find related conferences

#2: expand the dataset even more

#3: refine the engineered features

Features

- Stats of all previous relevance scores (std, sum, mean, median, min, max)
- Previous relevance scores computed in windows from previous year up to 4 years ago
- Stats of previous relevance scores (std, sum, mean, median, min, max) computed in **windows** from previous year up to 4 years ago

---

- Drift trend of previous relevance scores
- Exponential weighted moving average of previous relevance scores with estimated smoothing parameter
- Exponential weighted moving average of previous relevance scores, computed with a fixed smoothing parameter

We replace the features based on weighted trends from the previous phase with a set of more accurate time series based trend measures. We use the drift trend of historical relevance scores, which captures the increase or decrease of the relevance over time according to the average change in the past samples. This gives us an estimate of the relevance for the current year which we wrap in a new feature. Also we create five more features based on simple exponential smoothing and we experiment with different smoothing parameters. The new features are one-step-ahead forecasts, based on **all** past relevance scores. Depending on the values of the smoothing parameter, we give exponentially less weight to older observations, making newer observations more important and vice versa. Yet another new feature finds the smoothing parameter value to best fit the data points.

Phase 3

Improve the model further

#1: find related conferences

#2: expand the dataset even more

#3: refine the engineered features

#4: tune the models

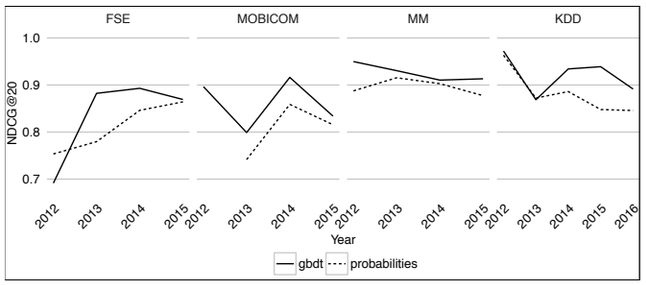


Figure: Results for the best features configuration

GBDT model:
Choose the model configuration (feature sets + #related conferences) which outperforms the baseline across all the years

Features

- Stats of all previous relevance scores (std, sum, mean, median, min, max)
- Previous relevance scores computed in windows from previous year up to 4 years ago
- Stats of previous relevance scores (std, sum, mean, median, min, max) computed in **windows** from previous year up to 4 years ago

---

- Drift trend of previous relevance scores
- Exponential weighted moving average of previous relevance scores with estimated smoothing parameter
- Exponential weighted moving average of previous relevance scores, computed with a fixed smoothing parameter

Finally we tune our models configurations by combining all the features in feature sets together with different numbers of related conferences. We test our models against the baseline for each configuration. From all the configurations we choose for each of the conferences the one which outperforms the baseline across all years. You can see in the plot these results: the Y axis shows the NDCG@20 score obtained by the GBDT model (solid line) versus the baseline (dotted line) across all years marked on the X axis.

Key points

- Used only the publicly available Microsoft Academic Graph data
- Started simple and gradually improved our models
- Grew our dataset in each phase and this improved the predictability
- Tuned the models so they outperformed an already known good baseline

- **Won the cup!**

THANK YOU

questions?