Modelling the Author-Paper Authentication Task

Yingzhen Li

May 11, 2013

We need to get the features of both the author and the paper, then computes the result with some model.
This model needs modification, just edit the tex code and re-sync it.

1 Features of the Author

1.1 Overview of the Author Table

Attributes: author id, name, affiliation
#author with distinct id: 247203
Some authors may have different id (I don't believe that there exist two or more people with the same name in some department!).

1.1.1 Name

May be abbreviated, e.g. A. Agliolo Gallitto, or, A. A. M. Negm.
Also it can be null!

1.1.2 Affiliation

Over 150,000 records have the affiliation with " value.
Can be university, department, company or some other organizations.

1.2 Features Related to the Author

Note: the authors in the validation set has no paper confirmed or deleted, hence we suggest here the PaperAuthor table is 'authentic'.

1.2.1 Topics

Find the paper of some author in PaperAuthor table then figure out the topics of their keywords and titles.
Get the research interests from the co-authors.

1.2.2 Journal & Conference

#Journal or #Conference the author has published paper in.
The topic of some journal (conference).

1.2.3 Co-author

Co-author list: figure out the co-authors of some paper in the PaperAuthor table.
Assign 'weights', e.g. by counting the number of papers co-written.
(Optional) Specify the topics of different co-authors.
(Optional) potential co-author

1.2.4 Other Statistics

#paper the author has published #paper the author has published in some year (high computation!)

2 Features of the Paper

2.1 Overview of the Paper Table

Attributes: paper id, title, year, conference id, journal id, keyword
#paper with distinct id: 2257249
#paper with title = '': 161935
#paperid with the same title (except ''): 1.0575087515292465

2.1.1 Keywords

The cut of the keywords: unigram, bi-gram, tri-gram ...
I suggest performing the keyword analysis in the paper sets of the same journal/conference:
1. for common keywords: still common!
2. for terminology: may have distinct meaning regarding to different research areas.
Analysis of the title:

2.1.2 Topics

We assign the result of the topic models to compose the topic set. See Section 3.1.1.

2.1.3 Journal and Conference

The topic sets of the paper and its publishing journal/conference may be intersected.
If the paper is with no keywords, topics of the journal/conference can help identifying the research area.
Statistics: #paper the author published in some journal/conference.

3 Journal and Conference

3.1 Overview of the Journal/Conference Table

Attributes: journal id (conference id), short name, full name, homepage
#journal with distinct short name/full name: 8284/15107
#conference with distinct short name/full name: 3692/4480

3.1.1 Topics

Extracted from the keywords (and titles) of its papers.
Topic models: a document contains the keywords of some journal/conference.
Using LDA (shallow topic extraction) or HDP/HLDA (hierarchical). Here note that we should see the perplexity to determine the number of topics assigned in advance.
Overcoming the pre-assign of #topic: Bayesian Nonparametrics, e.g. the Chinese Restaurant Franchise model.

4 Affiliation (Implied)

Containing authors: they may have similar research interests.
Hierarchical: e.g. University of Cambridge - Department of Engineering.

5 Topics (Implied)

Can be viewed as a kind of object in graphic models.
Can be viewed as a feature vector.

6 Connections between Objects

We only discuss some paths with lengths no longer than 3 here.

6.1 Direct Connections

Length 1:
author - (confirmed/deleted) paper (in (TrainConfirmed/TrainDeleted) PaperAuthor)
author - affiliation (in Author)
paper - conference/journal (in Paper)

6.2 Indirect Connections

Length 2:
author - paper - author (co-author)
author - paper - journal/conference (journal/conference of the author's preference)
paper - author - paper (maybe these papers described very similar topics)
Length 3:
author - paper - author - paper (maybe another collaboration)
author - paper - journal/conference - paper (maybe the work co-related)
paper - author - affiliation - author (potential co-author, weak implication)

6.3 Indirect Connections Adding Topics

Length 2:
author - paper - topic (get the research interests)
author - affiliation - topic (weak implication! we may not use it)
author - topic - paper (whether the author wrote the paper or not)
author - topic - author (potential co-author)
author - topic - journal/conference (journal/conference the author may publish paper in)

paper - journal - topic (if missing the keywords)
Length 3:
author - paper - author - topic (the co-author's topics may be the author's topics)
author - paper - journal/conference - topics
paper - author - paper - affiliation (may not exist usually, but some researches of an author can be highly co-related)
author - topic - journal/conference - paper
...

6.4 Path Weight Decadence

Linear: $weight = 1 - \alpha(length - 1)$
Non-linear:

7 Prediction with Graphic Models

For author A and paper P we compute the grading which indicate $P(A\ wrote\ P)$:

7.1 Just Find the Paths

1. figure out all the paths from A to P
2. $P(A\ wrote\ P) \propto \#path$

7.2 Shortest Path

1. assigning weights of each edge (path of length 1)
2. figure out the shortest path

7.3 Additive Path

1. figure out all the path and assigning weights (decade regards to the length, or just add up the weights)
2. sum up the weights of the paths

7.4 Probabilistic Models

1. first-order Markov chain: $P(path) \propto \prod_i P(object_{i+1}|object_i)$
2. (k-order) Markov model: $P(path) \propto \prod_i P(object_{i+1}|object_{i-k}, ..., object_i)$
3. $P(A\ wrote\ P) = 1 - \prod(1 - P(path))$