Name Disambiguation - aiming at disambiguating WhoIsWho

Zhang Xvdong

December 16, 2019

Abstract

In many applications, the eponymous disambiguation(Name Disambiguation - aiming at disambiguating WhoIsWho) has been regarded as a challenging issue, such as scientific document management, character search, social network analysis, etc., and with the proliferation of scientific literature, the resolution of the problem has become more difficult and urgent. Although the eponymous disambiguation has been extensively studied in academia and industry, the problem has not been solved well due to the clutter of data and the complexity of the eponymous scenario.

1 Introduction

Online academic search systems (such as Google Scholar, Dblp and AMiner, etc.) that include a variety of papers, have become the most important and popular academic exchange and paper search platform in the global academic community. However, due to the limitation of the paper distribution algorithm, there are a large number of papers distribution errors within the existing academic system, in addition, a large number of new papers will enter the system every day. Therefore, how to accurately and quickly assign papers to the existing author files in the system and maintain the consistency of author files is an urgent problem to be solved in the existing online academic system. Because of the huge amount of data within the academic system (AMiner has a file of approximately 130,000,000 authors and more than 200,000,000 papers), the author's eponymous scenario is complex, and there are still significant obstacles to resolving the problem of the eponymous disambigfoot quickly and accurately. The competition hopes to propose a problem-solving model that can distinguish papers of the same name belonging to different authors based on the details of the paper and the link between the author and the paper, and obtain good results from the paper. The good results of demystifying results can also affect other related fields by ensuring the effectiveness of expert knowledge search, high-quality

content management of digital libraries and personalized academic services in academic systems.

2 Background

Data mining is to obtain useful regular information according to the algorithm model from a large amount of data for subsequent work and decision-making. So the general data mining steps are: data acquisition, data preprocessing, data modeling.

Question Category First determine whether the problem is a classification problem or a regression problem. The incremental disambiguation of the paper is a classification problem.

Feature Selection Methods ***Embedded selection***: combining feature selection with a learner to allow automatic feature selection during model training, such as various tree models; ***Wrapped selection***: using the model's final learning performance as a feature set Evaluation criteria, choose a good feature set, but because it needs to train multiple models, it is expensive; ***Filter selection***: calculate the correlation between each feature and the corresponding variable, filter out the less relevant features, but the actual In the application, the relationship with the predictor is generally analyzed on a feature-by-feature basis. In Pandas, there are many related functions (describe, value_counts (), etc.) that can clearly show the relationship between the two. ***Dimension reduction***: Use related algorithms to process the data set, rank the importance of features, and take the features with high importance, such as PCA.

Model Establishment The establishment of a model refers to the various algorithms used in order to mine useful information, whether it is a traditional machine learning algorithm or a popular deep learning algorithm in recent years. According to different learning methods, machine learning algorithms can be divided into: supervised learning, unsupervised learning, semi-supervised learning, and reinforcement learning. Facing different data situations, different algorithms such as classification, regression, clustering, and association analysis can be used. At present, no single model can get good results, and all adopt the concept of ensemble learning to fuse models.

3 Technique

Feature generation is particularly critical in data mining. In the disambiguation of the same name in the paper, TF-IDF algorithm is used to extract features. TF-IDF is a statistical method to evaluate the importance of a

word to an article or an article in a corpus. The importance of a word increases proportionally with the number of times it appears in the file, but at the same time decreases inversely with the frequency of its appearance in the corpus. The main idea of TF-IDF is that if a word or phrase appears frequently in an article with a high TF and rarely appears in other articles, it is considered that the word or phrase has a good class discrimination ability and is suitable for To classify. TF-IDF is actually $TF * IDF$, where TF (Term Frequency) indicates how often a term appears in the document Document; IDF (Inverse Document Frequency), the main idea is that if there are fewer documents containing a word Word , The more differentiated the word is, the greater the IDF. For how to get the keywords of an article, we can calculate the TF-IDF of all nouns appearing in this article. The larger the TF-IDF, the higher the discrimination of this noun to this article. Take TF-IDF Larger words can be used as keywords in this article.

4 Solution

After previous analysis, the detailed steps of disambiguation of the same name of the paper are as follows:

Generate Training Data First extract the author's paper information in the training set and the meta-information on the training set from train _author.json and train_pub.json. Then filter the training set, and only take names with 5 or more authors of the same name as the training set. 500 training examples, one training example contains the paper, the real author of the paper, and 5 negative authors, the ratio of positive and negative examples is 1: 5.

Generate Feature Feature extraction is particularly critical in data mining. Good features can better reflect the core properties of the extracted content. Here, we use the TF-IDF algorithm to extract features from the title, abstract, and keywords of the literature. After data cleaning, we generate all positive and negative features.

Training With SVM After constructing positive and negative SVM examples, use SVC for training.

Load and Process Test Fata Similar to generating training data, first extract the author's paper information in the training set and the meta-information on the training set from whole_author_profile.json and whole _author_profile_pub.json, and load the set of papers to be assigned; then use the same feature extraction method Extract features; finally generate a list of candidate authors.

Use Trained Models to Predict Results Use the trained SVM model to score all candidate authors for each paper to be assigned, and use the highest score as the prediction result.

5 Evaluation

The final score is 63 points.

Reference

[1]. Jie Tang, Jing Zhang, Limin Yao, Juanzi Li, Li Zhang, and Zhong Su. ArnetMiner: Extraction and Mining of Academic Social Networks. In Proceedings of the Fourteenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (SIGKDD'2008). pp.990-998.

[2]. Yutao Zhang, Fanjin Zhang, Peiran Yao, and Jie Tang. Name Disambiguation in AMiner: Clustering, Maintenance, and Human in the Loop. In Proceedings of the Twenty-Forth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'18).

[3]. Jie Tang, A.C.M. Fong, Bo Wang, and Jing Zhang. A Unified Probabilistic Framework for Name Disambiguation in Digital Library. IEEE Transaction on Knowledge and Data Engineering (TKDE), 2012, Volume 24, Issue 6, Pages 975-987.

[4]. Xuezhi Wang, Jie Tang, Hong Cheng, and Philip S. Yu. ADANA: Active Name Disambiguation. In Proceedings of 2011 IEEE International Conference on Data Mining (ICDM'11), pages 794-803.

[5]. https://biendata.com/competition/scholar2018/data/

[6]. The Microsoft Academic Search Dataset and KDD Cup 2013

[7]. Wang, F. , Li, J. , Tang, J. , Zhang, J. , & Wang, K. . (2008). Name Disambiguation Using Atomic Clusters. Web-Age Information Management, 2008. WAIM '08. The Ninth International Conference on.