ATLANTIS PRESS

Advances in Intelligent Systems Research, volume 133

2nd International Conference on Artificial Intelligence and Industrial Engineering (AIIE2016)

A General Gender Inference Method Based on Web

Hong Yang[1,*] and Yali Yuan[2]

[1]Department of Computer Science and Technology, Tsinghua University, China

[2] Institute of Computer Science, Goettingen, Germany

[*]Corresponding author

***Abstract*—Gender information, as a crucial part of human demographics, is valuable for its abundant connotations and potential applications. Though much effort has been made on the problem of gender inference, most existing methods are highly dependent on data from specific sources, like Twitter, and are difficult to be generalized to other tasks. In this work, we propose a general Web-based method for gender inference. We show that our model significantly outperforms state-of-the-art without much human workload or any limits on specific scenarios. Based on that, we also present a voting framework to efficiently incorporate several methods to further improve performance. Experiments show that our voting framework can achieve 96.9% accuracy.**

Keywords-component; data mining; gender prediction; demographic; big data

I. INTRODUCTION

Human demographic is drawing a raising attention for its great value in a bunch of applications like marketing decisions and human resource management. Gender information, as a fundamental part of all demographic attributes, is even more interesting for its extended connotations leading to some undersurface insights. For instance, a report by [1] showed that women were underrepresented in STEM fields in education and employment. [2] and [3] offered a deeper understanding of such gender disparities, including their causes and influences. All these works are based on a convincing gender database, which is quite hard to gather and label by human experts. This calls for a general method for accurate and efficient gender inference from unlabeled data.

Despite much effort on gender inference under different settings, most existing methods focus on the task in specific fields-like Facebook, Twitter, Linked-In [4,5,6,7] and can hardly be generalized to other tasks. The state-of-the-art method maintains a name list containing frequent names for males and females, and achieves promising prediction performance by matching the user's name with the list. However, the performance is highly sensitive to the quality and size of the limited list, which considerably hurts its expected skyline.

Human experts can easily distinguish a Web page regarding a male user from that describing a female, by recognizing representative keywords in the lines. Luckily, in big data era, countless Web pages generated and indexed by search engines everyday, allows us to access abundant Web data about a strange name, from which we can train a machine to automatically find relevant keywords and infer the gender information from it.

In this work, we propose a general and accurate method for gender inference using Web data. Given a target user with her name and affiliation, we leverage a smart query construction to find the relevant Web pages, which are most likely to reveal her gender information, from a search engine like Google. Representative keywords are then extracted as features to feed a classification algorithm for gender inference. Experiments show that, even with very simple features and basic classification models, our method can easily outperform the state-of-the-art. Besides, we present a voting framework to incorporate all existing methods we know to further improve the performance to around 96.9%.

II. RELATED WORK

Traditional methods based on statistic are time-consuming and not suitable for the big data. Many automated approaches have been proposed to solve this problem. As far as I know, there are four types in the existing approaches: content-based, visual-based, name-based and other.

Content-based approach is widely used in the gender inference. This method extracts features from the text of users' post or the content of users' profile. Early works try to capture stylistic behavior of author's writings for classifying gender from blog posts [8]. Since social media became popular around the world and provide researchers a huge dataset, a growing number of papers focused on the users' gender inference based on these social media sites like Twitter, Netlog, Linked-In and Facebook [4,5,6,7]. [9] inferred gender from only the content of users' tweets. The features they used including word unigrams, hash tags, and psychometric properties. [10] proposed an approach of Twitter users' latent attributes inference including gender, age, and political affiliation from the Twitter profiles and postings of her friends. However, content-base approaches in gender inference have an high computational complexity due to the number of features generated from the text. Besides, these works rely on special data or website, it does not adapt to other applications.

Visual-based method is language independent, which is different from content-based method. As far as I know, visual information can be divided into two types. [11] and their later work [12] predicted gender using five color-based(e.g., the background color in a user's profile page) features extracted from Twitter profiles. The other type exploit the images posted by users on social media. [13] implemented gender inference by processing images in tweets. Even though those visual-based results indicated that there are differences between male

and female users, the performance are not enough to predict one's gender in real problem. And these methods also rely on the specific area or crowd.

Researchers employed name-based approach always need to prepare a dictionary of names in the first place. The work in [14] predicted user's gender by comparing the user self-reported name with the popular male names or female names. The main drawback of these approaches is that you cannot build a name list including all the names, different region and races have different naming style. Moreover, some names are difficult to predict gender directly.

Other approaches like [15] combined the content sentiment and name features; [6] combined the image information and text information to improve the performance by leverage many features, the drawback is the same with the last type approaches.

In summary, few existing studies have considered measures gender using a general approach to fit global situation, in this paper we proposed a novel solution to solve these problem. And our performance achieved around 97%, which is much higher than exist approaches.

III. General Gender Inference

In this section, we first introduce the Web-data based method to solve the gender inference problem, and then explain the voting framework in detail.

A. Web-Based Gender Inference

1) Basic idea: Given a person "*v*", referred to as query person, our goal is to infer the gender (male or female) of the person with high credibility. We aim to design a general method to automatically infer the gender from the Web. The method should be also flexible enough to extend to the entire group.

It is usually difficult to directly extracted gender information from the Web. Fortunately, some implicit evidences hidden in the big data from the Web can imply person's gender to some extent. A human expert can easily distinguish a Web page describing a male user from that of a female one, by recognizing some representative keywords. Based on this intuition, we construct a "smart" query for the query person to get relevant Web pages that are most likely to contain those representative keywords, and then apply a supervised classification model for the inference task.

2) Smart query construction. We construct the query by automatically identifying representative keywords of gender, then combine person name and representative keywords together as the query. To find the representative keywords, we first collect several person names (e.g., 1000) from professional websites such as AMiner and LinkedIn. We then submit the corresponding person names as queries to search engines like Google to obtain top-k (e.g., 10) snippets. Among all the words in the snippets, we identify the most representative keyword as that with the highest TF-IDF scores [16]. The TF-IDF score of a word "*w*" in a category "*c*" (male, female) is calculated as follows:

$$\text{TF-IDF}(w,c) = (1+\log n(S_c,w))\log(1+\frac{|S|}{n(S,w)}) \quad (1)$$

where "*Sc*" denotes the snippets that belongs to category "*c*". Notation "*n(Sc, w)*" denotes the number of snippets in category "*c*" that contains the word "*w*". Notation "*n(S, w)*" indicates the number of snippets in all the categories that contains the word "*w*" and "|*S*|" is the number of all the snippets in all the categories.

Using the above method, we found that the most representative keyword is "her" for females, and is "his" for males. The query is then constructed as "name his OR her".

B. Feature Definition

In our experiment, we employed binary several features such as how many snippets contain "his/her" in the search results, whether the title in the search results contain person name and snippet contains the word "his/her" at the same time, whether "his/her" appears in the snippets of the top 3 returned search results, and the number "his/her" in all the search results. Each of the feature is actually 2 features: "his" associated with males and "her" associated with females.

C. Train and Classification

We employee support vector machines (SVMs) [17] as our train and learn model. Parameter values and kernel choices for the SVM are discussed in the source paper.

D. Voting Framework

In order to improve the performance as higher as possible, we combine two exist methods in our framework - Name List method and face recognition. Thus, voting framework contains three predictors:

1) Facebook generated name list predictor (FGNL). We use a method proposed by [18] as the name list predictor (to hereafter refer to as: FGNL). Most state of the art methods for inferring Gender depend on a list of common names for males and females. In [18], the authors proposed an approach that used data from Facebook to construct an expanded and high-quality name list. They matched the user's first name with the list to make the inference. If the first name is matched with one of the male names, the user is treated as a male, and vice versa. We while if the first name is found in neither the male names nor the female names, or in both the name lists, they make a random guess about the user's Gender.

2) Face recognition predictor (Face). People's gender would be easy get from their personal avatars. Based on this assumption, given a query person, we use the name of the person as query and then retrieve relevant pictures from Web image search engine. In our approach, we take the first images as the query person's avatar.

We employ the Face++ as the gender recognition tool. Face++ support developers an easy use API and provide a

higher accuracy among the face detection algorithms, which won the "300-Face in the wild" challenge during the workshop of ICCV2013. Given an image, the Face++ API will return the faces information including age, gender, race, etc. If there is only one face in the image, we take the gender information as the query person's gender. Otherwise, the result is labeled as "unknown".

3) Web-based gender predictor(WebGP): Model described in subsection A.

Each predictor is expected to give an inference result as its "vote". However, the FGNL model cannot promise to give an answer when the query name isn't shown in the name list; same thing happens when the Face model cannot recognize any human face in the picture. A predictor will give up its vote under any of above situations. After this voting process, we choose the most "voted" gender label as our prediction.

IV. Experiment Result

In this section, we evaluate the effectiveness of our approach and compare to other exist approaches with real data set. Our approach has already been applied to an online academic search and mining system AMiner.org to infer the gender for researchers.

A. Experiment Setup

1) Dataset: To construct a ground-truth dataset for quantitative evaluation, we randomly choose 2,700 researchers from AMiner.org [19]. Several human annotators help annotate the gender for those 2,700 researchers. For disagreements in the annotation, we conducted "majority voting". Finally, for the 2,700 researchers, we got 1,416 female candidates and 1,284 male candidates. Specifically, in order to infer gender, we search the Web by querying the person name and the word "his OR her", and then generate 2,700 gender documents for the task, each document contains the {Title, URL, Snippet} which extracted from the search results. To construct the avatar data, for each person, we parse his/her images from the web, and then select the first picture as the representative, thus we generate 2,700 images.

2) Evaluation metrics. To quantitatively evaluate our model, we divide the dataset into training set and test set. We perform five-fold cross-validation and report the extraction performance in terms of precision, recall, and F1-score.

3) Comparison methods. We now compare our approach with several existing methods which mentioned in the last section for the task of gender inference on the ground-truth dataset.

- Facebook Generated Name List.
- Face Recognition.

All experiments are conducted on a Macbook Air with Intel Core i7 CPU 1.7GHz(2 cores) and 8 GB memory. In all the experiments, we search top 10 results by Google search and Google Image search, and conduct a five-fold cross validation for each method.

B. Evaluation Performance

Figure I shows the accuracy of gender inference with different methods. It can be clearly seen that the accuracy performance improves significantly using Web data. In addition, voting framework can further improve the performance to around 96.9%. The method using face recognition is the worst because the accuracy of the method depends on the pictures quality. Fuzzy images, group photos may lead the gender inference failed.

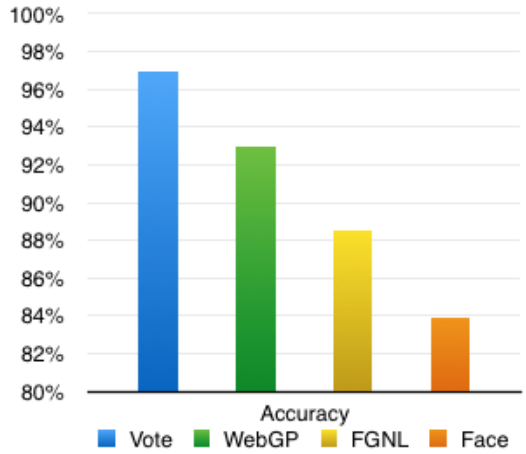

FIGURE I. ACCURACY COMPARISON OF GENDER INFERENCE.

TABLE I. PERFORMANCE COMPARISON OF GENDER INFERENCE (%)

Method	Precision	Recall	F1-score
Face	79.33	89.36	84.03
FGNL	92.98	82.53	87.12
WebGP	93.38	91.62	92.47
Vote	96.94	96.59	96.76

Table I shows the detail classification performance of gender inference by different methods. We can see that our Web-based gender predicts method achieves better overall performance than the baseline. And the voting framework further improves the overall performance. From Table I, we can infer that our Web-based gender predicts method and voting framework performs much better than the FGNL method in recall (+9.09%, +14.06%). This is because the FGNL method depends greatly on the name list. However, you can never list all those names, no matter how large the list is. On the contrary, our approach can automatically find the representative keywords for documents describing a user with specific Gender, and infer Gender from the big Web data with less limitation. So we seldom have the problem that the FGNL has to face when they cannot find the name in their list.

V. Conclusion

In this study, we proposed a general and accurate gender prediction method using Web data. For a given person, the

approach first constructs a meaningful query to retrieve information from a search engine. Without knowing his/her blog or social media content, we employ a classification model and infer the target person's gender from the search results. We test the proposed method on real data sets. Our experiments show that the proposed method significantly improves the gender inference performance in comparison with baseline methods. Besides, we present a voting framework to incorporate three predictors to further improve the performance to around 96.9%.

REFERENCES

[1] K. Roberts, "Engaging more women and girls in mathematics and stem fields: The international evidence," 2014

[2] S.-J. Leslie, A. Cimpian, M. Meyer, and E. Freeland, "Expectations of brilliance underlie gender distributions across academic disciplines," Science, vol. 347, no. 6219, pp. 262–265, 2015.

[3] R. van der Lee and N. Ellemers, "Gender contributes to personal research funding success in the netherlands," Proceedings of the National Academy of Sciences, vol. 112, no. 40, pp. 12 349–12 353, 2015.

[4] J. D. Burger, J. Henderson, G. Kim, and G. Zarrella, "Discriminating gender on twitter," in Proceedings of the Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2011, pp. 1301–1309.

[5] C. Peersman, W. Daelemans, and L. Van Vaerenbergh, "Predicting age and gender in online social networks," in Proceedings of the 3rd international workshop on Search and mining user-generated contents. ACM, 2011, pp. 37–44.

[6] A. Kokkos and T. Tzouramanis, "A robust gender inference model for online social networks and its application to linkedin and twitter," First Monday, vol. 19, no. 9, 2014.

[7] A. Panchenko and A. Teterin, "Detecting gender by full name: Experiments with the russian language," in Analysis of Images, Social Networks and Texts. Springer, 2014, pp. 169–182.

[8] A. Mukherjee and B. Liu, "Improving gender classification of blog authors," in Proceedings of the 2010 conference on Empirical Methods in natural Language Processing. Association for Computational Linguistics, 2010, pp. 207–217.

[9] C. Fink, J. Kopecky, and M. Morawski, "Inferring gender from the content of tweets: A region specific example." in ICWSM, 2012.

[10] F. Al Zamal, W. Liu, and D. Ruths, "Homophily and latent attribute inference: Inferring latent attributes of twitter users from neighbors." ICWSM, vol. 270, 2012.

[11] J. S. Alowibdi, U. A. Buy, and P. Yu, "Language independent gender classification on twitter," in Advances in Social Networks Analysis and Mining (ASONAM), 2013 IEEE/ACM International Conference on. IEEE, 2013, pp. 739–743.

[12] J. S. Alowibdi, U. A. Buy, and S. Y. Philip, "Say it with colors: Language-independent gender classification on twitter," in Online Social Media Analysis and Visualization. Springer, 2014, pp. 47–62.

[13] X. Ma, Y. Tsuboshita, and N. Kato, "Gender estimation for sns user profiling using automatic image annotation," in Multimedia and Expo Workshops (ICMEW), 2014 IEEE International Conference on. IEEE, 2014, pp. 1–6.

[14] A. Mislove, S. Lehmann, Y.-Y. Ahn, J.-P. Onnela, and J. N. Rosenquist, "Understanding the demographics of twitter users." ICWSM, vol. 11, p. 5th, 2011.

[15] W. Liu and D. Ruths, "What's in a name? using first names as features for gender inference in twitter." in AAAI Spring Symposium: Analyzing Microtext, vol. 13, 2013, p. 01.

[16] R. Baeza-Yates and B. Ribeiro-Neto, Modern Information Retrieval. ACM Press, 1999.

[17] T. Joachims, "Making large scale svm learning practical," Universitat Dortmund, Tech. Rep., 1999.

[18] C. Tang, K. Ross, N. Saxena, and R. Chen, "Whats in a name: a study of names, gender inference, and gender behavior in facebook," in Database Systems for Adanced Applications. Springer, 2011, pp. 344–356.

[19] J. Tang, J. Zhang, L. Yao, J. Li, L. Zhang, and Z. Su, "Arnetminer: extraction and mining of academic social networks," in Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2008, pp. 990–998.