Machine Learning-based State-of-the-art Approach to Identifying the Person behind an E-mail ID and Generating a Ingenuity Score based on PredictionIO

Anu B Nair
M.E Student
Dept. of Computer Science and Engineering
Gnanamani College of Technology
Pachal, Namakkal, Tamil Nadu
India 637018

R Umamaheswari
Asst. Professor
Dept. of Computer Science and Engineering
Gnanamani College of Technology
Pachal, Namakkal, Tamil Nadu
India 637018

P Kuppusamy
Head of the Department
Dept. of Computer Science and Engineering
Gnanamani College of Technology
Pachal, Namakkal, Tamil Nadu
India 637018

Abstract- As the technology grows, the data is available over the web and the demand of data extraction has been increased dramatically. As the amount of data increases, knowledge discovery become tedious. As these data is being generated by global data sources such as social media, the companies and the researchers need to collect these information to extract the details about a person for generating sellable data about a person. Mainly eCommerce companies are investing lot of money for these data and build there own recommendation engine, and they can easily sell there product. This paper proposes a simple solution to find information about a person just from email-ID, by data mining, web scraping and from social media platform. It also provides a ingenuity score based on the information collected. So the collected this score can be used for fraud prevention in business.

Keywords – Data mining, Prediction, Social Media, Machine Learning.

I. INTRODUCTION

With the growth of technology, the billions of internet users generate an ever increasing amount of data. The data is being generated continuously from various data sources such as social media, web sites, mobile application etc. Without proper processing of these data, this may be useless. If these data is properly analyzed and put to use, it has an immensive value. Here, we are finding a person behind an email-ID, also we will be providing a score based of the data.

eCommerce companies have enough resources to invest in their own machine learning solutions, so that they can sell appropriate product to the customer. Few companies are even selling this information to eCommerce companies. These information is being used for their recommendation engine. Other than the eCommerce companies, few companies use this information to find if fraudulent transactions. Some companies use for their business analysis

and optimization. This paper proposes a simple solution to find the person behind an email from big data and providing a score based on the data we found about the person by giving the data to Predictive Engines.

II. RELATED WORK

There are a number of works available in the literature that relate to Machine Learning techniques. Of these, a few which might be relevant for this work were identified, as follows:

Zheng Lin et al. [1] suggested traditional recommendation methods are limited functional in SNS because they only consider the content in web when try to find target customers. This paper proposes the use of the network structure to derive which actors in the SNS are more influential, then using word of mouth thread to fulfill the recommendation.

Hui-Ju Wu et al. [2] proposes to analyze the characteristics of users in social networking Web sites and in Web sites. This research expects to use the techniques of social network analysis and Web mining to discover the interest groups. This work in a way is based on machine learning technique and has been vitally inspiring toward our work.

Panos Fitsilis et al. [3], in described how social networks can be used for project management. The social network analysis is based on the probability-based model which is vital during any social network analysis and hence relevant to our work.

Wang Yong-gui et al. [4], in , discussed how semantic web and web mining works and proposes a standard framework for semantic web analysis. This framework for an analytic agent is the backbone of any mining system that runs on the internet. Agent in this work is inspired by the framework proposed by Wang Yong-gui et al.

There are three main types of machine learning [5, 6, 7, 8, 9]. They are:
(a) Supervised Learning, here the training set is labeled. The training data consist of training examples.
(b)Unsupervised Learning, here the training set is labeled.
(c) Semi-Supervised Learning, It is a type of supervised learning techniques that uses unlabeled data for training.

Jun Lu [10] proposes a predictive control strategy for engine control. The engine is being developed by mapping the rotation speed as function. The simulation results are used for prediction and the rotation speed is able to be maintained during abrupt changes in the load.

III. PROPOSED METHODOLOGY

In the proposed approach, we search the email in social media services, web search, people search services and blogs and the data is giving to data processing algorithm. Here the data is being processed and the wrong data about the person is eliminated. Then the processed data is giving to predictive engine and the score is being generated.

Fig. 1 shows the skeleton of the proposed ingenuity score calculation system. First we need to build a predictive engine server for Ingenuity Score generation. We have used PredictionIO for Building a Predictive Engines. We will create an event server by feeding sample events. By default, the deployed engine binds to http://localhost:8000. Once the engine is ready to use, by using the python client, we will be able to query to get the score. We then query the score using REST API. Predicted result is getting generated from classification engine.

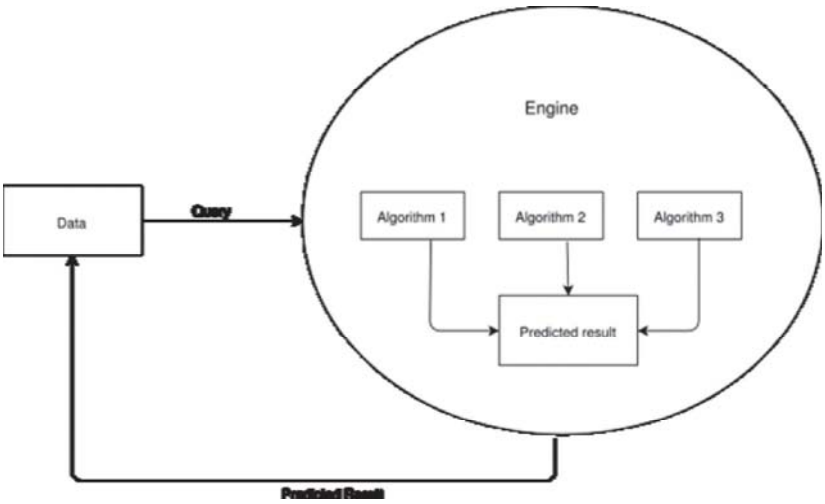

Fig 1. Logic of Ingenuity Score Generation and Predictive engine

Then the data is stored in database. Since all the search results are expensive, in-order to avoid repetitive search of same email the proposed system will store result in database. Storage is inexpensive comparing to CPU intensive process. Will store the data into memcache kind of storage (stored in RAM to reduce the number of times an external data source (such as a database or API) must be read.) Results are stored in json format, and it is stored as key-value pair (email as key).

IV. ARCHITECTURE DIAGRAM

Fig. 2 conceptualizes the architectural details of the proposed system. It shows interfaces of communication as well as the sources of data.

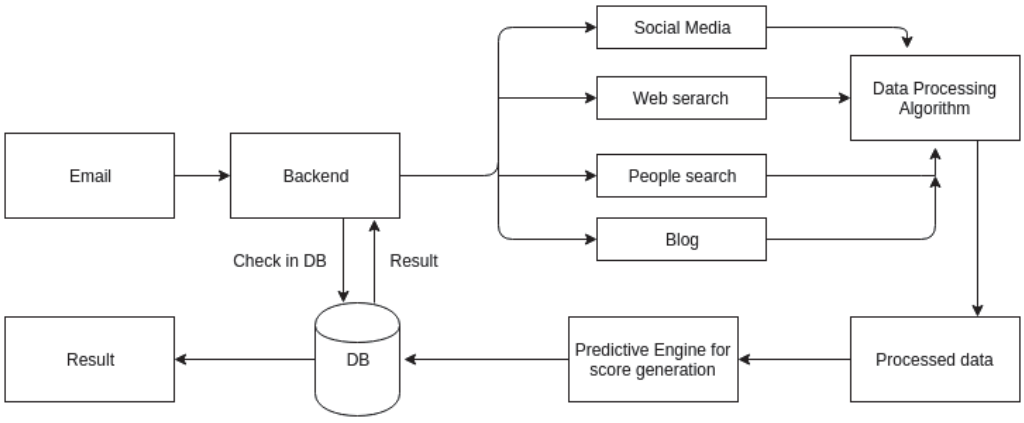

Fig 2. Architecture for search by email analysis system

V. ALGORITHM

The skeleton of the proposed algorithm is given in Algorithm 1.

Algorithm 1 The Proposed Algorithm

Input: An Email ID.
Output: A json file that contains the information about the person behind the email ID, collected from the open APIs of social networking sites.

1: **procedure** FIND–PERSON–DETAILS(EmailID)
2: Identify the social networks to probe into.
3: Adapt the system to send and retrieve requests/reply from/to the APIs of the network sites.
4: Retrieve all publicly available data related to the email ID and store it. Parse the tokens related to the information obtained and build meaningful database.
5: Identify the friends/connections of the person behind the email ID obtained in Step4.
6: Repeat Step3 for all such friends/connections obtained in Step5.
7: Give the processed data to predictive engine for score generation.
8: Store data and score in database.
9: **end procedure**

VI. ADVANTAGES OF THE PROPOSED METHODOLOGY

In the proposed approach, the main aim is to reduce cost and time. The advantages may be briefed as follows:

- It does the search of multiple resources parallel in-order to do with an acceptable time.
- Stores the search results in memcache storage (Stored in RAM to reduce the number of times an external data source, such as a database or API, must be read).

VII. RESULTS AND DISCUSSIONS

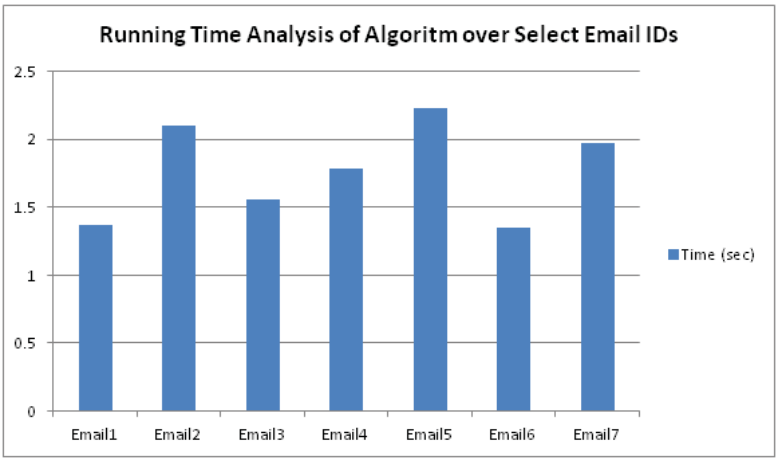

Fig 3. Running Time Analysis of Select Email IDs

The skeleton of the suggested approach was tested against a set of test cases and the positive results were obtained. A sample output is given below.
Input: An email ID (anoopvalluthadam@gmail.com)
Output: A json, which contains, list of matched result of email id. Sample output is given below:


```
{
    "score" : 2,
    "matched_result": [
        {
            "gplus": {
                "profile_pics": [
                    "https://lh3.googleusercontent.com/-
hGHDX6GgXq8/VkDZbU5pGSI/AAAAAAAAZuY/3fsVNOBwuew9D7lTKZoUh1gsuvrSNpJGACL0B/s0-d/5997aaf5-
6db1-497b-b362-b88ec9025741",

                ],
                "user_id": "107317830754480282820",
                "skills": "C, Python, Network Programming.",
                "gender": "Male",
                "apps": {},
                "profiles": {
                    "http://www.facebook.com/Anoop.Valluthadam":
"//s2.googleusercontent.com/s2/favicons?domain=www.facebook.com&amp;alt=p",
                },
                "birthday": "May 18",
                "followers": "384",
                "location": "Bangalore",
                "full_name": "Anoop Valluthadam",
                "twitter_handle": "anoopvalluthada",
                "education": [
                    [
                        "VMKV Engg.College",
                        "2007 - present"
                    ],
                    [
                        "NHSK",
```

```
                    ""
                ]
            ],
            ],
            "wordpress_url": "",
            "occupation": "Software Developer"
        }
    },
    {
        "youtube": {
            "profile_url": "http://www.youtube.com/user/anoopvalluthadam",
            "videos": []
        }
    },
    {
        "picasa": {
            "user_id": "107317830754480282820"
        }
    },
    {
        "vimeo": {}
    },
    {
        "gravatar": {
            "mobile": "9495326611",
            "profile_pics": [
                "http://0.gravatar.com/avatar/d81fe4c1b430e634dc60a87161f83b95"
            ],
            "profile_url": "http://gravatar.com/anoopvalluthadam",
            "full_name": "Anoop Valluthadam",
            "bio": "I am a Computer Engineering Student,Interested in Free and  Open
source Technologies and Photography,Cricketer."
        }
    },
    {
        "whois": {
            "related_domains": []
        }
    },
    {
        "flickr": {}
    },
    {
        "twitter": {
            "profile_pics": [

"https://pbs.twimg.com/profile_images/668739214935089152/WmQXVuCq_400x400.jpg"
            ],
            "followers": "333",
            "screen_name": "anoopvalluthada",
            "full_name": "Anoop Valluthadam",
            "bio": "Programmer, Linux/Free software/Opensource Enthusiast ,Cricketer.
Interested in Photography(anoopvphotography.in)",
            "url": {},
            "following": "519",
            "location": ""
        }
    }
```

```
    ]
}
```

Table1: Comparison of run-time of Select Email IDs.

Email	Time (s)
Email1	1.4
Email2	2.1
Email3	1.6
Email4	1.8
Email5	2.3
Email6	1.4
Email7	2

The proposed algorithm was tested against 70 emails. Fig 3 shows an analysis of running time of the algorithm across 7 emaillDs. These 7 emailID were selected for plating running time analysis graph.. Fig. 3 shows an analysis of the running time of the algorithm over select seven input Email IDs. This running time analysis is given in Table 1.

VII. CONCLUSION AND FUTURE SCOPE

This paper proposes an effective method to identify all details of person from an email ID. It also provides a Ingenuity score based on prediction. In-order to get the satisfactory result, the prediction engine is trained by giving more data.

The machine learning approach has been taken only for predicting the score. Future modification would be giving learning capability to the computers so that inaccurate data about the person can be eliminated. Also, this can be used for a recommendation system based on the data already available may be developed as part of future enhancements.

REFERENCES

[1] Zheng Lin, Lubin Wang, and Shuhang Guo. Recommendations on social network sites: From link mining perspective. In Management and Service Science, 2009. MASS '09. International Conference on, pages 1–4, Sept 2009.

[2] Hui-Ju Wu, I-Hsien Ting, and Kai-Yu Wang. Combining social network analysis and web mining techniques to discover interest groups in the blogspace. In Innovative Computing, Information and Control (ICICIC), 2009 Fourth International Conference on, pages 1180–1183, Dec 2009.

[3] P. Fitsilis, V. Gerogiannis, L. Anthopoulos, and A. Kameas. Using social network analysis for software project management. In Current Trends in Information Technology (CTIT), 2009 International Conference on the, pages 1–6, Dec 2009.

[4] Wang Yong-gui and Jia Zhen. Research on semantic web mining In Computer Design and Applications (ICCDA), 2010 International Conference on, volume 1, pages V1–67–V1–70, June 2010.

[5] T.A. Arunanand, K.A. Abdul Nazeer, M.J. Palakal, and M. Pradhan. A nature-inspired hybrid fuzzy c-means algorithm for better clustering of biological data sets. In Data Science Engineering (ICDSE), 2014 International Conference on, pages 76–82, Aug 2014.

[6] Yaochu Jin and B. Sendhoff. Pareto-based multiobjective machine learning: An overview and case studies. Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on, 38(3):397–415, May 2008.

[7] J. Srivastava. Data mining for social network analysis. In Intelligence and Security Informatics, 2008. ISI 2008. IEE International Conference on, pages xxxiii–xxxiv, June 2008.

[8] Jason Brownlee. A Tour of Machine Learning Algorithms. http://machinelearningmastery.com/a-tour-of-machine-learning-algorithms/, 2013. [Online; accessed 19-December-2015].

[9] Mauro Ribeiro, Katarina Grolinger, and Miriam A.m. Capretz. Mlaas: Machine learning as a service. 2015 IEEE 14th International Conference on Machine Learning and Applications (ICMLA), 2015.

[10] J. Lu, W. Ma, Y. Han, Y. Gao, Z. Guo, X. Li, and H. Niu. Model predictive engine control using support vector machine. In Cyber Technology in Automation, Control, and Intelligent Systems (CYBER), 2015 IEEE International Conference on, pages 1569–1573, June 2015.