International Journal of Computer Applications (0975 - 8887)
Volume 136 - No.5, February 2016

A Machine Learning-based State-of-the-art Approach to Identifying the Person behind an E-mail ID

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

With the growth of internet and related technologies, data available over the web has increased dramatically. As the volume of data increases, the challenge to the computer scientists arises, as knowledge discovery becomes tedious. One of these discovery techniques, which would be widely required soon, would be to identify people and retrieve information about them through social media, via email IDs. In this paper, a state of the art technique is presented, based on Natural Language Processing, to identify details of a person behind an email ID, by scraping social media platforms.

Keywords

Data Mining, Social Media, Machine Learning

1. INTRODUCTION

As technology advances, billions of internet users generate an ever increasing amount of data. Data is everywhere. If these data is analysed, and put to use, it can bring in immense value. For example, finding a person behind an email id is challenging, and nowadays, eCommerce companies are paying huge amount of money for getting information behind an email, so that they can sell appropriate product to the customer.

Facebook is earning lots of money by selling personalized information about all of us, to the eCommerce companies [1, 2]. According to eCommerce companies, they can easily invest money allocated for marketing to personalized recommendations rather that preferring generalized recommendation. Also, lot of companies, especially payment gateway companies prefer to check email age (how old is the email), to detect credit card fraud. All logic behind this is to find person behind an email. To do this, giving learning capability to the computers is a way of work. This can be done by machine learning. Its a research work, continuously dedicated to making the computers more intelligent, and, also, understanding how to handle such big data.

2. 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.[3] suggested various recommendations that one should take care of while creating and maintaining social networks. They layout the commercial aspects of social networking and how the social data could be potentially used for monetary and business purposes, and details the concept of Link Mining in social networks.

Hui-Ju Wu et al.[4] proposed a methodology to combine the concepts of web mining in the blogspace spread across the internet to identify user 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., in [5], described how social networks can be used for project management. This social network analysis research work is probability-based model which is vital during any social network analysis and hence relevant to our work.

Wang Yong-gui et al., in [6], 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.

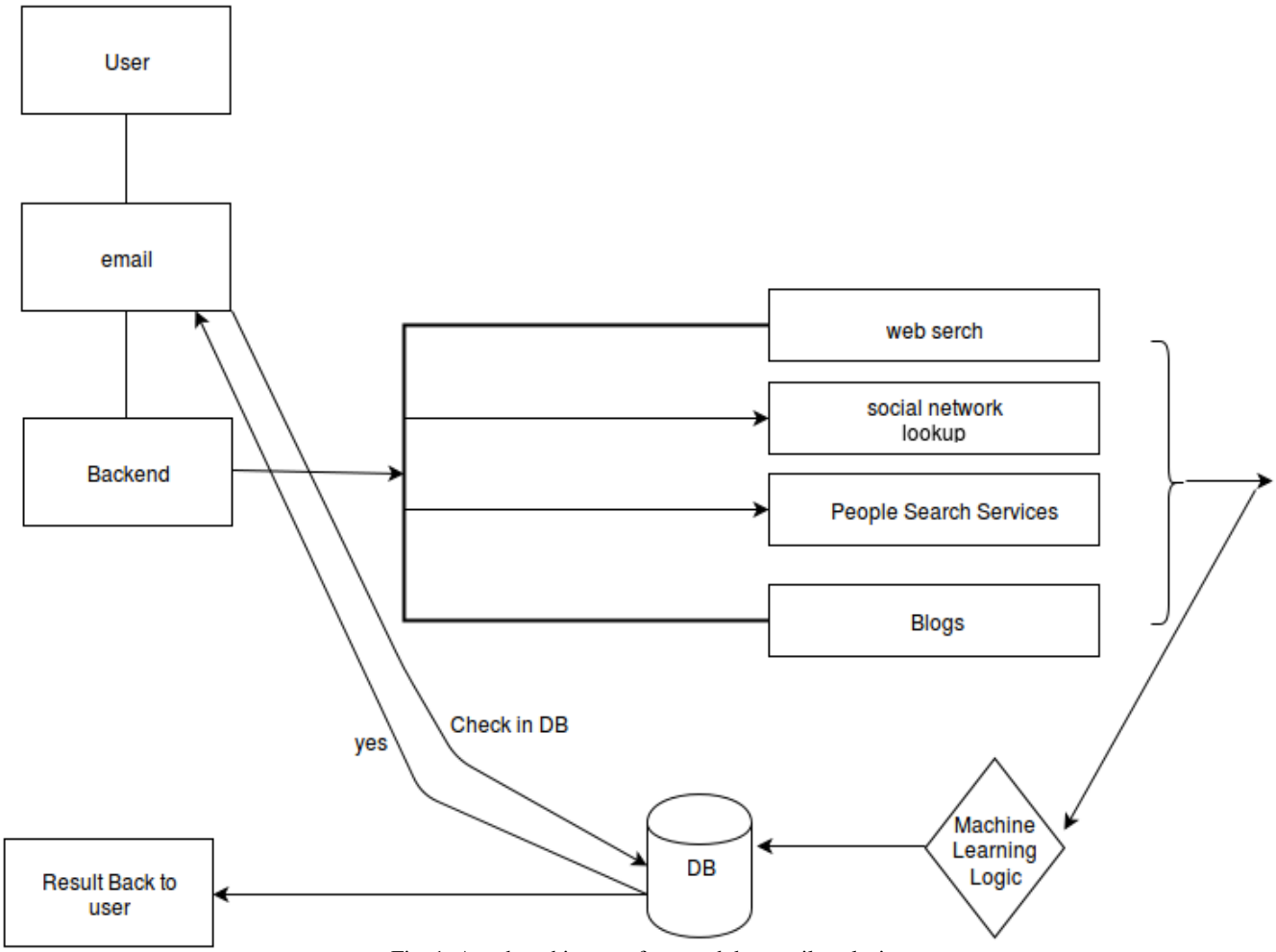

Fig. 1: A web architecture for search by email analysis.

3. MACHINE LEARNING ALGORITHMS

Machine Learning algorithms are normally grouped into three based on the Learning Style [7, 8, 9, 10]. They are:

—Supervised Learning

—Unsupervised Learning

—Semi-Supervised Learning

Of these, we use Supervised and Semi-supervised learning techniques, which are discussed briefly below.

(1) Unsupervised Learning
The training data will be fed to the unsupervised learning systems as input. This will have a known label or result. A model is readied through a training procedure where it is required to make predictions and is augmented when those results aren't right. The training and augmentation process proceeds until the model accomplishes a wanted level of exactness and quality.
In this work, the input will be an email id, gathered information is validated and training process continues till gets an accurate result. The predictions are corrected if it is wrong.

(2) Semi-Supervised Learning
In this kind of learning, feeder data is a blend of marked and unlabeled samples. A coveted prediction exercise would be defined and the model must understand the structures to arrange the input and in addition make prediction.

4. PROPOSED APPROACH

In the proposed approach, 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 compared to CPU intensive process. The date will be stored in memcache type 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).

5. ARCHITECTURE DIAGRAM

Fig. 1 conceptualizes the architectural details of the proposed system. It shows interfaces of communication as well as the sources of data.

6. ALGORITHM

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

Algorithm 1 Skeleton of 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 Step4 for all such friends/connections obtained in Step4, using Supervised and Semi-supervised learning.
7: **end procedure**

7. ADVANTAGES OF THE PROPOSED METHOD

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).

8. EXPERIMENTAL RESULTS

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 (anubnair90@gmail.com) Output: A json, which contains, list of matched result of email id. Sample output is given below:


```
{'result': [{'gplus': {'profile_pics': ['
    https://lh3.googleusercontent.com/-
    n7z9HAWhodU/VJD52DtvSbI/AAAAAAAAFZc/3
    WzQXvsL_Vc/s0-d/IMG_20141013_091846.jpg
    ', 'https://lh3.googleusercontent.com/-
    WoB9A8XmPLA/U3w7HFzkSYI/AAAAAAAAB-0/
    LeDGH0VzjyU/s0-d/photo2.jpg', 'https://
    lh3.googleusercontent.com/-YcIUpIv7knI/
    Thg5kqRm52I/AAAAAAAAALs/5AGcEJMuKGY/s0-d
    /ann.png']}]}
```


Table 1. : 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

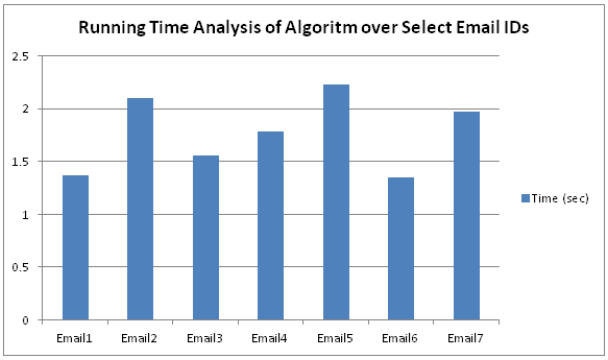

Fig. 2: Running time analysis of Algorithm over Select Email IDs.

The algorithm was tested against a set of 70 Email IDs leading to satisfactory results. The Google profile details of those seven Email IDs were fetched by the program in the form of json. Fig. 2 shows an analysis of the running time of the algorithm over select seven input Email IDs. These seven email IDs were selected to plot the graph, as a representative figures as regard to closeness in terms of running time.

9. CONCLUSION AND FUTURE SCOPE

In this paper, an efficient methodology to identify the personal, professional and social details of a person behind an email ID, with the help of social networking sites and their open APIs is suggested. These recommendations were tested against a limited number of test cases, and yielded satisfactory results.

The peculiarity of the idea presented in this paper is its machine learning based approach. Future modifications to this approach would be to add more relevant features to extract more data of a person behind the email ID. Also, feature extraction in more social networking websites may be incorporated and a recommendation system based on the data already available may be developed as part of future enhancements.

10. REFERENCES

[1] Ching-Yung Lin, L. Wu, Zhen Wen, Hanghang Tong, V. Griffiths-Fisher, L. Shi, and D. Lubensky. Social network analysis in enterprise. *Proceedings of the IEEE*, 100(9):2759–2776, Sept 2012.

[2] Dong Liu, Li Wang, Jianhua Zheng, Ke Ning, and Liang-Jie Zhang. Influence analysis based expert finding model and its applications in enterprise social network. In *Services Computing (SCC), 2013 IEEE International Conference on*, pages 368–375, June 2013.

[3] 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.

[4] 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.

[5] P. Fitsilis, V. Gerogiannis, L. Anthopoulos, and A. Kameas. Using social network analysis for software project manage-

ment. In *Current Trends in Information Technology (CTIT), 2009 International Conference on the*, pages 1–6, Dec 2009.

[6] 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.

[7] 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.

[8] 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.

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

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