Жамият ва инновациялар –
Общество и инновации –
Society and innovations
Journal home page:
https://inscience.uz/index.php/socinov/index

Applying data science for the categorization of plain text email spam

Azam NURULLAEV[1]

University of Cumberlands in KY

ARTICLE INFO

Article history:
Received September 2024
Received in revised form
15 October 2024
Accepted 25 October 2024
Available online
25 December 2024

Keywords:
email spam,
clustering algorithm,
machine learning,
spam message classification,
genetic algorithm.

ABSTRACT

Over the last few years there has been offered diverse techniques and methods for removing email spam and some of them are intended to divide and classify them into different subgroups. A group of software engineers has developed several techniques, including a genetic algorithm, K-NN algorithm (which will find a set of K-nearest neighbors), and clustering method (classifying spam messages into several subclasses) to deal with these problems. However, the main function of all the above-mentioned techniques is to promote the user interface and experience of email spam messages by dividing them into subgroups or removing and blocking non-requested messages. This research project will explain algorithms related to eliminating email spam messages and put forward a new suggestions/methods to the problem.

2181-1415/© 2024 in Science LLC.
DOI: https://doi.org/10.47689/2181-1415-vol5-iss11/S-pp166-175

Маълумотлар фанини қўллаш орқали оддий матнли электрон почта спамларини категориялаш

Калит сўзлар:
электрон почта спам,
Кластерлаш алгоритми,
Спам хабарларини
таснифлаш,
Генетик алгоритм.

АННОТАЦИЯ

Сўнгги бир неча йил ичида электрон почта спамларини олиб ташлаш учун турли техникалар ва методлар таклиф қилинди, уларнинг баъзилари спам хабарларини турли кичик гуруҳларга бўлиш ва таснифлашни мақсад қилади. Бир гуруҳ дастурчилар бир нечта техникаларни ишлаб чиқдилар, жумладан, генетик алгоритм, К-НН алгоритми (К-энг яқин қўшниларни аниқлаш), кластерлаш усули (спам хабарларини бир нечта кичик синфларга таснифлаш), бу муаммоларни ҳал қилиш учун. Бироқ,

[1] Master's Graduate, University of Cumberlands in KY, USA. E-mail: nurullayevazam4@gmail.com

юқорида санаб ўтилган техникаларнинг асосий вазифаси фойдаланувчи интерфейсини ва электрон почта спам хабарларининг тажрибасини яхшилаш, уларни кичик гуруҳларга бўлиш ёки сўров қилинмаган хабарларни олиб ташлаш ва блоклашдир. Ушбу тадқиқот лойиҳаси электрон почта спам хабарларини йўқотиш билан боғлиқ алгоритмларни тушунтиради ва муаммони ҳал қилиш учун янги таклиф/усулни илгари суради.

Применение науки о данных для категоризации спама в виде обычных текстовых электронных писем

Ключевые слова:
спам в электронной почте,
Алгоритм кластеризации,
Машинное обучение,
Классификация спам-сообщений,
Генетический алгоритм.

АННОТАЦИЯ

За последние несколько лет было предложено множество различных методов и техник для удаления спама в электронной почте, некоторые из которых предназначены для разделения и классификации спама на разные подгруппы. Группа инженеров-программистов разработала несколько методов, включая генетический алгоритм, алгоритм K-NN (который находит набор K-ближайших соседей), метод кластеризации (классификация спам-сообщений на несколько подклассов) для решения этих проблем. Однако основная цель всех вышеупомянутых техник заключается в улучшении пользовательского интерфейса и опыта работы с сообщениями электронной почты, разделяя их на подгруппы или удаляя и блокируя нежелательные сообщения. Этот исследовательский проект объяснит алгоритмы, связанные с удалением спама в электронной почте, и предложит новый метод решения этой проблемы.

INTRODUCTION

In recent years, mail has gained tremendous popularity in our society. Because in a really good and cheap tool to flourish one's business. Business owners can send emails to their clients and keep them updated about the products and discounts that they offer to attract more customers. Albeit, this fame also caused the birth of unexpected and unnecessary commercial messages that distract people by giving not valid sources of information. Currently, in every second millions of spam messages are sent to people's mail addresses worldwide, and consequently, both the market and client started to get affected by this trend. These spams are not only a serious problem for the market, but also it's one of the biggest headaches of ISPs.

Research conducted by a group of data scientists and software engineers from Charles Darwin University in Australia (on 12 June 2019) showed that more than 270 billion email addresses have been exchanged in a market and out of them almost 57% of them were spam that provided users not valid source of information. Even hackers started to utilize mail spam messages to hack customers' login & password credentials by sending spam consisting of login pages of their Apple accounts, reported and published by Bloomberg on 19 January 2016[1].

Even though so many scientists and engineers created so many algorithms that try to remove spam messages, there still remain millions of uncontrolled messages.

Therefore, alternative and efficient algorithms are being created by data scientists around the world while taking this obstacle and problem on a global scale. Some engineers offered a solution by dividing all these messages into 2 subgroups (black and white list), whereas blocked IP addresses, their locations, and other blocked spam go to the black list, while the rest go to the white list. But, in this fast-growing society, it is an almost impossible and never-ending job. Therefore, other alternative solutions were put forward. One of them is applying machine learning by testing on millions of spam and acquiring certain knowledge and skills to distinguish which mail is valid and which one is a spam message.

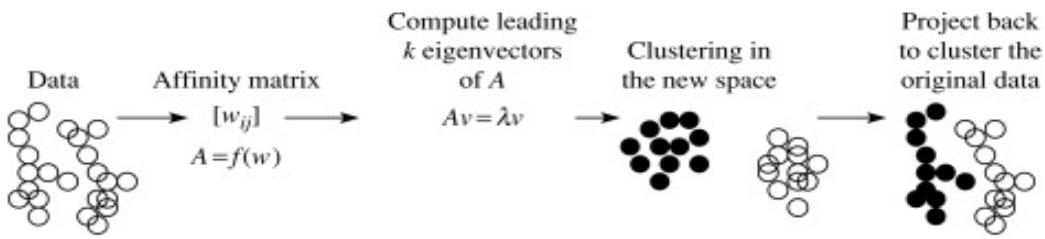
This is the goal of this report to explain and show an example of efficient antispam algorithms used by giant internet companies and also put forward a new solution/suggestion to the problem.

Related Works

Since scientists started looking at spam on a global scale, several algorithms offered by engineers around the world. One of the efficient methods was to classify mail spam. This method is also called as clustering method. In an annual conference in 2007, Project Honey Pot outlined to manage all bulk mail and luckily it was successful and worked very productively. This project takes spam messages as a whole network tree and traces these messages till it finds this tree's nodes and blocks them [2]. This project was managed and conducted by Unspam Technologies, Inc. Later on, in our research report, we show an example of how clustering works and what its algorithm looks like.

Another successful technique had been offered and it was called Salton's vector space model. This model is still widely used in our society by many internet companies. The reason behind it is very simple. This model is mainly created on linear algebra/Calculus and has several advantages over the Honey Project. For example, all elements of this model weigh not a binary term. In addition, it uses vector analysis as one of the base elements. However, this model crashes when it comes with large data and keywords must be exact, otherwise, this function returns the opposite Boolean value.

It needs to be highlighted more information about classifying Email messages or in other words clustering method because it more advanced version of clustering is based on the combination of the neural network and machine learning. This technique has already overcome traditional methods. Classifying has covered several core elements, like affinity matrix more data manipulation, eigenvector, and coming back to the same data, but in categorized subgroups.



Even Google is advancing method and if you take a look at Gmail you can see that spam messages are categorized and divided into several subgroups.

A new generation of clustering is a genetic method that is based on complex logic and computations. For instance, when data scientists conducted a survey over genetic methods in combination with clustering, it showed that in this algorithm more precise accurate information was displayed.

In this report, we with teammates decided to utilize and show one simple example of classifying method and for the conclusion of it, filtering email spam produced more accuracy even in large data. You can see one simple example in section 4.

External damages caused by Spam

In the 21st century, email has been one of the crucial aspects of our life and work causing about 100 billion emails to be sent to valid email addresses around the world. Unfortunately, most of these emails are not valid information, in 2010 approximately 88 percent of this email traffic was spam (2010 Symantec; 2011 MAAWG) [3]. Under current laws, nearly all of these spam emails are illegal as they don't give the choice for customers to decide. On the other hand, they don't provide any kind of value as a return to an advertisement they share with customers unlike companies YouTube, Facebook, Instagram, WeChat, and many more which do legitimate advertisements.

The external effects of spam are also considerably high when we take the ratio of external costs(damage) to private profit. It is estimated that almost 20-50 billion a year is sent by customers and private firms in America because of spam, and It is possible to imagine how huge these numbers get if we estimate it on a global scale. But as a return spam-advertised merchants make only about $200 million in America and $15 billion worldwide. This makes the externality ratio even more than 100, which shows that it causes a drastic amount of damage to make a little benefit for the community of spammers. These estimated values are taken based on computer science professionals who monitored the activities of spammers over the curse of time in 2008 and 2011 (Holz, Vigna, Stone-Gross Paxson, Kanich, and so on). And most of the time this mass amount of spam being sent online can be described as digital pollution. It causes problems to customers to keep their emails ordered and organized and make them lose their concentration on their job related daily emails by making their inbox totally mess [4]. On the other hand, email providers also get damaged as the value of their service is decreased by the external factor of these spam. Of course, spam is not the only one out there causing digital pollution, there are some others like telemarketing, billboards, junk email, invalid emails, and spying apps that track your browsing history and send them to marketing companies. However, these types of digital pollution don't have such high externality and are not in use as much as spam.

Now to give how high this externality is we give some examples to compare.

Activity	*Revenue/benefit*	*Cost*	*Externality ratio*
Driving automobiles	$0.60 per mile	$0.02–0.25 per mile[a]	0.03–0.41
Stealing automobiles	$400–1200 million per year	$8–12 billion per year	6.7–30.3
Email spam	$160–360 million per year	$14–18 billion per year[b]	39–112

Sources: The source for the first row is Delucchi (1997), for the second row, Field (1993). (The *FBI Uniform Crime Report* (2010) places the vehicle value extracted by criminals in the same range as Field 1993.)

[a] Air pollution costs.

[b] Cost to end users.

Even though the external damage of spam emails sent is so high and not so efficient, it is still in use because of its very low cost compared to other types of advertisement. Here we provide a table showing the difference [5].

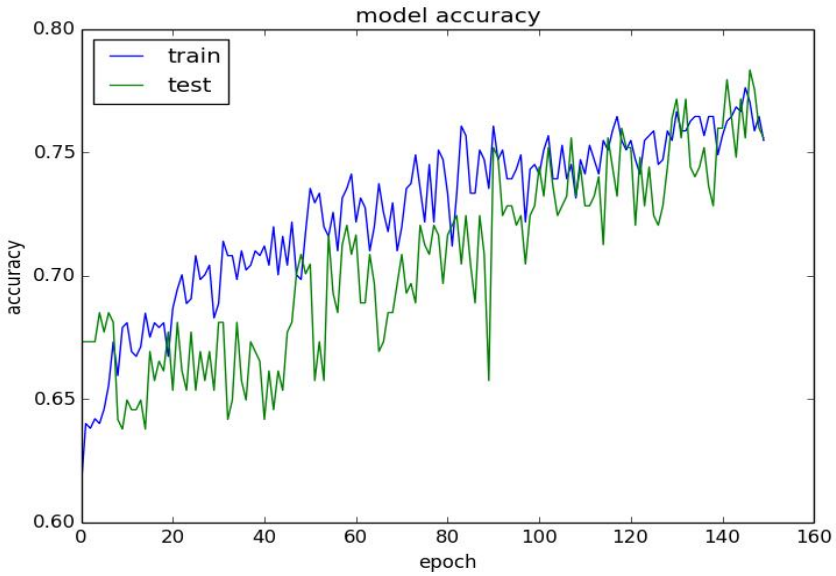
(Sources: U.S. Postal Service website, For Super Bowl advertising, **"http://money.cnn.com /2011/02/03/news/companies/super_bowl_ads?index.hrm"**)

Cost of Spam Advertising Relative to Other Advertising Media
(cost per thousand impressions (CPM))

		Breakeven conversion with marginal profit = $50.00	
Advertising vector	*CPM*	*Percent*	*Per 100,000 deliveries*
Postal direct mail	$250–1,000	2–10%[a]	2000
Super Bowl advertising	$20	0.04%	40
Online display advertising	$1–5	0.002–0.006%	2
Retail spam	$0.10–0.50	0.001–.0002%	0.3
Botnet wholesale spam	$0.03	0.00006%	0.06
Botnet via webmail	$0.05[b]	0.0001%	0.1

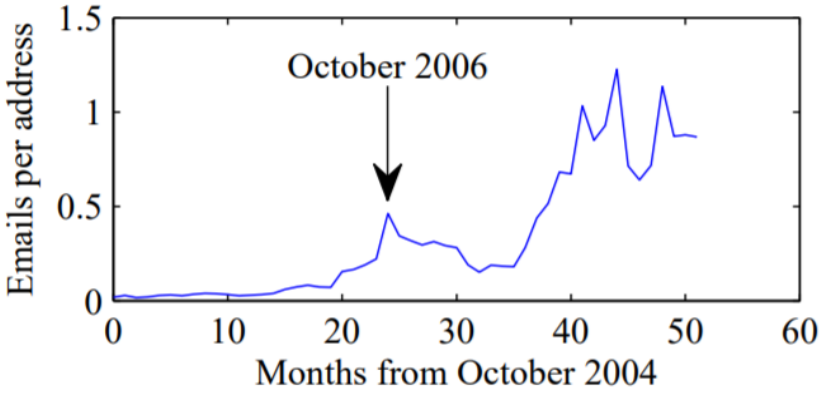
This spam problem was due to the behavior of SMTP protocol used to send an email over the network, which uses the push method, unlike most other protocols which use pull requests on demand. But at the same time, it was the main benefit of using this protocol, so spammers could not be completely erased by implementing new methods or protocols. However, Internet administrators improved authentication protocols: where previously one only had to type a password to collect one's incoming mail, now most had to authenticate themselves by providing a password to send outgoing mail. And lots of other methods developed over time including, crowdsourcing, IP blacklisting, machine learning, and data science to detect suspected spam messages and either reject them from being delivered or send them to a junk mail folder. It is important to note that in recent years machine learning and data science methods gained dominance. The reason for this is as data gets large efficacy of other methods drops drastically, for example, you cannot keep up blacklisting IP addresses manually because they need to be inspected for some time to conclude that suspected email addresses are actually in the business of spam. Moreover, they also cost a lot more and require more work making companies lose their attention to their main objective.

The machine-learning approach first began in the late 1990s. A typical machine-learning implement uses 'ground truth' data on a subset of observations to learn rules to classify the remaining data. They are at first trained with test data to differentiate between spam and valid mail and then as they continue to work they improve themselves.



Now we specifically point to the clustering method of machine learning and data science. Clustering helps to divide spammer accounts into groups save some additional data and identify the personals who are using these accounts for spamming purposes. By doing so it is possible to identify not only currently in use fake accounts but also future accounts that will be created by these personals. Other methods mostly concentrate on just detecting spam emails and blocking them but in fact, this is not enough these days. Because, as anti-spam methods develop, spammers also evolve their tactics and most of the time create new fake accounts repeatedly over some time. This is some sense that makes the process of finding spam accounts useless because they will not be used by spammers anyway after some time and immediately will be replaced by other fake accounts for the same purpose [6].

But with the clustering method, you can actually store some data about fake accounts, such as their domain, number of emails sent, advertisement type, linked sites, and many more. And by processing them you can find out some useful data as an output. The most typical ones include finding locations and personals that actually running this business. In addition to this, these people most of the time work as a group, and their accounts and spam email most of the timeshare similar content. And the behavior of email account usage can be a sign showing this account a suspect for spam or target of spam emails. The below graph shows the change of email usage over months.



Climax points can be seen as periods that spam email being generated in vast amounts, cause it is not usual to see exponential changes in the number of emails sent. And these time periods can be investigated and be linked to some processes happening. And as at this point data gets extremely large it is hard to manage large groups of data. However, with the clustering method, it is possible to subgroup large chunks of data and work with them closely.

Here some graphical illustration of clustering output:

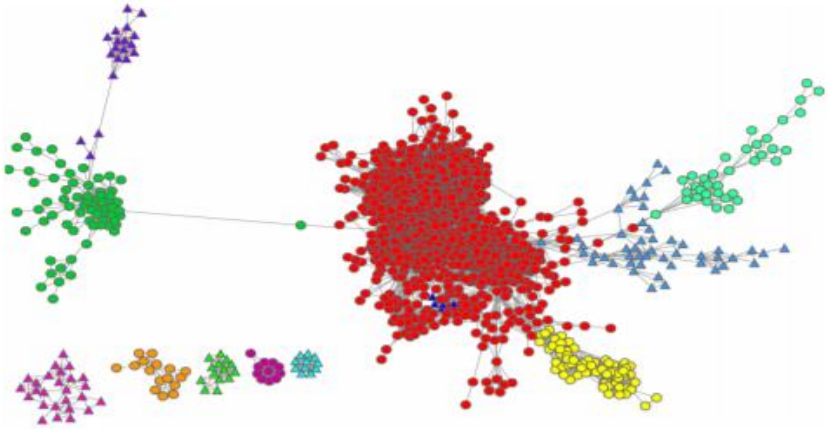
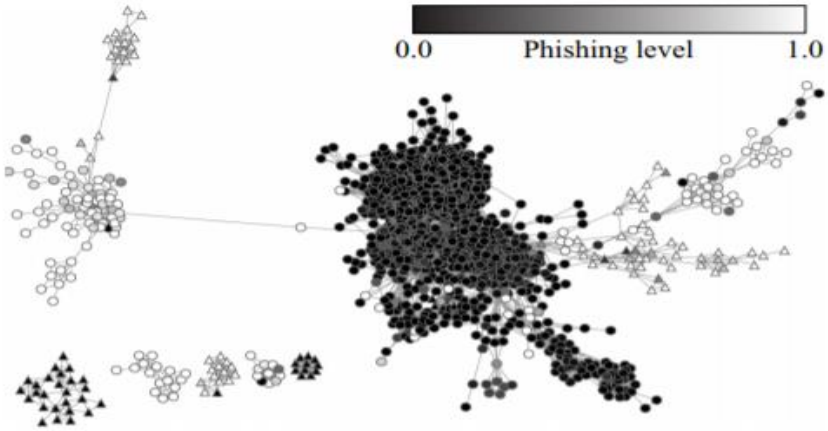


Fig. 3. Social network of harvesters formed by similarity in spam server usage in October 2006 (best viewed in color). The color and shape of a harvester indicate the cluster it belongs to.



Problem Statement

In our research project about email spam and effective algorithms to deal with this problem. We decided to demonstrate a simple, but very effective algorithm to deal with this problem. Let's get started!

Let's suppose that in our mail address, we have a collection of non-requested messages, and we need to deal with it via filtering. As we said in previous sections we are going to use the clustering technique to divide messages into categorized subgroups. Firstly, if we think like data scientists, we need to go through all the information about this spam and be aware of everything, like IP address, from which domain it was sent and of course, its location. Once, we went through all credentials we got a collection of messages and let's refer it as

Q = {q1, …, qn} and K = {k1, … km} that contains basic info to filter it. If you remember we said in our report that one of advantages of classifying is it is weighted.

For every qi where have its weight $w_q = \{w_{q1}, w_{q2}, \ldots. w_{qn}\}$

Once we did above mentioned steps, now it's time to do a simple math calculation:

$$\operatorname{sim}(s_i, s_j) = \frac{\sum_{l=1}^{m} w_{il} w_{jl}}{\sqrt{\sum_{l=1}^{m} w_{il}^2 \cdot \sum_{l=1}^{m} w_{jl}^2}}, \quad i, j = 1, \ldots, n.$$

This formula might be quite complex and requires some logic, but it will produce an accurate result[7].

We know that several algorithms exist worldwide, and every method can be efficient based on the data type, such as if your data is big or does it needs to filter your mail messages with exact matches or not.

CLUSTERING

Clustering in our case used as data clustering and it is the process of dividing data into subclasses or subclusters where items or we can say elements of a data set in the same cluster are max analogous and elements of a data set in different clusters are unrelated. It is highly dependent on the data nature and the aim for which classification is being used, disparate measures of likeness should be used to place data elements into clusters, so that resemblance measures control how the classification is formed. Knowing two of the cluster forms that are hard and fuzzy classification. In fuzzy (in some references it is known as soft clustering), data items can be related to more than a class, and connected with each relation is a set of membership levels. In the latter, information is divided into different subclasses, from which all of the attributes of a class is only be the child of the same parent.

Lets say we have a collection of spam messages and we denote it as 'S' so that

S={s1, s2, ….. , sn}

divided into non-overlapping clusters 'C' so that C= {C1, C2, ….. Cm}, where m>1, reason to do that is to maintain of highly likeness among elements of different parents to will refer to a certain topic and max destination among clusters. Below, we see how hard clustering should take place.

$$C_p \neq \varnothing \quad \text{for } p = 1, \ldots, q,$$

$$C_p \cap C_z = \varnothing \quad \text{for } p \neq z,\ p, z = 1, \ldots, q,$$

$$\bigcup_{p=1}^{k} C_p = S.$$

Let see the designations for this case:

$$O_{kNN}(s_i) = \left\{ s_j \mid \operatorname{sim}(s_i, s_j) \geq \operatorname{sim}(s_i^k, s_j) \right\}$$

Here we look at the set of *k* nearest neighbors of spam receiving message Si, where we consider S^k_i as k th nearest neighbor of spam message Si

$$u_{ij} = \begin{cases} 1 & \text{if } s_j \in O_{kNN}(s_i), \\ 0, & \text{otherwise,} \end{cases}$$

$$v_{ij} = \begin{cases} 1 & \text{if } s_i \in O_{kNN}(s_i), \\ 0, & \text{otherwise.} \end{cases}$$

Here we look at on two conditions of our function when it is equal to 1 and 0 so we can implement here Boolean operation on solving these problems.

$$x_{ip} = \begin{cases} 1 & \text{if } s_i \in C_p, \\ 0 & \text{if } s_i \notin C_p, \end{cases} \qquad i = 1, \ldots, n; \; p = 1, \ldots, q.$$

By taking a look at above conditions, we can simply say that both of the condition need to undergo through complex logic and mathematical calculations:

$$f(x) = \sum_{p=1}^{q} \sum_{i=1}^{n} \sum_{j=1}^{n} \left(u_{ij} + v_{ij}\right) \operatorname{sim}\left(s_i, s_j\right) x_{ip} x_{jp} \longrightarrow \max.$$

This is supposed that each class contains at least one spam message and does not contain all spam messages than we are not considering.

$$1 < \sum_{i=1}^{n} x_{ip} < n, \quad p = 1, \ldots, q,$$

Where X_{ip}

$$x_{ip} \in \{0, 1\} \quad \text{for any } i, p.$$

As we described in the above conditions spam messages are represented like the Boolean values that return either True or False [8]. These problems are called NP-problems, which require much time to process the problem and compute the result. In the next section, we will consider the algorithms to solve clustering problems. However, some algorithms solve huge computing expenses, the genetic algorithm is outstanding in this case of a problem that will be more confident to compute the number of spam messages.

Algorithm for Solving the Clustering Problem

Considering algorithms that solve large problems that take big time and a lot of resources, it should be taken a genetic algorithm that is most suitable to our case and easy to apply. However, they are not sure to find optimal solutions for the problem. The starting point in this algorithm is the coding of solutions to make it suitable in the formation of chromosomes that depend on the nature of a solved problem.

$$x_{ip} \in \{0, 1\} \quad \text{for any } i, p.$$

Where we consider values 0 or 1. For that coding, the number of chromosome equal to *n*q,* where the first *q* position corresponds to the first spam message, following *q* position to the second spam message.

Indeed, the genetic algorithms are most suitable for our problems, but when solving constrained problems genetic algorithms are faced by a problem of occurrence of Hedin solutions.

By applying the penalty function, we can remove unrelated results produced by operators of genetic algorithm and that will give more chances of save clear decision, this functions method allows to speed up process of convergence of algorithm. This is because the functions method at getting of not acceptable chromosomes does not demand performance of additional operations.

$$o_{pj} = \frac{1}{n_p}\sum_{d=1}^{n_p} w_{dj}, \quad p = 1,\ldots,q;\ j = 1,\ldots,m,$$

where n_p is a number of points in class and compactness of cluster Cp is calculated by the following formula:

$$r_p = \frac{1}{n_p}\sum_{i=1}^{n} \mathrm{sim}\left(s_i, O_p\right) x_{ip}.$$

The likeness of cluster that we consider here, we define it below as follows:

$$R_p = \frac{1}{q-1}\sum_{z=1}^{q} \mathrm{sim}\left(O_p, O_z\right), \quad p = 1,\ldots,q,$$

Where Op and Oz are resemblance between the centers of the clusters C [9].

CONCLUSION AND FUTURE WORK

In this report, we outline side effects of email spam can cause and also show some algorithms and their usage. As we said previously, all algorithms can be efficient and accurate in its way. For example, some algorithms are good for small amounts of data, whereas others get crashed when it comes to large amounts of data. Last, but not least, it has been a very controversial topic that encouraged data scientists and software engineers to do some experiments on this topic and create an alternative method that can be suitable for small and large amounts of data while producing high accuracy.

In future works, as we said as our own new suggestion combining the classifying method along with finding the nearest possible neighbor algorithm can be very efficient since it will be suitable for any sized data and filtrations will not be based on an exact matching system.

REFERENCES:

1. A group of scientists from MIT (Massachusetts Institute of Technology) University in their report The Community Behavior of Scammers".

2. Spectral clustering by Honey Pot security tool. IEEE International Conference on Communications.

3. Salton – Vector model. "Information Processing and Management" book, [513-523]

4. International journal regarding IT advances and technologies called "Journal of Automation and Information Sciences" journal [52-58]. ISSN: 10642315

5. "On clustering validation techniques" Journal of Intelligent Information Systems [107–145]

6. "A review: accuracy optimization in clustering ensembles using genetic algorithms" by Ghaemi [287-300]

7. Methods in Hard and Fuzzy Clustering "Soft Computing and Human-centered Machines"

8. Clustering or Classifying method usage: "Soft and Hard clustering in International Journal of Computer Trends and Technology" [108-113] ISSN:22312803.

9. Applying effective algorithms "Automatic Hard Clustering Using Improved Differential Evolution Algorithm" [137-174]