ISSN No. 0976-5697

Volume 3, No. 1, Jan-Feb 2012

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

Near Duplicate Matching scheme for E-mail Spam Detection using Spam Trees

Ch.Vijaya Kumar*
PG Student
Sri Sai Aditya Institute of Science and Technology
Surampalem,Andhra Pradesh, India
vkumar.ch@gmail.com

G.Santi
Assistant Professor, Dept of CSE
Sri Sai Aditya Institute of Science and Technology
Surampalem,Andhra Pradesh, India
santhi.ssait@gmail.com

Abstract: One of the major problems that the users of Email in the internet are facing is spam mails or e-mail spam. In recent years there are so many schemes are developed to detect the spam emails. The basic idea is to have a similarity matching scheme for spam detection by maintaining a known spam database, formed by users feedback, to block the subsequent near-duplicate spam's. We propose a novel e-mail abstraction scheme, which considers e-mail layout structure to represent e-mails using HTML content in email which effectively captures the near duplicate phenomenon of Spam mails. To detect near duplicates and duplicate spam mails faster, we propose a new approach SimHash.

Keywords: Spam mails, Emails, Near Duplicate, SimHash, Spam Trees

I. INTRODUCTION

Internet is the most widely used area. In internet most widely used are E-mails. E-mails play a major role for the communication between the people .The people who are using emails cannot verify the duplicate and near duplicate web documents creating the more problems on the web search engines. These documents will increase the space required to store the index, slow down the searching results and the annoy users. According to the data availability on the internet, the huge data are shorts texts such that mobile phone short messages, instant messages, chat log, BBS titles etc.

The statistical information is given by the Information Industry Ministry of china that more than 1.56 billion mobile phone short messages are sent each day in Mainland China. You already know how much of email is spam, but here are a bunch of other factoids as per [1] you may not be aware of:

a. **90%** of spam is in English. A year ago it was 96%, so spam is getting more "international."
b. **88%** of all spam is sent from botnets (networks of compromised PCs).
c. **91%** of spam contains some form of link.
d. Unsolicited newsletters are increasing and are now the second most common type of spam.
e. Spam from webmail services like Gmail and Hotmail isn't as common as you might think. Only **0.7%** of spam is sent from webmail accounts.
f. **1 in 284** emails contain malware.
g. **1 in 445** emails are phishing emails.
h. As many as **95 billion** phishing emails were in circulation in 2010.

Unfortunately, the status of duplicate and near duplicate messages is very complex. Among these especially near duplicates and spam mails.

These differences may result from several causes:

a) Same contents appearing on different sites are all crawled, processed and indexed.
b) Mistake introduced while parsing these loosely structured and noisy text (HTML page may contain ads. and it is known as shorting of semantics useful for parsing)
c) Manual typos (all information on Internet are created by people originally) and manual revising while being referred and reused
d) Explicit modification to make the short message suitable for difference usage.

Checking can be easily done when the repository of spam mails is small like hundreds or thousands of instances. When the size and the number of instances increasing to millions and more, it becomes impossible for human beings to check them one by one, which is tedious, costly and prone to error. Resorting to computers for such kind of repeatable job is desired, of which the core is an algorithm to measure the difference between any pair of short messages, including duplicated and near duplicated ones.

In Section 2, we define near duplicate and the construction of SP Tree and in section 3 we describe how SimHash works, in section 4 simhash advantages and disadvantages a brief review of conventional work is presented in Section 4, followed by conclusion in Section 5.

II. PRELIMINARIES

A. *Near Duplicate:*

Near-duplicate [1] spam detection is to exploit reported spams and to subsequently block one which has similar content. The definition of similarity between two e-mails is diverse for different forms of email. representing e-mails based mainly on content text, we represent e-mail using an HTML tag sequence, which depicts the layout Structure of e-mail, and look forward to more effectively capturing the near-duplicate phenomenon of spams.

a. Near-Duplicate:

Let I ={t1, t2, ...tn} be the set of valid HTML Tages with two types of newly created HTML tages <mytext/> and <anchor/>. An e-mail abstraction derived as <e1, e2, . . ei; . . . , em>, which is an ordered list of tags, where ei Є I. The definition of near duplicate is: "Two e-mail abstractions A=<a1, a2, . . . ,ai, . . . , an> and B=<b1, b2, . . . , bi, . . . , bm> are viewed as near-duplicate if for all ai= bi and n = m.

B. Related Works:

Since the e-mail spam problem is increasingly serious various techniques have been explored to solve the problem. They can be categorized into the categories [2]:

a. content-based methods
b. non content-based methods.

Content based methods analyze e-mail content text and model this problem as a binary text classification task. Naive Bayes[3] and Support Vector Machines (SVMs) methods comes under this category. Naive Bayes[3] methods train a probability model using classified e-mails, and probability is assigned for each word in e-mails for making a key work as a suspicious spam keyword. SVM [4], is a supervised learning method, which is an efficient and high performed text classification method. Markov random field model [5], logic regression [6] and nueral network [7], and certain specific features, such as URLs and images have also been taken into account for spam detection.

The other group analyzes non content information such as e-mail header, e-mail social network, and e-mail traffic [8] to filter spams.

Collecting notorious and innocent sender IP addresses or email addresses from e-mail header to create blocked list of senders and allowable mail list.

C. Structure Abstraction Generation:

Structure Abstraction Generation [2] generates the e-mail abstraction using HTML content in e-mail. SAG [2] is composed of three major phases, Tag Extraction Phase, Tag Reordering Phase, and <anchor> Appending Phase. In Tag Extraction Phase, the name of each HTML tag is extracted, and tag attributes and attribute values are eliminated. In addition, each paragraph of text without any tag embedded is transformed to <mytext/>.

Since the arrangement of HTML tags are arranged in pairs, various sequential patterns of tags are contained in e-mails. In the worst case, if we consider two e-mail abstractions which have the same tag length and differ only in their last tags, the difference cannot be detected until the last tags are compared. To handle this problem, we destroy the regularity by rearranging the order of tag sequence to lower the number of tag comparisons. Note that this process ensures that the newly assigned position numbers of e-mail abstractions with the same number of tags are completely identical.

In Tag Reordering phase each tag is assigned a new position number (PN denotes for position number) with following expressions,

$b = L^{1/2}$

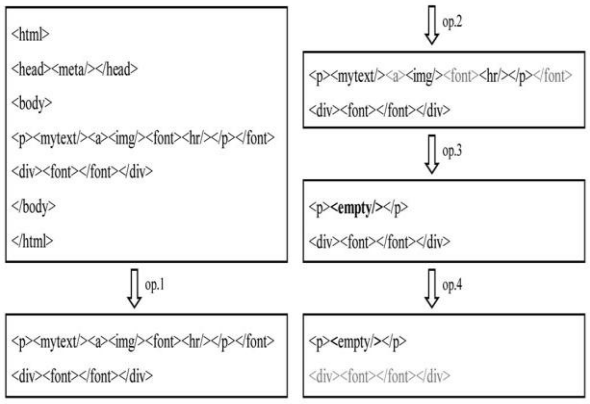
$r = (PN_{orgi}-1) \%b$

$q = (PN_{orgi}-1)/b+1$

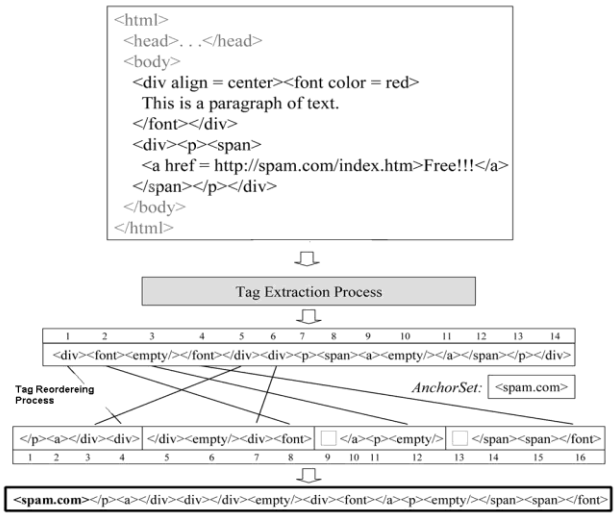
$PN_{new} = (b*r) + (b-q+1)$

Where L is the tag length of an e-mail abstraction, and PN_{orig} is the original position number. Variable b is the number of buckets. Variable r indicates which bucket should be placed and variable q is the number of shift counts from the end of this bucket

An example of the preprocessing step in Tag Extraction Phase of SAG.



Procedure flow of Structure Abstraction Generation



D. Design of Spam Tree:

SP tree [2] is a data structure to facilitate the process of near-duplicate matching. SpTable and SpTrees (sp stands for spam) are proposed to store large amounts of the e-mail abstractions of reported spams. Several SpTrees are the kernel of the database, and the e-mail abstractions of collected spams are maintained in the corresponding SpTrees. According to near duplicate d efinition , two e-mail abstractions are possible to be near-duplicate only when the numbers of their tags are identical.

For efficient matching Sp Trees are designed to be binary trees. The branch direction of each SpTree is determined by a binary hash function. If the first tag of a subsequence is a start tag (e.g.,), this subsequence will be placed into the left child node. A subsequence whose first tag is an end tag (e.g.,) will be placed into the right child node. Since most HTML tags are in pairs and the proposed e-mail abstraction is reordered in SAG, subsequences are expected to be uniformly

distributed. Moreover, on level i of each SpTree (with the root on level 0), each node stores subsequences whose tag lengths are equal to 2_i. For instance, as shown in Fig, the subsequence <spam:com> is placed into level 0, the subsequence </p><a> (whose tag length is 2_1) is placed into level 1, and so forth.

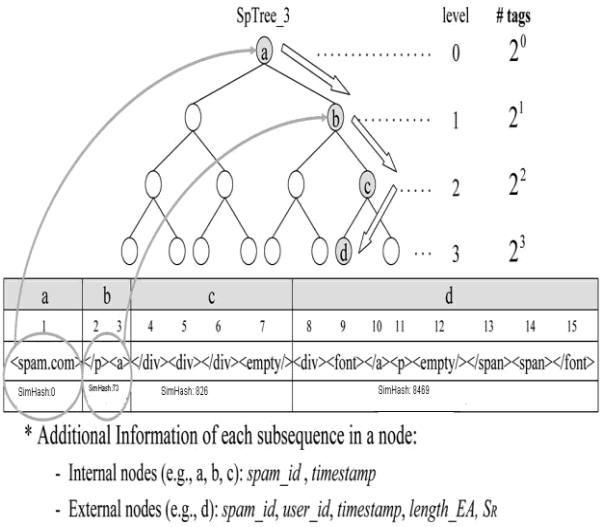


Figure1: Illustration of SP Tree with an example

III. NEAR-DUPLICATE DETECTION BY SIMHASH

Charikar's SimHash[9], actually, is a fingerprinting technique that produces a compact representation of the objects may be documents or images. So, it allows for various processing, once applied to original data sets, to be done on the compact sketches, a much smaller and well formatted (fixed length) space. With documents, SimHash works as follows: a Web document is converted into a set of features, each feature tagged with its weight. Then, we transform such a high dimensional vector into an *f bit* - fingerprint where *f* is quite small compared with the original dimensionality.

The calculation of the hash is performed in the following way:

a. Document is splitted into tokens (words for example) or super-tokens (word tuples)
b. Each token is represented by its hash value; a traditonal hash function is used
c. Weights are associated with tokens
d. A vector V of integers is initialized to 0, length of the vector corresponds to the desired hash size in bits
e. In a cycle for all token's hash values (h), vector V is updated: i^{th} element is decreased by token's weight if the i^{th} bit of the hash h is 0, otherwise i^{th} element is increased by token's weight if the i^{th} bit of the hash h is 1
f. Finally, signs of elements of V correspond to the bits of the final fingerprint

Sample program to show how SimHash works:

```
public class HtmlSimhash {
private static final Logger LOG = Logger.getLogger(
HtmlSimhash.class );
public static void main(String[] args) {
Tap inputTap = new Hfs( new TextDelimited(new
Fields("docid","body"), " " ),args[0] );
Tap outputTap = new StdoutTap();
// create the flow
Flow simhashFlow = Simhash.simhash(inputTap,
outputTap, 1, HtmlText.tokenizer(3));
simhashFlow.complete(); // or add to your Cascade, etc
 }
}
```

In this paper, we show that SimHash is indeed Effective and efficient in detecting both duplicate (with $k = 0$) and near-duplicate (with $k > 0$) (see the two typical examples in TABLE II.) among large short message repository. However, we also notice that due to the born feature of short messages, $k = 3$ may not be an Ideal parameter for. For example, as shown in TABLE III. , $k = 2$ is enough to detect the one-character difference, but k has to be 5 to detect the same pair of messages with two-character difference. Besides, with the same one-character difference, short messages require larger k for effective detection. This may be explained by an observation, that the same difference, e.g. having one different character on the same position of two spam messages, would be more influential to short text than to long text.

This is a paper focusing on discussing Solution for real application.

Firstly, we demonstrate a series of practical values of SimHash-based approach by experiments and our experience.

Secondly, we point out that $k = 3$ may be suitable for near-duplicated spam mail detection, but obviously not suitable for short messages.

Thirdly, we propose one empirical choice, $k = 5$, as applied on our Online short message search.

Table 1. Typical near-duplicates of spam mails with differences highlighted in grey

1. International Monetary Fund congratulate you as our Ten(10) Star Prize Winner in our 2011 End of Year IAP held in London.This makes you a cash prize of £750,000.00 GBP 2. IMF congratulate you as our Ten(10) Star Prize Winner in our 2011 End of Year IAP held in London.This makes you a cash prize of £750,000.00 GBP
1. Pay Rs 1079 for an XXL Bean Bag worth Rs 1800 at Cozy Bean Bags. Sit back & relax! *2.* Pay Rs 1079 for an XXL Bean Bag worth Rs 1800 at Cozy Bean Bags. Sit back, relax?

Table 2. Example: detect duplicate with k =0 and near-duplicate with k >0 (with differences highlighted in gray)

K=0	1.Great Opportunity -- IT Professionals only IIPM LOOKING FOR INDIAN PROFILES 2.Great Opportunity -- IT Professionals only IIPM LOOKING FOR INDIAN PROFILES
K>0	**1. Your e-mail has won you, (£750,000.00.Pounds) from COCA COLA NATIONAL LOTTERY On our 2011 charity bonanza** **2. Your e-mail has won you, ($750,000.00.Dollors) from COCA COLA NATIONAL LOTTERY On our 2011 charity bonanza**

Table 3. Example: detect same long text but more diffrence requires larger *k* (with differences highlighted in gray)

K=2	1.We are Pleased to inform you that you have won a prize money of GBP750,000.00 2.We are Pleased to inform you that you have won a prize money of INR750,000.00
K=5	1) Your e-mail address attached to Winning number 20-12jan-2010-02MSW, serial number S/N-00168, drew the lucky numbers 887-13-866-37-10-83 (2) Your e-mail address attached to Winningnumber20-12DEC-2010-02MSW, serial number S/N-00168, drew the lucky numbers 887-13-865-37-10-83

Table 4. Example: detect same diffrence but shorter text requires larger k (with differences highlighted in gray)

K=2	**1. Your e-mail has won you, (£750,000.00.Pounds) from COCA COLA NATIONAL LOTTERY On our 2011 charity bonanza** **2. Your e-mail has won you, ($750,000.00.Dollors) from COCA COLA NATIONAL LOTTERY On our 2011 charity bonanza**
K=5	1. Great Opportunity --- IT Professionals only IIPM seeing FOR INDIAN PROFILES! 2.Great Opportunity -- IT Professionals only IIPM LOOKING FOR INDIAN PROFILES *

IV. ADVANTAGES AND DISADVANTAGES OF SIMHASH

Sim Hash has several advantages for application based on our experience:

a. Transforming into a standard fingerprint makes it applicable for different media content, no matter text, video or audio;
b. Fingerprinting provides compact representation, which not only reduces the storage space greatly? but allows for quicker comparison and search.
c. Similar content has similar SimHash code, which permits easier distance function to be? Determined for application.
d. It is applicable for both duplicate and near duplicate Detection, with $k = 0$ and $k > 0$ respectively.
e. Similar processing time for different setting of *k* if via the proposed divide-and-search mentioned above, and this is valuable for practice since we are able to detect more near duplicates with no extra cost.
f. The search procedure of similar encoded objects is easily to be implemented in distributed environment based on our implementation experience.
g. From the point of software engineering view, this procedure may be implemented into standard module and be re-used on similar applications, except that the applicants may determine the related parameters themselves.

V. CHALLENGES TO DETECT SPAM E-MAILS

Now a day, spammers are becoming more and more sophisticated. They are finding ways to trick people into thinking that their unsolicited junk messages are worth the time you spend reading them. Some users may understand it as a spam and sends it to spam box but some users consider it as worthy and opens it.

We specify some of the rules for specifying a mail as a spam mail

A. It is placed in Spam Folder:

Sometimes we unknowingly categorize a legitimate email as spam, and emails from certain websites end up in the spam folder. We must deal with issue on a case-by-case basis to determine whether the mail is a legitimate or garbage into your inbox.

B. By seeing Email Address:

Legitimate companies send emails through a server based out of their company website like support@ companyname.com. If we have a long string of numbers in front of the @ sign or the name of a free email service before the .com or any other domain, we need to question the legitimacy of the email.

C. Content of the mail:

Some times mails may be consisting of content which tells us to do something with in a period of time like hours or days and it may consist of links that may be leading us to some other website. Most companies tell you what to do, but they never direct you to where to do it with a link. Mails contains spelling mistakes purposefully have the chance to be a spammer. Spammers don't care enough about the actual messages they're sending to take the time to make them make sense.

D. Spam's ask for personnel Information:

Legitimate institutions never ask for personal information in an email. They don't need to ask you for your personal information anyway because they usually have it on hand. So, if you get an email that asks you for any personal information, no matter how legitimate it might seem, delete it right away. Personal information is only meant to be entered in secure, encrypted forms, not emails where anyone and everyone can get their hands on your information.

E. By Seeing Greeting in the mail:

When you receive a genuine email, the sender addresses you directly, using either your first or last name. If you receive an email where they refer to you as a "Valued Customer" or as a member of some company, its spam. Senders of your genuine emails want to get your attention, so they always address you directly.

VI. CONCLUSION AND FUTURE WORK

Uses an innovative tree structure, SpTrees, to store large amounts of the e-mail abstractions of reported spams. To achieve efficient matching with balanced tree structure, SpTrees are designed to be binary trees. The branch direction of each SpTree is determined by a binary hashfunction.

The improvement is limited since we map each subsequence in a node of an SpTree to a hash value. Therefore, the subsequences that have some prefix tags in common still can be differentiated with one comparison. In this paper, Instead of mapping each subsequence in a node of an SpTree to a hash value using a binary hash function we propose to replace it with a special hash function, namely Simhash.

The advantage of this over other hash functions is that it sets a minimum on the number of members that the two sets must share in order to match. This mitigates the effect of extremely common set members on data clusters.

SimHashbased approach is Fast, Flexible, Customizable (HtmlSimhash), Scalable and is patented.

VII. ACKNOLEDGEMENT

We are greatly delighted to place my most profound appreciation to all my friends and faculty for encouragement and kindness in carrying out the paper. Their pleasure nature, directions, concerns towards us and their readiness to share ideas rejuvenated our efforts towards our goal. We also thank the anonymous references of this paper for their valuable comments.

VIII. REFERENCES

[1]. http://royal.pingdom.com/2011/01/019/email-spam-statistics

[2]. Chi-Yao Tseng, Pin-Chieh Sung, and Ming-Syan Chen "Cosdes: A Collaborative Spam DetectionSystem with a Novel E-Mail Abstraction Scheme" IEEE transactions on knowledge and data engineering, vol. 23, no. 5, may 2011

[3]. V. Metsis, I. Androutsopoulos, and G. Paliouras, "Spam Filtering with Naive Bayes—Which Naive Bayes?" Proc. Third Conf. Email and Anti-Spam (CEAS), 2006.

[4]. E. Blanzieri and A. Bryl, "Evaluation of the Highest Probability SVM Nearest Neighbor Classifier with Variable Relative Error Cost," Proc. Fourth Conf. Email and Anti-Spam (CEAS), 2007.

[5]. S. Chhabra, W.S. Yerazunis, and C. Siefkes, "Spam Filtering Using a Markov Random Field Model with Variable Weighting Schemas," Proc. Fourth IEEE Int'l Conf. Data Mining (ICDM), pp. 347-350, 2004.

[6]. M.-T. Chang, W.-T. Yih, and C. Meek, "Partitioned Logistic Regression for Spam Filtering," Proc. 14th ACM SIGKDD Int'l Conf. Knowledge Discovery and Data mining (KDD), pp. 97-105, 2008.

[7]. A.C. Cosoi, "A False Positive Safe Neural Network; The Followers of the Anatrim Waves," Proc. MIT Spam Conf., 2008.

[8]. R. Clayton, "Email Traffic: A Quantitative Snapshot," Proc. of the Fourth Conf. Email and Anti-Spam (CEAS), 2007

[9]. M. S. Charikar. Similarity estimation techniques from rounding

[10]. Algorithms. In Proc. 34th Annual ACM Symposium on Theory of Computing, pages 380–388. ACM, 2002.