ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 7, July 2017

Classification of Webpages of Public Sector Banks in India Based on Security and Marketability

Sanghamitra Samanta[1], Dr. G. Charles Babu[2]

P.G. Student, Department of Computer Science and Engineering, MallaReddy Engineering College (A), Secunderabad, Telangana, India[1]

Professor, Department of Computer Science and Engineering, Malla Reddy Engineering College (A), Secunderabad, Telangana, India[2]

ABSTRACT: While a considerable amount of attention has been devoted to webpage classification in general, understanding the importance of web content in the homepages of banking websites receive a very little focus from researcher community. With the growing trend in digitization, banking organizations have invested their money and effort for bringing their business into the world of World Wide Web (WWW). Though each bank has their own choice, the designs of their websites are often outsourced to specialized web developers. Therefore, the use of security and reachability related features (like use of online social networks for direct customer engagement) in their homepages is often controlled by third party, and hence a verification of certain expected and mandatory features in the homepages is essential. Machine learning is a powerful tool that is often used to categorize web content in an efficient way. Such a tool can also help to maintain the quality and integrity of webpage's in light of constant changes in their content. In this work, we have studied the security and reach ability features of the homepages of public sector banks in India using well known supervised and unsupervised machine learning models. The aim of this study is to increase the efficiency and effectiveness of web content in banks home pages. We have collected the homepages of all 26 public sector banks at different time periods (one month apart) and investigated the consistency in the use of security based protocols (i.e., like https) and hyperlinks to online social networks (like face book and twitter). Our analysis shows that the use of http based hyperlinks is the only major distinguishing factor for the banking homepages. Also, we have observed that many of the banks are resistant to the use of online social media, and hence, it has significant impact on the classification.

KEYWORDS: Security, Public sector banks, Reachability, homepages.

I. INTRODUCTION

Classification is very important in many information management and retrieval tasks for better efficiency and in avoiding manual effort. In particular, classification of web content helps to perform focused crawling, to assisted development of web content, to topic-based hyperlink analysis, to online advertising based on context, and to analyse the structural properties of the pages, to name a few use cases. Web search is often improved with the help of both query classification as well as content classification. The content of web in today's Internet becomes uncontrolled in the sense that a same piece of information is presented from many different perspectives and in different forms, which imposes additional challenges to web page classification and integrity of web pages. In general, classification of web pages is more challenging compared to that of traditional text classification due to interconnected nature of it which, often, collectively define the context or meaning off the page. Classification can also play an important role to maintain integrity and reliability of web pages, specifically, when it comes to sensitive websites like banking websites in presence of consistent evolution.

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 7, July 2017

Banks have adopted global digital media, i.e., World Wide Web (WWW), since quite some time for publishing their business and make the customers directly access the merits of the products offered to them. In banking vocabulary, a product refers to their offerings like home loan, education loan, savings account, online transaction facilities, and so on. Web has made customers not to worry of visiting the branch offices physically often to know about any such products. While information availability has a been a major issue in the past decades, it is now important to protect and maintain integrity of web content, and keep customers reliability always increasing on their respective banks, especially, in the current era of cyber crime. This has led to the choice of using https protocol (hyper text transfer protocol with transport layer security) over http (hypertext transfer protocol) for web content transfer from server to web browser. In addition, to make the web pages more interactive and responsive, an html (hypertext markup language) page is accompanied by several scripts and images that help the web content to appear dynamically. As business decisions are changing from time to time, web pages, in particular homepages, of the banks tend to update themselves in order to accommodate new policies and remove deprecated ones. Therefore, the home pages are needed to be in constant and automatic monitoring to ensure integrity.

Specifically, it becomes very important for the administrator of the websites to be in confident that the changes made to the web pages are as per their stated policy (possibly internal). Manpower in banking service sector is more focused towards their day-to-day business activities and managing customers which make them less conversant with the peculiarities of the technologies used by them in their business process. For example, which protocol is more reliable to ensure the security and privacy of customer information, or whether to use javascript in an html page is not really a well-focused concern of the banking authority. Hence, most of the banking authorities end up in outsourcing such works to external organizations. However, it is now important for the web administrator in banks to remain in confidence that a modification to the existing system does not violate any integrity or reliability of their system. One of such modifications that can takes place rather often is the change in their homepage as policies mature or new one introduced. But, no tool exists to date to automatically verify the consistency of the web pages.

With the popularity of online social media, like facebook, twitter, etc., the controlling authority of Indian banking community, Reserve Bank of India (RBI), repeatedly advised the banks to make use of such a popular and engaging media to directly involve customers with banking services [4, 5]. Online social media is a powerful platform that can enable customers to exchange their views of a product directly in an easy and efficient way. In fact, current generation is more interested to know about things from such social media and spread information on this platform comfortably and much rapidly. Therefore, in order to catch hold of future customers and to obtain reliable feedback on some certain products even before its actual launch, the use of online media is highly recommended. However, it is very difficult to know for any banking authority whether such an influential media is properly utilized by any of the banks. Therefore, we aim to build an automatic system which can evaluate and rank the banks depending on the utilization of social media and bring the advantages of overall banking system to the current and next generation of customers by already making them a part of the banking ecosystem.

In this paper, we have considered all the public sector banks in India to analyse the security and reachability related features, and find a clustering of their homepages. Our aim is to investigate the degree of alteration in the homepages, and see if they can constantly be identified as the same banks homepage over a period of time. We plan to achieve this goal by using a supervised machine learning framework. Also, we cluster the homepages based on the hyperlinks used in the homepages with the help of an unsupervised machine learning model, which can lead to effective ranking of the banks depending on either security, or use of social media, or both of the features.

The contributions of this study are as follows.

1. A supervised machine learning based framework is designed and implemented to automatically download the homepages of the banks, to extract effective features, to apply random forest (a supervised machine learning algorithm) classifier, and finally to predict the homepages having potential change in place.

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 7, July 2017

2. We extract a total of 21 features related to the use of different types of tags, like script, image, anchor tags, and rank them according to their importance for classifying the web pages. Our analysis shows that the use of http based hyperlinks is the most important (more than 10%) feature for classification. The importance of the hyperlinks to online social networks is about half that of most important one which is significantly higher than half of the features.
3. We perform unsupervised learning on the homepages and our analysis shows that six clusters can suitably accommodate the 26 web pages while using only the http and https based hyperlinks. And, four clusters are suitable while using only the hyperlinks that are pointing to online social networks.
4. Finally, our supervised classification shows that about 96\% accuracy can be achieve using random forest classifier while using all the features. We have observed that the homepage of Bharatiya Mahila Bank got classified as Bank of Baroda which is the only wrong classification, indicating web pages need not be mutually exclusive in their structure.

The paper is organized as follows. Related works are described in Section 2.The system model and the classification framework is described in Section 3. Section 4 talks about the data sets used in our experiments and the performance metrics for analysis along with results of our experiments. Section 5 concludes the paper and points our possible future work.

II. RELATED WORK

The use of social media for the well-being of banking has been one of the major area that Reserve Bank of India has advised since long back to the banks in India. It has prescribed a simple six steps procedure to adopt and take advent age of online social media, starting from assessing a product idea without actually deploying it to measure success rate of a product [4, 5]. The report in [5] has prescribed a set of possible area where social media can be used to potential betterment of banking sector. However, it does not really enforce to implement that by any of the banks, neither it has taken any accountable measure for the degree of usage of social media. Nevertheless, the popular online social media sites that the banks have started to utilize are limited to facebook, twitter, and linkedin. The use of security based protocols like https for hosting web content over Internet has been foremost important especially for banking sector. However, there is no controlling authority that regulates the use of such security protocols. We focus on the design of an automated tool that can learn and classify the homepages of public sector banks in India depending on the usage of social media and the use of security protocol.

Classification of Webpages has been studied since long time focusing on the improvement of search efficiency primarily. The study in [9] has focused on the classification efficiency of the models like support vector machine (SVM), K-Means, Naive Bayes classifiers etc. In particular, it has investigated the impact of having imbalance distribution of the number of signatures to the available classes where a few classes exist with relatively small number of signature sand other have a very high cardinality. The work shows that their method works efficiently for low cardinality classes as well as high cardinality classes. The study in [10] has focused on classifying the webpages into a number of semantic groups automatically. Semantical text processing tools as well as machine learning models have been utilized together to classify a set of webpages. Webpage classification often suffers from a problem of lack of information content in the webpages which lead to the problem of misclassification. The studies have proposed to augment the information for such webpages from additional sources similar in terms of content and syntax.

Many web content in today's Internet are based on simple url links. For example, in online sites like facebook, twitter, etc., users use short comment followed by hyperlinks (i.e., urls) to support their comments. An increasing problem is to classify such webpages which dynamically grow every now and then without having meaningful content (only links are available to certain sites). The study in [1] has suggested a method to classify the urls, without requiring the information present in their associated webpages. Unlike web pages, urls content very concise information with limited phrases, and therefore classification algorithms face challenges to classify them properly. The proposed method can also be applied to a classification problem where text content is very short. The report in [3] has shown the

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 7, July 2017

efficiency of two learning strategy, namely subject oriented and genre oriented learning for web page classification. Subject based classification is also known as topic based classification where a set predefined set of topic (like, complex systems, physics, biology, etc.) are assumed and the webpages are considered to belong to one or more of those subjects.

This type of classification is often used for context based web search. On the other hand, genre based classification is often called as style based classification, where a set of properties of some styles are presumed, like advertisement, online shopping, call for papers, homepage, etc. The webpages are classified to one or more of these types. Genre based classification is relatively a newer approach where a webpage may not belong to a specific or a set of subjects. This paradigm has begun to draw attention as the popularity of Internet for general purpose use like online news paper or online advertising is devised. The report basically examines the effectiveness of different types of classification approach for diverse classification tasks. While a large set of features can represent a webpage with greater information, every single feature may not be so important to achieve better accuracy. Hence, discarding relatively less important features has been one of the important research areas. The study in [2] has focused on such a problem where a set of reduced and important features are extracted for webpage classification. Several existing classification algorithms, like ant colony based classifier and firefly algorithm, have been used with this set of features and it is shown that the optimized set of features produce near similar accuracy as that of considering all of them. When a large number of webpages are used for training a classification model, it often takes a significantly large amount of time. Sometimes a large number of features can also create such problems. Hence, the amount of computing resources used for this purpose becomes a serious concern, particularly, while using certain type of classification models. The study in [8] has formulated the problem of resource utilization for a classification problem as a optimization problem that helps to find an optimized set of features for different types of classifiers. In particular, the study has evaluated the amount of resources used for naïve bayes, support vector machine, and random forest classifiers.

Performance of a classification model significantly depends on the number of positive as well as negative examples almost equally. For example, to classify webpage as a homepage of a website, one needs to have a number of examples of webpages that are actually homepages (positive examples), and a number of examples of webpages that cannot be a homepage of a website (negative examples). Usefulness of negative examples becomes extremely important when it is a problem of binary classification, and obtaining real negative examples for a given class is in general a laborious and hard task. The study in [12] has demonstrated a method of training a classifier without the need of negative examples. It uses positive and unlabeled examples to train a classification algorithm that achieves a significantly good accuracy comparable to traditional classifiers like support vector machine. In our study, we have lack of negative examples as no duplicate homepages of banking websites can be found perhaps due to legal issues. However, in this study we have used traditional classifiers to obtain their performance in our classification problem, and consider the used of specialized models as a future work.

The study in [7] has investigated the purpose of creating hyperlinks in academic webpages. Several universities are considered in order to rank their academic organizations with respect to their research outcomes. Use of hyperlinks has been one of the primary basis that led to extract the features from the webpages. Moreover, the content in the hyperlinked pages are also considered for their evaluation of the performance of the research groups. The study has shown that the hyperlinks alone can provide significant amount of information for their academic achievements. And, this study has influenced our work to consider hyperlinks as one of the major basis for homepage classification. Classification of webpages has many applications in the arena of information retrieval in the context of World Wide Web. Needless to mention that the different types of applications have different requirements which can only be satisfied through a non-unique set of features across the applications. A review of a number of classification algorithms along with their application domain are discussed in [11]. The study has analysed the rationale of selecting different sets of features. We refer interested reader to this paper and the links therein for an in depth understanding of features and their usability in the context of web page classification.

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 7, July 2017

The only study, to the best of our knowledge, in [6] has considered the webpages of banks in India. The study has analysed the problem of a non-functional requirement, i.e., accessibility, of the web pages. The accessibility of a web page refers to the ease of accessing the webpage by any kind of customers irrespective of their physical disabilities. The study has shown that none of the websites (of public and private sector banks) are completely accessible by every kind of customers having some disabilities. In fact, in terms of accessibility, no significant difference could be observed among the web pages. As in one hand it important for a banking website to be accessible by almost every type of customers, it is also important for such a website to be more secure as well as widespread to reach a larger community of users. Our study is based in the second observations that identify and ranks the homepages based on their reliability and reachability.

III. SYSTEM MODEL AND CLASSIFICATION FRAMEWORK

We consider a scenario where a web administrator wants to keep track of the changes made to the web pages (in our case it is the homepage of the web site)of its own organization where the web pages are developed or modified by a group of specialized developers. Administrator need to automate the verification of the web pages depending on the content and the structure of the previous versions of the pages. He also needs to observe the security features adopted in the new or modified page, and approve it to be used in the actual system. Occasionally, administrator may also need to find the ranking of its own webpage depending on certain parameters like use of social media, among the other competitive organizations, i.e., fellow bankers. Hence, it needs to have an account of the publicly available information from those competitors. The administrator would prefer to have a machine learning based automated system to quickly comprehend or evaluate it entire web site on a page by page basis or as a whole. As the problems are different in nature, one supervised model can be employed to identify the pages leading to maintaining the integrity and one unsupervised model can be used to group the pages with an aim of getting a ranking of the page among other similar and competitive pages. The unsupervised model is capable of considering features of user choice to obtain a clustering depending on the purpose, like security or reachability.

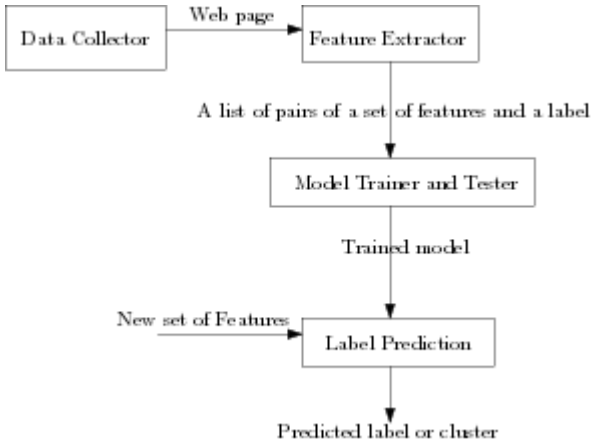


Fig. 1: Classification framework.

Our proposed classification framework is shown in Figure 1. The framework has four modules, data collector, feature extractor, model trained and tester, and label/group predictor. Data collector module collects the homepages of the other banks including itself automatically through appropriate interface, and provide each of those pages along with a label (in this case, bank name) to the feature extractor module. Feature extractor module extracts (possibly using some html parser) a set parameters related to the structure of the page, and then extracts a set of features from

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 7, July 2017

those parameters using some simple statistical methods. It stores each of those features and the associated label in a stable storage, like light-weight database system. The model trainer and tester module then takes the set of features and the labels from the data base and train a supervised model, which is in our case random forest classifier. This module also train an unsupervised model which in our case k-means clustering and cluster the features into group and then assign a label to the groups. Note that, unlike supervised learning, the group labels in this case are algorithm defined instead of user defined. The model is tested through various sampling of the entire set of features with/without labels. The importance of the features is extracted from the model training phase, which is then used for features selection. Finally, the trained model is provided to label predictor module. The label predictor module takes a set of features extracted from a new web page and the label/cluster is predicted for this set of features.

Data Collector: This module contains a number of python-based web crawling tools. This module takes a list of banks whose home pages are of interest to the administrator. This module make legitimate request to the respective web server for each of the banks and obtain the home pages from them. For example, this module needs the URL, https://www.pnbindia.in/, to download the home page of Punjab National Bank. The Python url lib package is used to open a communication link with the URL. The link can then be used to download the raw content of the web site using request package.

Feature Extractor: Before explaining the features extracted from a html file, we briefly discuss the format of a html file used to build a typical homepage of a bank. Figure 2 shows a couple of samples from an arbitrarily selected homepage (namely, pnbindia).As shown in the figure, the file starts with a script type tag followed by some meta information and linking some style sheets and external images. In general, we have observed that every homepage uses a lot of javascripts (using script tags) to provide better responsiveness to the web users. Some of the banks use a number of http or https hyperlinks to some external web services, as well as to connect to the some social networking sites in order to provide better visibility of their products and services. We use these tags at a very high level to extract certain features from each of the homepages. In general, we use two types of features, count and ratio, on the number of tags in a web page. Altogether 21 features have been extracted as described below.

1. N (t∗) - the number of tags (of any type) found in a homepage.
2. N (ta) - the number of tags of type anchor, a, that are used for provide hyperlinks, i.e., < a > ... < /a >.
3. R (ta , t∗) - the ratio between the number of anchor tags and the number of any type of tags, i.e.,
   R (ta , t∗) = N (ta)/N (t∗)
4. N (ts) - the number of tags of type script, i.e., < script > .. < /script >,that are used to provide better responsiveness.
5. R (ts , t∗) - the ratio between the number of script tags and the number of any type of tags, i.e.,
   R (ts , t∗) = N (ts)/N (t∗)
6. N (ti) - the number of tags of type image, i.e., < img > .. < /img >, to better explain the concepts described using texts. Note that use of more images in a webpage would essentially make the page slow to download.
7. R (ti , t∗) - the ratio between the number of image tags and the number of any type of tags, i.e.,
   R (ti , t∗) = N (ti)/N (t∗)
8. N (thp) - the number of anchor tags that provides hyperlinks to http enabled ages, i.e., < a...href = “http : //....00 >. In general, such hyperlinks are used to point to some sites that may not be security sensitive. For example, a http based hyperlink is used to connect to www.ignou.ac.in, as shown in Figure 2, which may not be as security sensitive as a banks homepage.
9. R (thp, ta) - the ratio between the number of http based hyperlinks and the total number of anchor tags, i.e.,
   R (thp, ta) = N (thp)/N (ta)
10. N (ths) - the number of anchor tags that use https based hyperlinks, i.e.,< a...href = “https : //....00 ... >. Such links are used to provide hyperlinks to the sites are relatively highly security sensitive. For example, a https based hyperlink is used to connect to gateway.netpnb.com as shown in Figure 2which can lead to access some secure sensitive information.
11. R (ths , ta) - the ratio between the number of https based hyperlinks and the total number of anchor tags, i.e.,
    R (ths, ta) = N (ths)/N (ta)

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 7, July 2017

12. N (to) - the number of anchor tags that neither use http based hyperlinks or https based hyperlinks, i.e., < ahref = "javascript : void(0)"... >. In general, such tags are used to connect to some external javascript methods. For example, in pnbindia, such tags are used to hyperlink < ahref ="javascript : void(0)"... > type of javascripts.
13. R (to, ta) the- ratio between the number of tags of type non-http and non-https and the number of anchor tags, i.e., R (to, ta) = N (to)/N (ta)

```
<!DOCTYPE html>
<html lang="En">
<head id="Head1">

<script>
  (function(i,s,o,g,r,a,m){i['GoogleAnalyticsObject']=r;i[r]=i[r]||function(){
  (i[r].q=i[r].q||[]).push(arguments)},i[r].l=1*new Date();a=s.createElement(o),
  m=s.getElementsByTagName(o)[0];a.async=1;a.src=g;m.parentNode.insertBefore(a,m)
  })(window,document,'script','https://www.google-analytics.com/
analytics.js','ga');

  ga('create', 'UA-84964116-1', 'auto');
  ga('send', 'pageview');

</script>

    <meta http-equiv="Content-Type" content="text/html; charset=utf-8" /><title>
        Punjab National Bank
</title><meta name="viewport" content="width=device-width, initial-scale=1,
maximum-scale=1, user-scalable=no" /><link rel="stylesheet" href="css/style.css?
i=3" type="text/css" media="screen" /><link rel="stylesheet" href="css/
vertical_menu.css" type="text/css" media="screen" /><link rel="stylesheet"
href="css/font-awesome.min.css" type="text/css" media="screen" />
    <!--css & js for banner slider Start-->
    <!--Favicon Code start here-->
    <link rel="shortcut icon" href="images/favicon.ico" type="image/x-icon" /
><link rel="icon" href="images/favicon.ico" type="image/x-icon" />
    <!--End here-->
    <link rel="stylesheet" href="css/tab-style.css" type="text/css"
media="screen" />
    <!--[if lte IE 8]>
   <link rel="stylesheet" href="css/belowIE.css" type="text/css" />

<![endif]-->
    <script type="text/javascript" src="js/1.8.2_jquery.min.js"></script>
    <script type="text/javascript" src="js/customfunction.js"></script>
        <p><a target="_blank" href="https://www.netpnb.com/web/L001/webpages/
annamalai.html" title="Annamalai university"><strong class="bullet"></strong>
<span>Annamalai university </span></a></p>
        <p><a target="_blank" href="http://www.ignou.ac.in/" title="IGNOU"><strong
class="bullet"></strong> <span> IGNOU</span> </a></p>
        <p><a target="_blank" href="https://gateway.netpnb.com/AGDAV/home.html"
title="AGDAV Delhi"><strong class="bullet"></strong> <span>AGDAV Delhi</span></a>
```

Fig. 2: Sample of a homepage html file, (A) the beginning of the file, and (B) arbitrarily selected portion from the middle of the file.

14. N (tr) - the number of hyperlinks to connect to RBI (i.e., Reserve Bank of India) website. Such links are used to obtain information related, for example (https://www.rbi.org.in/currency/Security%20Features.html), security related features in currency notes in India, in general.
15. R (tr , ta) - the ratio between the number of hyperlinks to RBI site and the number of anchor tags, i.e.,
R (tr , ta) = N (tr)/N (ta)
16. N (tf) - the number of hyperlinks to facebook site. Usually, such links are used to obtain some understand customers sentiment towards the banks on various products, like mobile banking app, internet banking options, etc. Also, decision makers of the bank get to know about the customer satisfaction label on certain products and take appropriate decision to improve or launch a new product.
17. R (tf , ta) - the ratio between the number of hyperlinks to facebook and the number of anchor tags, i.e.,
R (tf , ta) = N (tf)/N (ta)
18. N (tt) - the number of hyperlinks to twitter site. These hyperlinks are used to obtain recent trend or sentiment on some certain products of a bank, rather than historical events which may be better obtained through facebook networks. And, this might drive some quick business decision and assess upcoming mood of customers on their recently launched products.

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 7, July 2017

19. R (tt , ta) - the ratio between the number of hyperlinks to twitter network and the number of anchor tags, i.e.,
R (tt, ta) = N (tt)/N (ta)
20. N (tl) - the number of hyperlinks to linkedin social network site. In many cases, such links are used to advertise some recruitment related information to be shared among prospected candidates. Considering linkedin as one of the social networking site for classified users, banking community sometimes use these hyperlinks to attract certain bonafied customers both to promote and create new customers for their products.
21. R (tl, ta) - the ratio between the number of hyperlinks to linkedin networking site and the number of anchor tags, i.e.,
R (tl , ta) = N (tl)/N (ta)
Thus, we consider a good mixture of features that can account for security aspects, reachability aspects, and webpage performance aspects at the same time. We plan to extract a few more features related to business performance and vulnerability of the web pages in future.

Web Page Annotator: In this work, we assume that every single bank adopts new business initiatives and policies on a monthly basis or a quarterly basis, and hence they update their homepages on a rather regular basis to reflect the recent changes. They can even change their base website names, for example, when they grow beyond a manageable size, or sometimes a number of banks get merged to profit from each others business spread. For example, recently, State Bank of India has acquired a number of individually functioning banks to take benefits of larger business capital, and hence the bank has considerably changed its homepage to reflect such a big event. In order to maintain uniformity, we assume a consistent label to each of the homepages which is used for training and testing purpose. Table 1 show the list of public sector banks in India with an identifier for each bank, whether using http/ https, and the labels assigned to them. For example, Punjab National Bank uses https to host their homepage and the page is labelled as pnbindia.

Table 1: Homepages, bank names and the labels assigned to them.

B. ID	Bank Name	http/https	Label
I	Punjab National Bank	https	pnbindia
II	Dena Bank	http	denabank
III	Allahabad Bank	https	allahabadbank
IV	Andhra Bank	https	andhrabank
V	Bank of India	http	bankofindia
VI	Bank of Baroda	http	bankofbaroda
VII	Bank of Maharastra	http	bankofmaharastra
VIII	Canara Bank	http	canarabank
IX	Central Bank of India	https	centralbankofindia
X	Corporation Bank	http	corpbank
XI	Indian Bank	http	indianbank
XII	Indian Overseas Bank	https	iob
XIII	Oriental Bank of Commerce	https	obcindia
XIV	Punjab and Sindh Bank	https	psbindia
XV	Syndicate Bank	https	syndicatebank
XVI	UCO bank	https	ucobank
XVII	Union Bank of India	http	unionbankofindia
XVIII	United Bank of India	http	unitedbankofindia
XIX	Vijaya Bank	https	vijayabank
XX	Bharatiya Mahila Bank	http	bmb
XXI	IDBI Bank	https	idbi
XXII	State Bank of Mysore	http	statebankofmysore
XXIII	State Bank of Patiala	https	sbp
XXIV	State Bank of Travancore	https	statebankoftravancore
XXV	State Bank of India	https	sbi
XXVI	Reserve Bank of Inia	https	rbi

Model Trainer: Model trainer is implemented using scy-kit learn package that implements the machine learning algorithms. This python based package is a popular one and used in a wide variety of problems across different disciplines. We use only two models, random forest classifier which is part of ensemble methods in the package, and

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 7, July 2017

kmeans algorithm which is part of cluster methods of the package. We briefly discuss the algorithms here for an idea of the working of the algorithms. K-means algorithm is used to cluster data by trying to separate samples in k groups of equal variance, minimizing a criterion known as the inertia or within-cluster sum-of-squares. This algorithm requires the number of clusters to be specified. It scales well to large number of samples and has been used across a large range of application areas in many different fields for unsupervised grouping or clustering. Specifically, the k-means algorithm divides a set X of n samples into a set C of k disjoint clusters. Each cluster cj is described by what is called centre, μj , which is a mean of the samples in that cluster. The centres are commonly called as the cluster centroids. The centroids need not be data points from actual data set X, although they belong to the same hypothetical plane. The K-means algorithm aims to choose centroids that minimize the inertia or within-cluster sum of squared. Inertia or the within-cluster sum of squares criterion is considered as a measure of internal coherentness of the clusters.

Random forest algorithm makes decision trees susceptible to high variance if they are not pruned. This high variance is reduced by creating multiple trees with different samples of the training data set and combining their predictions. A good random forest is the one that generalizes well and have higher accuracy by each tree, and higher diversity among the trees. The importance of each individual features is computed by what is called feature contribution which is an absolute impact on the classification. As the existing studies have shown that this classifier is suitable for large data set as well as large set of features, we consider in this study for supervised classification.

IV. EXPERIMENTAL RESULTS

In this section, we describe the experimental setup and the data set used in our experiments followed by results. We have collected three sets of data where each and every set contains homepages of the websites of 26 public sector banks in India including Reserve Bank of India (RBI), and they are collected at different time periods roughly one month apart. A summary of the data sets is shown in Table 2.Both Set I and Set II contains 26 home pages of 26 public sector banks, but Set III contains 24 home pages. The home pages in Set III is collected on 28th April when State Bank of India merges its associated banks into one bank from 1st April, 2017, and therefore websites of three associated banks of State Bank of India do not exist, namely, State bank of Mysore, State Bank of Patiyala, and State Bank of Travancore. During the collection of Set I and Set II, IDBI bank homepage could not be collected using a program agent as the corresponding server blocked the download request, possibly assuming abotnet attack. However, during the collection of Set III, the same homepage could be downloaded with same program agent. But, the homepage of RBI could not be downloaded using a simple program agent. In both the cases, we have manipulated the download requests by altering the request header to make it look like a browser agent, rather than a program agent, and hence both the homepages were downloaded without any further problem. Set I and Set II are considered as training set and Set III is considered as testing set. So, training set has got 54 signatures, and testing set got 24 signatures. We understand that the number of signatures in our experiments is limited during the course of this work (we plan to collect have better experiments with larger signatures by collecting the homepages again and again over one year period of time).Nonetheless, interesting to find that the homepages do show changes over time in both their content and links to outside pages, like adding links to social network sites and other government sites. Thus, we consider that this set of data can have some important variations already which form the basis of our preliminary investigation.

Table 2: Datasets description.

Set	Collection Time	#Banks	Avg. Size (KB)	Total Size (in MB)	Avg. Tags
I	17^{th} Feb. 2017	27	100	3.0	827
II	25^{th} Mar. 2017	27	90	2.4	853
III	28^{th} Apr. 2017	24	95	2.7	912

Classification problems are perhaps the most common type of machine learning problem and as such there are a myriad of metrics that can be used to evaluate a prediction process used for these problems. In this work, we use one simple metric that measures an overall accuracy of the classification performance.

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

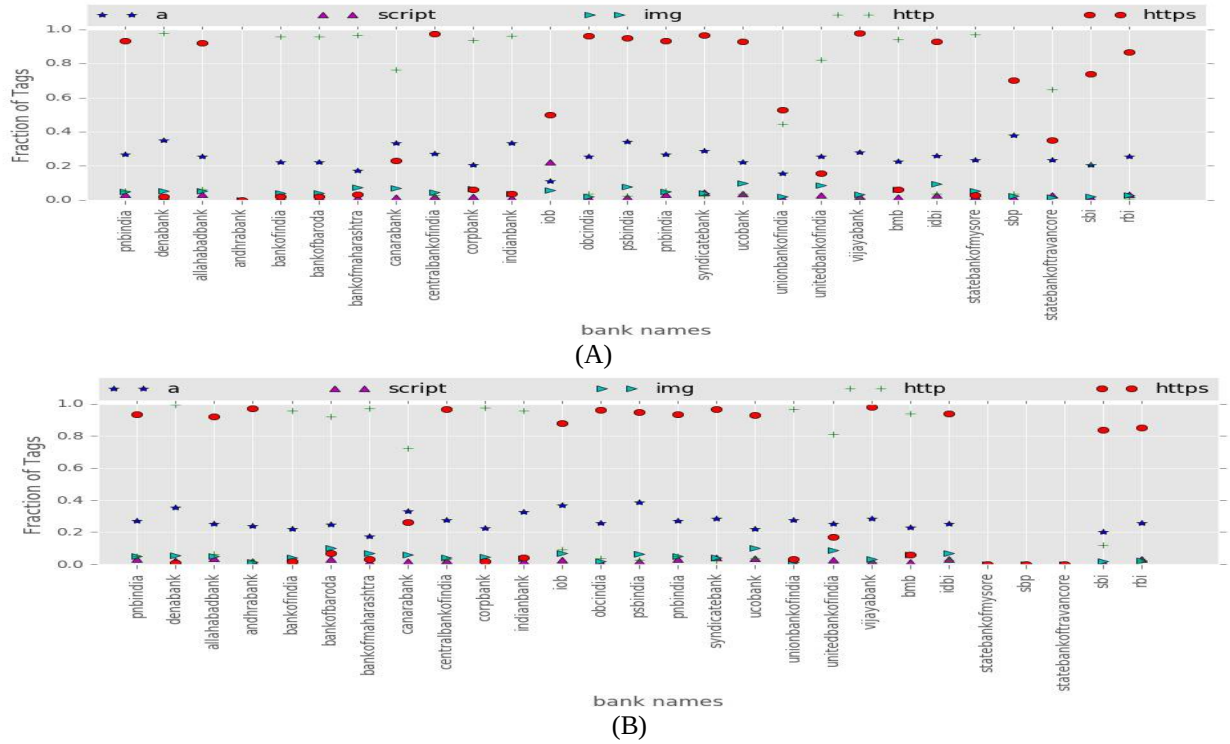
Vol. 6, Issue 7, July 2017

1. Accuracy – The ratio of the number of correct predictions to the total number of predictions.

This metrics can be derived from the confusion matrix of any classification model. We shall report the confusion matrix.

Before going to our results on classification, we at first present the variation of different properties of the homepages used here for experiments. Figure 3 shows the variation of the fraction of tags that are anchor tags, script tags, image tags, http and https hyperlinks in the three data sets. The figure shows that 18-39% tags are anchor tags in any of the data sets. A maximum of10% tags are image tags in all of the homepages, and less than 5% tags are used for scripts. Interesting to observe that there is a increasing trend of using https hyperlinks as compared to using http links. Out of 26 homepages, 11 homepages use more (> 70%) http links in data Set I. And, 3 banks have increased the use of https tags to more than 60% of all hyperlinks, and 2 banks have reduced the use of http links to less than 30% of total links. About 10 banks have less than10% of links with https which can be alarming in the era of digital hacking and exploits.

One can observe that the use of http and https based hyperlinks can be a potential indicator for clustering the web pages. Because we have assumed that the use of https is a potential indicator of a better security aspects of the web pages as compared to http links, the use of https links can be used for learning purpose to rank the pages and can be considered as a ranking based on security features. However, the use of script tags may not be a good parameter for classification in any case as they have almost same amount (about 5%) of those tags in all the homepages.



(A)

(B)

Fig. 3: Variation of fraction of tags that are anchor tags, script tags, image tags, http hyperlinks, and https hyperlinks in the three data sets in subplots (A), and (B) respectively.

Considering the variations of individual features, we proceed to perform the classification of the homepages using a supervised learning mechanism. As mentioned before, we consider only Random Forest classifier for our classification as it has been shown to produce good accuracy in various other fields of study (see Related Work section, Section 7). We train our model using two data sets Set I and Set II, and the data Set III is considered as test set. Recall that our aims to predict the label (i.e., bank name) of a homepage in presence of consistent changes to them. Results of our

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 7, July 2017

analysis are presented in terms of confusion matrix, as shown in Table 3. A total of 26 signatures of homepages obtained from data set III which is used to evaluate the performance. Results of 20 predictions are presented in the Table 3. It has been observed that the predication accuracy is 25*100/26 =96%. However, in some cases, the similarity of a page with other homepages is quite higher (about 60% similar). The homepage of Bharatiya Mahila Bank got predicted as the homepage of Bank of Boroda, which is the only wrong prediction. Nevertheless, we have seen that in majority of the cases, the homepages has not been changed significantly, and this has a huge impact on the prediction accuracy being so higher.

Table 3: Confusion matrix of random forest classifier. Columns and rows indicate the number of the actual labels and the predicted labels respectively. B. IDs are the bank identifiers as shown in Table 1.

B. ID	I	II	III	IV	V	VI	VII	VIII	IX	X	XI	XII	XIII	XIV	XV	XVI	XVII	XVIII	XIX	XX
I	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
II	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
III	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
IV	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
V	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
VI	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
VII	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
VIII	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
IX	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
X	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
XI	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
XII	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
XIII	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
XIV	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0
XV	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
XVI	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1
XVII	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
XVIII	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
XIX	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
XX	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1

We have further investigated the importance factor of the features for classification. Table 4 presents the importance values. The importance values are obtained from random forest classifier while performing training phase. One can see that the feature of number of http based hyperlinks (i.e., N (thp)) is the most important distinguishing factor for the homepages, contributing above 11%. The contribution of majority of the features lies between 5% to 7%, accounting to eleven features and remaining features have less than 4% contribution to the classification, for example, the ratio of hyperlinks of type “others” to all hyperlinks contributes the least. Interesting to observe that the hyperlinks to social networking sites have significant importance for the classification in our case this indicates that there exist some banks that consider the use of social networking sites as one of their policy to enhance their business, e.g., to engage online community to obtain their opinion on mobile banking or internet banking. Both the number and the ratio of https based hyperlinks have been seen to have average importance (about 5.5%). This indicates that almost a fixed number of https based hyperlinks are used by almost all the banks and such hyperlinks provide similar services to all the banks. A closer inspection on the html files shows that every bank uses such type of hyperlinks when it comes to provide online banking services for example.

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

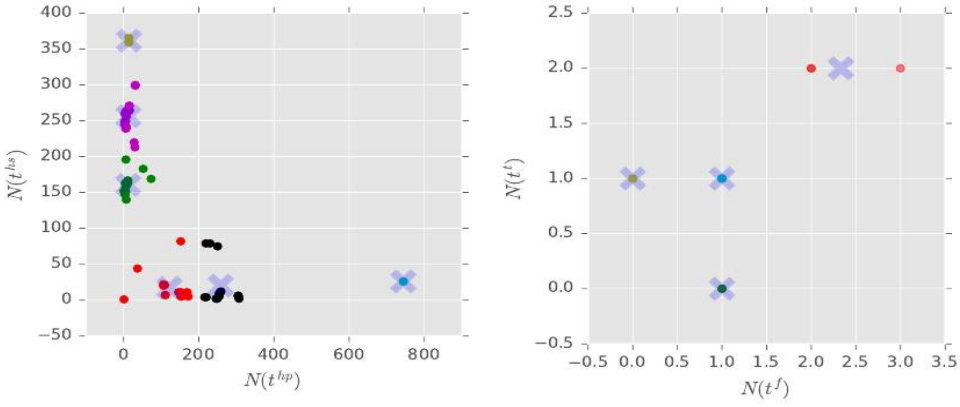
Vol. 6, Issue 7, July 2017

Table 4: Features importance using Random Forest classifier.

Feature	$N(t^*)$	$N(t^a)$	$R(t^a, t^*)$	$N(t^s)$	$R(t^s, t^*)$	$N(t^i)$	$R(t^i, t^*)$	$N(t^{hp})$
Importance	0.055	0.031	0.060	0.061	0.064	0.066	0.070	0.114
Feature	$R(t^{hp}, t^a)$	$N(t^{hs})$	$R(t^{hs}, t^a)$	$N(t^o)$	$R(t^o, t^a)$	$N(t^r)$	$R(t^r, t^a)$	$N(t^f)$
Importance	0.053	0.055	0.054	0.038	0.027	0.037	0.055	0.028
Feature	$R(t^f, t^a)$	$N(t^t)$	$R(t^t, t^a)$					
Importance	0.051	0.037	0.032					

We finally performed unsupervised learning on all the three data sets. In this case, only the features are used without their labels. The aim of this clustering is to rank the homepages depending on their a) security properties, i.e., the usage of https based hyperlinks, and b) the reachability properties, i.e., the use of hyperlinks to facebook and twitter online social networks. We have used K-means clustering with various values of k (indicating the number of user defined clusters). Clustering is performed on every pair of features, but we focus on the hyperlink based features. Two pairs are mainly considered, first, R (thp , ta) and R (ths, ta), and then R (tf , ta) and R (tt , ta). Note that all the features have been used in supervised learning case, but in this case only a subset of features is used.

Figure 4 shows the clusters of homepages, (A) using http and https based features, (B) using facebook and twitter based features. In both the cases, we performed clustering with 2 to 6 predefined clusters. In http and https based clustering, we have observed n = 6 produces best clustering compared to other values of n. Sum of http and https based hyperlinks roughly equals to the total number of hyperlinks, excluding a few hyperlinks for javascripts using neither http nor https hyperlinks. One of six groups has got a very few homepages that used a high number of http based hyperlinks compared to a very small number of https based hyperlinks, and this is the homepage of Union Bank of India. Three groups of homepages uses a very high fraction of https based links compared to http based links. And, the other two groups are seen to have a balanced combination of them. Banks like Central Bank of India, Oriental Bank of Commerce, Syndicate Bank are the ones that use more than 95% of their hyperlinks to be https based ones. Banks like Dena Bank, Bank of India, Bank of Baroda, Bank of Maharashtra, Indian Bank, State Bank of Mysore, and Bharatiya Mahila Bank are the one that use more than 95% of the hyperlinks with http, i.e., without using security protocols, in their homepages.



(A) (B)

Fig. 4: Clustering of homepages with (A) http and https based hyperlinks, and (B) hyperlinks to facebook and twitter. Cross symbols in both the plots indicate the centre of the corresponding clusters.

Finally, we have investigated the clustering using facebook and twitter hyperlinks based features. We see that n = 4 produces best clustering in this case. In general, 1 to 3 hyperlinks have been used in a homepage to connect to facebook network. State Bank of India have used the highest number of such hyperlinks (=3). About 50% of the banks have not

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 7, July 2017

used any such hyperlinks to facebook network in their homepages, for example, Dena Bank, Allahabad Bank and Andhra Bank. Only 9 of the banks have used hyperlinks to twitter network, with State Bank of India and Punjab National Bank using highest number of such links (=2). All other banks have used only one such hyperlink to connect to this social networking site. Thus, clearly showing 4 clusters are sufficient to accommodate the homepages that have used social networking sites. The Bank shaving no hyperlinks to any of these social networks are excluded from consideration here (i.e., about 13 of the banks). Our investigation shows that almost all the banks are under-utilizing the facilities of online social networks. At the same time, some of the banks are not sufficiently using security based protocols, essentially indicating less awareness of secure access of banking information among the banks community.

V. CONCLUSION

In this study, we have considered the problem of classifying homepages of public sector banks in India depending on their security and socially reachable features. We have proposed a simple classification framework that extracts a set of important features, uses a supervised and a unsupervised classifier, and obtain a good clustering of the homepages. We have obtained an important set of features that can optimize the resource utilization. In particular, we have considered homepages of 26 public sector banks, collected at different period of time, and extracted features based on the hyperlinks and a number of html tags. Our aim is to observe the change in the homepages of each of the banks as business needs evolve over time, and their impact on the classification. We have used random forest classifier, a supervised learning algorithm, and k-means clustering, an unsupervised learning algorithm in this work. Our results show that the changes in the homepages are not very significant, i.e., the change in a homepage did not make it to be classified as a different homepage in majority of the cases. However, the features that have very high importance is very limited in number. For example, the feature that counts the number of http based hyperlinks in a homepages is the only features that have a very high importance (greater than 10%), while the use of script tags are much less importance. Interestingly, we have observed that the use of hyperlinks to online social media sites like facebook and twitter have a significant contribution (about 4-5%) to the classification, the importance is similar to that of the number of https based hyperlinks. Our experiment on using unsupervised learning algorithm shows that there exist six clusters of homepages when we consider only security based features i.e. the use of http and https based hyperlinks. And, there are four clusters when we consider the reachability related features extracted from the use of hyperlinks to online social media. We plan to further investigate the impact of using http and https based hyperlinks in the bank homepages on their vulnerabilities in a future work.

REFERENCES

[1] T. A. Abdallah and B. de la Iglesia. Url-based web page classification: Withn-gram language models. volume 553 of Communications in Computer and Information Science, pages 19–33. Springer, 2014.
[2] K. Bhatt, A. Singh, and D. Singh. An improved optimized web page classification using firefly algorithm with nb classifier (wpcnb). International Journal of Computer Applications, 146(4):15–21, Jul 2016.
[3] B. Choi and Z. Yao. 5 web page classification. Louisiana Tech University, Ruston, LA 71272, 2008.
[4] C. for Advanced Financial Research and Learning. Uses of social media by banks. http://www.cafral.org.in/sfControl/content/LearningTakeaWays/1216201531637PM-CAFRAL-LT-Uses-of-Social-Media-by-Banks.pdf.
[5] Social media framework forindian banking sector. http://www.idrbt.ac.in/assets/publications/BestPractices/Social Media Framework (2013)
[6] A. Kaur and D. Dani. Banking websites in india: an accessibility evaluation. CSITransactions on ICT, 2(1):23–34, 2014.
[7] P. Kenekayoro and M. Buckley, Kevanand Thelwall. Automatic classification of academic web page types. Scientometrics, 101(2):1015–1026, 2014.
[8] I. Mahadevan, S. Karuppasamy, and R. Ramasamy. Resource optimization inautomatic web page classification using integrated feature selection and machinelearning. Int. Arab J. e-Technol., 1(1):19–28, 2009.
[9] S. T. Marath, M. Shepherd, E. Milios, and J. Duffy. Large-scale web pageclassification. In Hawaii International Conference on System Sciences, pages1813–1822, Jan 2014.
[10] J. Materna. Automatic Web Page Classification. In Workshop on RecentAdvances in Slavonic Natural Language Processing, pages 84–93, 2008.
[11] X. Qi and B. D. Davison. Web page classification: Features and algorithms.ACM Comput. Surv., 41(2):12:1–12:31, feb 2009.
[12] H. Yu, J. Han, and K. C. C. Chang. Pebl: Web page classification withoutnegative examples. IEEE Transactions on Knowledge and Data Engineering, 16(1):70–81, Jan 2004.