Natarajan Meghanathan
Dhinaharan Nagamalai (Eds)

Computer Science & Information Technology

International Conference on Computer Science & Engineering
(CSEN - 2014)
Bangalore, India, October 25 ~ 26 - 2014

AIRCC

Volume Editors

Natarajan Meghanathan,
Jackson State University, USA
E-mail: nmeghanathan@jsums.edu

Dhinaharan Nagamalai,
Wireilla Net Solutions PTY LTD,
Sydney, Australia
E-mail: dhinthia@yahoo.com

ISSN: 2231 - 5403
ISBN: 978-1-921987-16-8
DOI : 10.5121/csit.2014.41001 - 10.5121/csit.2014.41006

Typesetting: Camera-ready by author, data conversion by NnN Net Solutions Private Ltd., Chennai, India

Preface

The International Conference on Computer Science & Engineering (CSEN-2014) was held in Bangalore, India, during October 25~26, 2014. The International Conference on Advanced Computing (ADCO-2014) collocated with the CSEN-2014. The conferences attracted many local and international delegates, presenting a balanced mixture of intellect from the East and from the West.

The goal of this conference series is to bring together researchers and practitioners from academia and industry to focus on understanding computer science and information technology and to establish new collaborations in these areas. Authors are invited to contribute to the conference by submitting articles that illustrate research results, projects, survey work and industrial experiences describing significant advances in all areas of computer science and information technology.

The CSEN-2014, ADCO-2014 Committees rigorously invited submissions for many months from researchers, scientists, engineers, students and practitioners related to the relevant themes and tracks of the workshop. This effort guaranteed submissions from an unparalleled number of internationally recognized top-level researchers. All the submissions underwent a strenuous peer review process which comprised expert reviewers. These reviewers were selected from a talented pool of Technical Committee members and external reviewers on the basis of their expertise. The papers were then reviewed based on their contributions, technical content, originality and clarity. The entire process, which includes the submission, review and acceptance processes, was done electronically. All these efforts undertaken by the Organizing and Technical Committees led to an exciting, rich and a high quality technical conference program, which featured high-impact presentations for all attendees to enjoy, appreciate and expand their expertise in the latest developments in computer network and communications research.

In closing, CSEN-2014, ADCO-2014 brought together researchers, scientists, engineers, students and practitioners to exchange and share their experiences, new ideas and research results in all aspects of the main workshop themes and tracks, and to discuss the practical challenges encountered and the solutions adopted. The book is organized as a collection of papers from the CSEN-2014, ADCO-2014.

We would like to thank the General and Program Chairs, organization staff, the members of the Technical Program Committees and external reviewers for their excellent and tireless work. We sincerely wish that all attendees benefited scientifically from the conference and wish them every success in their research. It is the humble wish of the conference organizers that the professional dialogue among the researchers, scientists, engineers, students and educators continues beyond the event and that the friendships and collaborations forged will linger and prosper for many years to come.

Natarajan Meghanathan
Dhinaharan Nagamalai

Organization

Program Committee Members

Ahmed Rashed	University of Menoufia, Egypt
Aiguo Song	Southeast University, P.R.China
Akhil Garg	Nanyang Technological University,Singapore
Ali Sahafi	University of Tabriz, Iran
Alireza Souri	Islamic Azad University, Iran
Allaoua Refoufi	University of Setif, Algeria
Bal Virdee	London Metropolitan University, UK
Ching-Nung Yang	National Dong Hwa University, Taiwan
Chunhua Wang	Hunan University, China
Cristina Alcaraz	University of Malaga, Malaga
Daniel Gilgen	Trier University of applied Sciences, Germany
David Naccache	Ecole Normale Superieure, France
Dayang Hjh Tiawa Awg Hj Hamid	Universiti Teknologi Malaysia, Malaysia
Debabrata Samanta	Acharya Institute of Technology, India
Evelyn R. Sowells	North Carolina A&T State University, USA
Fan Hong	University of Limerick, Ireland
Farhad Shahnia	Curtin University, Australia
Farhan Khan	University of Kent, United Kingdom
Giancarlo	University of Roma, Italy
Hossein Jadidoleslami	MUT University, Iran
Huashan Liu	Donghua University, China
Himanshu Aggarwal	Punjabi University, India
Huiyu Zhou	Queen's University Belfast, United Kingdom
Hyung-Woo Lee	Hanshin University, Rep of Korea
I-Cheng Chang	National Dong Hwa University, Taiwan
Ioannis Dassios	The University of Edinburgh, UK
Isa Maleki	Islamic Azad University, Iran
Juntao Fei	Hohai University, China
Kaamran Raahemifar	Ryerson University, Canada
Kannan Govindarajan	MIMOS Berhad, Malaysia
Khaled Merit	Mascara University, Algeria
Manjunath A E	R V College of Engineering, India
Mehdi Nasr	Islamic Azad Univerity,Iran
Mingqian Liu	Xidian University, China
Mohammad Al_Selam	University of Technology, Iraq
Mohammad Rafiqul Haider	University of Alabama, United States
Mohammed Amin	Higher Colleges of Technology,UAE
N.K. Srinat	Visvesvaraya Technological University,India
Natarajan Meghanathan	Jackson State University,USA
Nirmalya Kar	National Institute of Technology,India
Norbert Herencsar	Brno University of Technology, Czech Republic
Sandeep Bhongade	Rajiv Gandhi Technical University, India
Smain Femmam	University of Haute Alsace, France

Tiago Araujo	University of Nebraska-Lincoln, United States
Xiaofei Zhang	Nanjing University, China
Yahya Almayali	University College of Humanity Studies, Iraq
Yangmin Li	University of Macau, P.R.China
Yi Cao, Henan	Univerisity of Technology, China
Zheng Yang	University of Illinois, USA

Technically Sponsored by

Computer Science & Information Technology Community (CSITC)

Networks & Communications Community (NCC)

Digital Signal & Image Processing Community (DSIPC)

Organized By

Academy & Industry Research Collaboration Center (AIRCC)

TABLE OF CONTENTS

International Conference on Computer Science & Engineering (CSEN-2014)

International Conference on Advanced Computing (ADCO-2014)

INCORPORATING SYNONYMS INTO SNIPPET BASED QUERY RECOMMENDATION SYSTEM

Megha R. Sisode and Ujwala M. Patil

Department of Computer Engineering, R. C. Patel Institute of Technology,
Shirpur, Maharashtra, India
Megha_sisode@yahoo.com
patil_ujwala2003@rediffmail.com

ABSTRACT

Recently, growth of internet has been increased for information retrieval though it is difficult to extract the relevant information in less time. Search engine sometime fails to understand user search intend. Query recommendation can be used to help user to state exactly their information need. Search engine can return appropriate result to meet users' information needs. There are various methods based on history of users and snippets to retrieve the information. But these methods fail to satisfy users need. Therefore in addition of history and snippets with synonyms will do better. Moreover user preferences can be used to build the user profile which will help in effective recommendations. Here for given query recommendation the synonyms are extracted on line. Synonym based method ranks the clicked URLs at the top of the result based on user profile. The performance of the system shows that the synonym based approach give better and effective recommendation for all queries as compared to previous methods.

KEYWORDS

query recommendation, synonym, snippets, information retrieval, user profile

1. INTRODUCTION

As the growth of World Wide Web is increased with increase of size and popularity and the assembly of large scale volumes of web data, thus it is difficult to extract the relevant information that have been used in wide range of application. Many novice users face the difficulty to get the desired information although they use most efficient search engines such as yahoo, google.

The search engine has gain more success and the growth of internet resources is increasing as the web is a repository of large scale updated information. Web search engine is the major platform to extract the needed information to user by posing a query. The web search engine helps user to exploit the required information based on user query. For this purpose search engine provide platform to the users to specify their information need in the form of queries simply as list of keywords. This keyword based user interface causes lots of troubles in search process.

User queries are the most important factors as they are only interface for users to access web pages that affect the performance of search engines. Although users' information needs are

Natarajan Meghanathan et al. (Eds) : CSEN, ADCO - 2014
pp. 01–13, 2014. © CS & IT-CSCP 2014 DOI : 10.5121/csit.2014.4101

complicated, their queries are usually simple, short and possibly ambiguous. Queries are simple because users are unable to organize complicated queries which can describe their information needs more exactly.

This causes a major challenge in current Web search techniques, which is the understanding of user's information need behind queries. Sometimes it becomes difficult for search engines to understand information need from only queries, so that click through behavior data can be utilized. Thus the query recommendation technique is proposed to present users with a list of possible choices whose information needs are relatively clear to search engines. By this means, users can exactly state their information need by clicking recommendation query links instead of inputting new queries [1].

With the analysis into altavista search engine's query logs it is found that the average length of user queries is 2.35 terms and mostly the user queries are short including around two terms per query, on the other hand the ambiguity of language play essential role, as the users often fail to organize appropriate terms for their search query, so as the search engine returns mismatch results to the same topic and also faces the problem for synonyms and polysemous words that exist in language. Thus query recommendation function helps user to recognize their short formed and possibly ambiguous queries and return appropriate result to satisfy the user information need [1].

Recently the query recommendation has been widely used by the users to satisfy their information needs. According to the survey, it has been observed that approximately 78% users will change their queries with search engine recommendation function if they cannot obtain satisfactory results for their query. So that it is mandatory for search engine to provide good quality recommendations which can express users actual information needs more exactly.

According to research their has been lot of work done for improving result of search engine based on users previous query log data and click behavior so that search engine can locate popular queries which are similar to current query either in content or in click context [3].

These methods end with suggesting user to adopt a similar and/or frequently adopted query also fails to exactly understand users' information need and also does not consider current users search intent into account as they believe that current user shares similar interest as other user with the same query. These methods also produce improper recommendation for low frequency query as not much candidate queries for them.

In order to solve these problems and give better recommendation results which can satisfy users information need, we have to understand the way to express the users' information need. For better recommendation the query must be formulate properly and well organized manner with more exact meaning. If we observe the way user search on web, then we will come to know that when user clicks certain search result returned by search engine, it does not always mean that user is interested in the resultant document as because she/he has not yet viewed the resultant document. Instead the assumption is the user must be interested in the snippets of the corresponding document because it is the snippets that are actually shown to and read by user.

The synonym based method follows that users' information needs are described in the interaction with search engine, specifically, in the snippets which they have been ever clicked for the results.

Thus on the basis of these assumption, a synonym based query recommendation framework with snippet click model is presented, which include global scale and local scale snippets. With these models keywords are extracted from clicked snippets to make effective recommendations with using synonyms of query word along with snippets and location information. Differently

synonyms based query recommendation methods gives effective and more accurate results as compared to the history and snippet based system.

Nowadays, google is a world most leading search engine with different language interfaces. There exist some limitations with the keyword based searching. One of the web search key issue is that user tend to insert very general queries. That leads huge amount of information to be returned for given query. There are various ways to deal with a huge amount of retrieved web pages for arranging with the proper meaning. Synonyms or word sense disambiguation can be used along with snippets. The synonym based query recommendation approach uses WordNet* for discovering the synonyms.

The rest of the paper is organized as follows. Section 2 gives literature survey on work done in query recommendation processing methods; section 3 explains the motivation for synonym based method; then section 3 gives the working of synonym based recommendation method. Section 4 describes experimental setup for synonym based method. Finally conclusion and the future work is explained in section 5.

2. LITERATURE REVIEW

Recent researchers have proposed various recommendation systems for online information retrieval using various approaches. A literature survey is done to examine different approaches in order to mine essential features from query log data of search engine.

Ricardo Baeza-Yates et al. had proposed a method for suggesting list of related queries to user based on a query clustering process. This method not only discovers the related queries, but also ranks them according to a relevance criterion. This notion of query similarity has several advantages that it is simple and easy to compute. Moreover, it captures semantic relationships among queries by relating queries that are worded differently but stem from the same topic [2].

Silviu Cucerzan et al. had presented a method to suggest queries based on mining into post-query browsing behaviors referred as “search trails”. They utilized user landing pages i.e. the ending pages of search trails to generate query suggestions. For each landing page of a user submitted query they identify queries from query logs that have these landing pages as one of their top 10 results and these queries are used for suggestions [3].

Shen Xiaoyan et al. proposed an effective approach for query suggestions. This approach accepts Chinese web query as input and the approach not only identify related queries already existed in the log of previously submitted queries of search engine but also use synonyms that are extracted from web based corpuses to construct new related queries. Also rank the queries according with degree of relatedness, freshness and effectiveness. This approach proves its effectiveness in recommending related queries for high frequency queries than that of low frequency query [4].

Qi He et al. proposed a novel sequential query prediction approach to grasp a users’ search intent based on her/his past query sequence and its resemblance to historical query sequence models mined from massive search engine log data. Differently from previous work done where only single preceding query is used for prediction, this work considers variable number of preceding query and effectively captures more complex context information for recommendation. The Results shows that the sequence-wise approaches significantly outperform the conventional pair-wise ones in terms of prediction accuracy[5].

---

* *http://wordnet.princeton.edu/*

Thus the work has one fundamental difference from all previous session-based approaches. As all previous work focuses on pair-wise query relations and uses only a single preceding query for query prediction, proposed method consider variable number of preceding queries and effectively capture more complex context information for query recommendation. Moreover, this approach can automatically determine the optimal context length to be used for query prediction [5].

Hamada Zahera et al. proposed a method for suggesting a list of queries that are related to the user input query based on previously issued queries by the users. Their method was based on clustering process in which groups of semantically similar queries posed by user are detected in order direct them toward their required information need. This method not only discovered the related queries but also rank the query according to a similarity measure [6].

C. Sumathi et al. proposed a session based approach where the proposed method is based on the users' navigational patterns and provide recommendations to fulfill the current users information need. This method had classified and matched an online user based on her/his browsing interests [7].

Poonam Goyal et al. had proposed a method to facilitate users with query recommendations in which the concepts related to the users information need are suggested to the users to satisfy their exact information need. In that they extracted the concepts from the web snippets and have used two weight functions to measure the relevance between query and concept. Related concepts with different meaning are selected and recommended as query suggestions to the users [8].

Ji-Rong Wen et al. had proposed an approach to cluster similar queries to recommend URLs for frequently asked queries of a search engine by using four notions according to: first, the context of the query; second, common clicked URL's between queries; third, string matching of keywords, and fourth is, the distance of the clicked documents in some pre-defined hierarchy. But result of this method generates very sparse distance matrices and this sparsity is diminished using large query logs. Thus string matching features are used to locate similar queries [9].

Osmar Zaiane et al. had used content similarity to recommend similar queries using Query Memory, a data structure that holds the collective query trace and also extra information pertaining to the queries that would help in measuring similarities between queries. Query trace is a log containing previously submitted queries. The major advantage of this method is that it suggests the queries when user is not satisfied by current search result but sometimes this method produces irrelevant result and leaves the choice up to user [10].

Eugene Agichtein et al. shows that incorporating user behaviour data can significantly improve ordering of top results in real web search setting. Also alternatives for incorporating feedback into the ranking process has been examined and explored the contributions of user feedback compared to other common web search features[12].

Yiqun Liu et al. had presented an approach to focus how to detect users actual information need by extracting the snippets. The snippet based approach considers that users information needs are better described in their interaction with search engine more specifically, in the snippets of the results which ever clicked by users. The key idea of that system follows that if user click certain result from list then it shows that the user has read that particular snippet and interested in that snippet and not in the result. But this system does not consider the location and synonyms also, which may be useful for improving the results of search [1].

3. MOTIVATION

Many query recommendation methods used to suggest related queries by extracting information from clicked documents because these documents are expected to contain users' preference and relevance judgments. Different methods use previous query data, history of snippets. But this method does not consider current user search intent exactly, also fail to recommend for low frequency query. In addition to the snippets of clicked documents, the synonyms extracted from the online synonym services can be integrate to improve the performance of search engine. Also user preference can be integrate to form current user profile to specify the current users search intent. So the synonym based recommendation system is expected to give more accurate recommendation based on input.

4. SYNONYMS BASED QUERY RECOMMENDATION SYSTEM

The synonym based query recommendation method is based on the assumption that users' information needs are described more specifically in snippets of the results which they ever clicked, this is because when user clicks a certain search result, it does not necessarily mean that she/he is interested with the result because she/he has not yet viewed the resultant document. It is probably mean that she/he is interested in the snippets of the corresponding resultant document because these snippets are actually shown to and read by users. According to this assumption, the synonyms based query recommendation uses clicked snippets also the synonyms extracted from the online synonyms service for considering the synopsis.

In addition with this synonym based recommendation system also considers location information. As in previous method if the user is at some specified location then the system does not consider the location of user and recommends without considering location but in synonym based recommendation method it improves the accuracy in result with considering location information along with synonyms extraction. Figure 1 shows the working of synonym based query recommendation method.

The synonym based recommendation is based on a snippet click models which tries to extract keywords appearing in users' clicked snippets as recommendation. According to Baeza-Yates et al. query recommendation is the method which is used to suggest alternative queries to users in order to help them to specify alternative related queries in their search process [2]. But believe is that the users not only specify alternative related queries but also try to express their information need in the form of query recommendations. Therefore, search engine should recommend queries which are most likely to represent users' information needs.

The synonym based query recommendation task try to rank snippets which are related to the original proposed query on the basis of user profile. Users are interested in the content of snippet because it contains keywords that are related to their information needs and these are actually shown to and read by user. Therefore, the major idea of synonym based query recommendation framework is to locate keywords that appear in snippets clicked by users and can describe users' information need. Differently with previous recommendation methods, it relies on information extracted from users' result click-through process instead of the historical queries fired by other users.

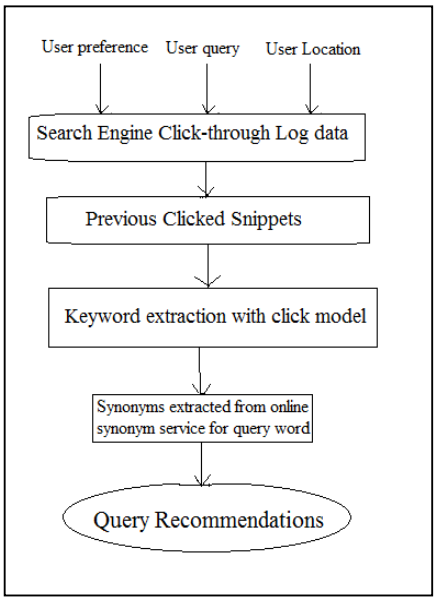


Figure 1. Synonyms based query recommendation system with considering user preferences and location information

4.1. Snippet click mosels incorporating with synonyms

Note that users click a certain resultant document because she/he actually views its corresponding snippet and also expects this document to meet her/his information need. Therefore, the probability of clicking a certain document is decided by both whether user views the snippet and whether user is interested in it. Because user is only able to view the snippet of the document before she/he actually click on the result, then the probability of clicking is decided by whether user is interested in the snippet of the result document; in other words, by whether this snippet meet user's actual information need or not. For synonym based recommendation both a global scale and a local scale snippet click models are used.

These local scale and global scale snippets are adopted to finish the task of query recommendation.

- **Global scale snippet model using synonyms**

For the global scale model, all the clicked snippets for a certain query are treated as a whole ''snippet document". Therefore, for all clicked snippets, it shows that if user has clicked certain snippet then user must be interested in it and that satisfies the users' information needs. Therefore, a simple TF-based model is used to extract keyword lists from the snippets. For each keyword in the snippets, the recommendation candidates are those with the largest term frequency value, where for a query word W, TF is defined as sum of all appearances of W in all related snippet.

The corresponding global scale snippet algorithm using synonym based recommendation (Algorithm 1) is as follows:

Algorithm 1. Query recommendation based on global scale snippet click model using synonym

QueryRecommendation (Original query Q, Users Click through pattern CLKPAT)

1. Find all the documents clicked for Q in CLKPAT and form a set of document called D;
2. Extract all the snippets of D for query Q by using search engine interface and form a snippet set called S;
3. For snippet set S, extract N keywords by using TF or other keyword extraction algorithms;
4. Extracts the synonyms for all the N keywords from online synonym service.
5. Return these N keywords as recommendation words with considering synonyms.

This algorithm generates a list of keywords for recommendation of query Q. Note that sometimes these keywords may not be directly used for recommendations because they should be combined with the original query to form complete information need. For example, keyword ‘‘free download” may be returned for query ‘‘Yahoo messenger”, it should be combined with the original query word to form a complete query recommendation word such as ‘‘Yahoo messenger free download”. However, even these keywords cannot be directly adopted as recommendations, these are supposed to meet users’ information needs.

- **Local scale snippet model using synonyms**

Differently in a local scale snippet click model each snippet is considered to be treated separately. With the bag-of-words model, a certain clicked snippet can be represented by a set of keywords each with having different TF values. As this consider each snippet separately, and not all keywords appears in the each clicked snippet. So many keywords will have term frequency value as zero, thus it may generate the sparsity problem. The smoothing technique can be used to avoid data sparsity problem and to estimate exactly the information need of user. After this we can consider the probability for each keyword is consider to describe users’ information need and the keyword having high probability to satisfy users information need are used to suggest for query recommendation.

The synonym based query recommendation algorithm for local scale snippet is as follows:
Algorithm 2. Query recommendation based on local scale snippet click model using synonyms

QueryRecommendation (Original query Q, Users Click through pattern CLKPAT, Users search interest SI)

1. Find all documents clicked for Q in CLKPAT and form a document set called D;
2. Extract all snippets of D for query Q by using search engine interface and form a snippet set called S;
3. Extracts users search interest SI by combining users profile P and location information L;
4. Recommendation candidate set CANDIDATE = { };
5. For each snippet S_i in snippet set S, if P(Click $_i$) is greater than threshold T , then put all words into CANDIDATE set;
6. For all words in CANDIDATE set, extracts the keywords from snippets and also after smoothing task, and form the equation E.
7. Solve E according to Gaussian elimination* or other methods.
8. Select N keywords with the largest probability values which indicate greater possibility of describing users’ information need;
9. Extracts the synonyms for all N keywords from online synonym service.
10. Return these N keywords as recommendation words with considering the synonyms.

Here also, as similar to Algorithm 1 these N keywords should be combined with the original query to form complete query recommendations. Except the keywords which only appear in snippets with having probability of clicking P(Click) values lower than that of threshold T.

5. EXPERIMENTAL SETUP AND RESULTS

For evaluation of the performance of the synonym based recommendation system, the synonym based query recommendation system is run on configuration having Windows 7 with 4GB RAM. The synonyms based recommendation method is implemented with java on android platform. For the system android works at front end and SQLite works at back end to store the database of application. Database is stored in the android device itself with the help of SQLite database. For this the Eclipse software development kit (SDK) is used, which includes the Java development tools to develop an android application, where Eclipse is an integrated development environment (IDE). An android emulator has been created with having query recommendation as an application on it working in contribution with SQLite database.

To evaluate performance of the synonyms based query recommendation framework, practical search engines database has been used and compared performance of synonym based recommendation with current search engine's query recommendation performances. The evaluation is different from most previous researches where the performance of query recommendation is evaluated by how many percentages of users actually clicked these recommendations in practical environment. The synonym based query recommendation method adopts human-annotation based precision-recall metrics for evaluation.

Precision and recall are the basic measures used in evaluating search strategies. Recall is the ratio of the number of relevant records retrieved to the total number of relevant records in the database. It is usually ex pressed as a percentage. Precision is the ratio of the number of relevant records retrieved to the total number of irrelevant and relevant records retrieved. It is usually expressed as a percentage. Recall and precision are inversely related. As the recall increases the value of precision get decreases, and vice-versa.

Click-through rate and user profile are adopted as metrics to evaluate the performance of recommendation algorithms. User profile is built from click patterns, users location patterns and user interest, all of which is available by user once the application is used by that user. For extracting the users location the latitude and longitude has been used. For the synonym based recommendation, click-through data of an application is tracked by the application itself and stored by using SQLite database.

Previously for snippet based query recommendation the experimental results show that the keywords generated by snippet query recommendation method are more preferred by users than the others. About one third of the recommendation is provided by Baidu and Sogou search engines match snippet based query recommendation algorithm results. In snippet based query recommendation method for all recommendations generated by search engines, some match recommendation keywords generated by snippet system while others not. The comparative results of click-through rate and average amount of user clicks are shown in figure 2 and figure 3[1].

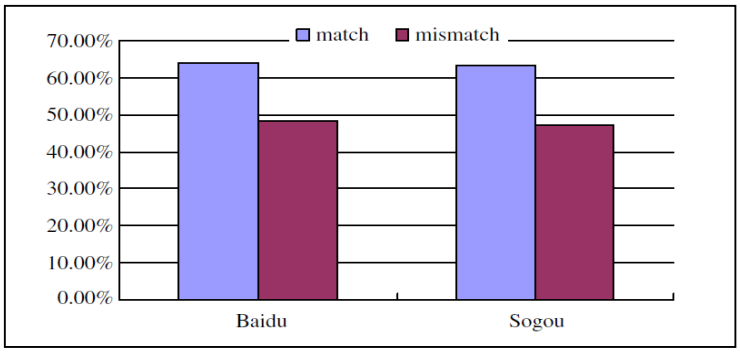


Figure 2. Comparison of click through rate between the recommendations that matches and does not match

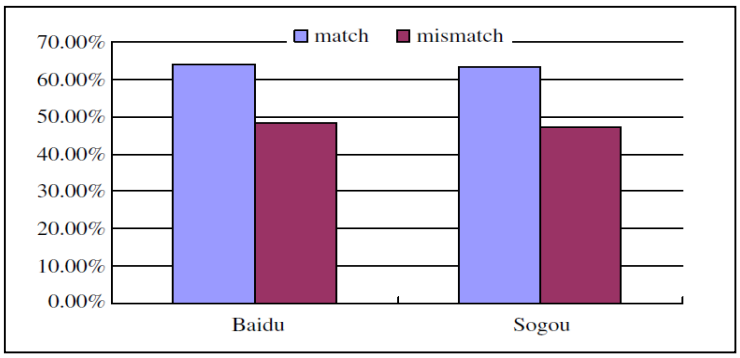


Figure 3. Comparison of average amount of user click between the recommendations that matches and does not match

The synonym based recommendation method have used google search engine database*, where global scale snippet is done by google itself. For global scale model all clicked snippets for certain query word are treated as a whole. In this users information need is supposed to be related with the snippets. For local scale model each snippet for a query word is treated separately. In this the system can use Algorithm 2 to estimate the probability of keyword in representing users' information need accurately.

The synonym based recommendation method is also incorporated with users' preferences. The method is also integrated with synonyms matching with the help of online synonym service that gives us the requested synopsis for the given words.

The results of the synonym based recommendation system are shown in the form of URLs. It gives URLs because it represents the links from where the snippets are fetched, and then for the result user have to click on the URLs to get the snippets. In case if the user clicks on a URL then the URL appears again in the result then the URL would be ranked at the top of the list. This concept of snippet ranking is based on users' profile.

For snippet based query recommendation the past experimental results show that the keywords generated by snippet query recommendation method are more preferred by users than the others. About one third of the recommendation is provided by Baidu and Sogou search engines match snippet based query recommendation algorithm results. In snippet based query recommendation method for all recommendations generated by search engines, some match recommendation keywords generated by snippet system while others not.

---

* *https://developers.google.com/web-search/docs*
https://developers.google.com/cloud-sql

In order to measure the performance of synonym based query recommendation the experimental setup is made with the current well known google search engine database with user profile, user preferences and click through data as metrics. For evaluation the same sample query has been fired to synonym based recommendation system and measured the performance by posing the same query in google search engine. The results are compared on the basis of results returned and delay time.

For example different sample queries are run, and compared with the returned results from google with the same sample query, and the precision and rank of returned results are measured for synonym based query recommendation system. Among these there are variation in time required. Also the synonym based system have more relevant results at the top as this is considering the synonyms returned by online synonym service for query word. Figure shows the representation of the result for sample query run on synonym based query recommendation.

From the observation it has been noticed that as compared with the history and snippet based methods; synonym based recommendation gives better results with synopsis. Also it ranks the link at the top on the basis of users preference and location. The synonym based recommendation method able to satisfy users' information need in less time. For the evaluation of the synonyms based query recommendation method different metrics have been applied such as first percent precision is calculated for differed sample queries and another parameter is used as number of related queries.

In figure 4, the vertical axis represents the % precision calculated for the query and the horizontal axis represents the number of results matched with the user search intent. Also the performance of the method incorporating synonyms with snippet based recommendation is measured with having percent of matching recommendation and the rank of the required result as a metrics of recommendation for different sample queries.

It can be observed from the results that the percentage of precision increases if the query presented by the user is high frequency query and also the time required to returns the result get increase with low frequency query.

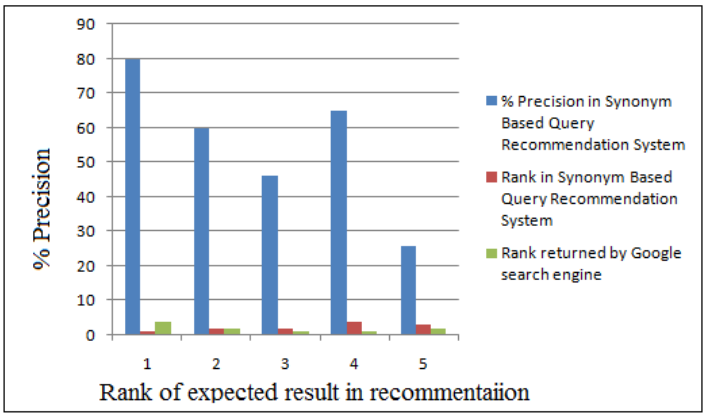


Figure 4. Result of synonym based query recommendation for few sample queries

- **SNAPSHOTS FOR SYNONYMS BASED QUERY RECOMMENDATION**

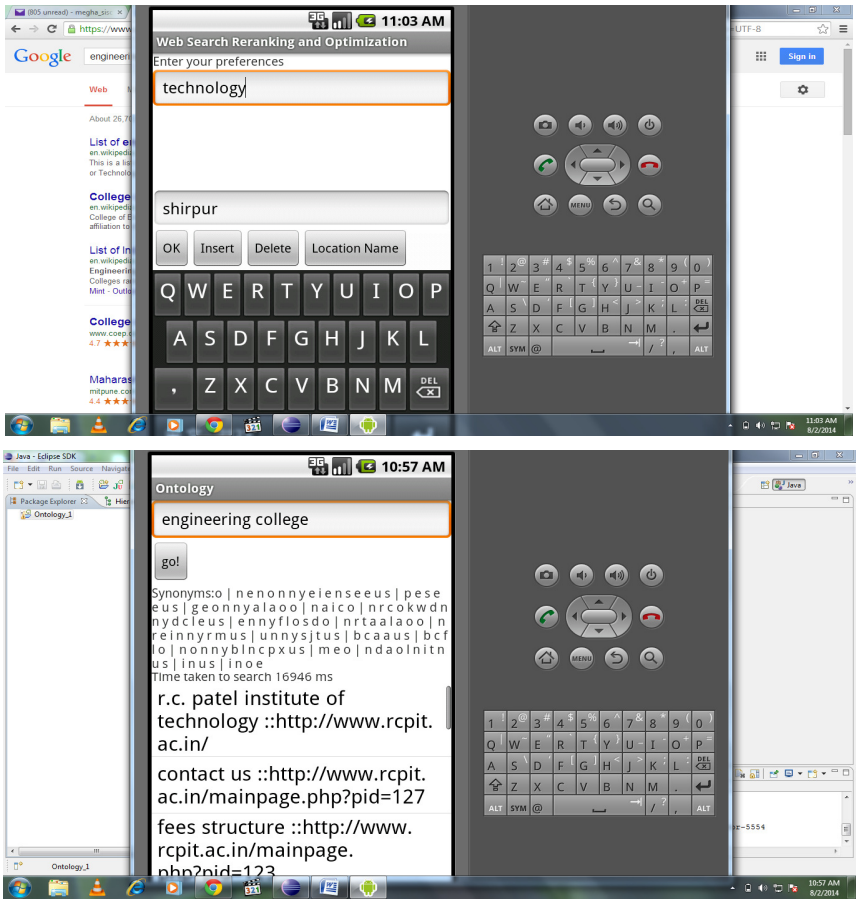


Figure 5. Results of synonym based query recommendation for one of the sample query

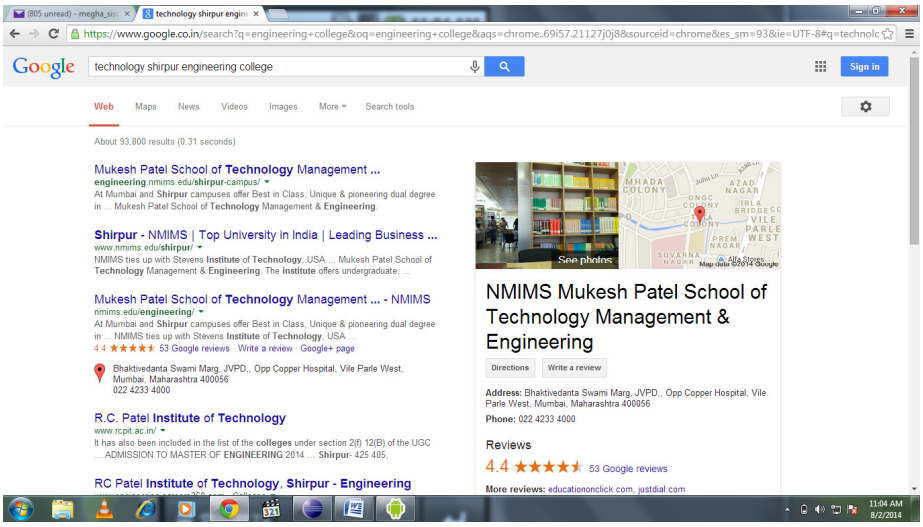


Figure 6. Results of Google search engine for the same sample query

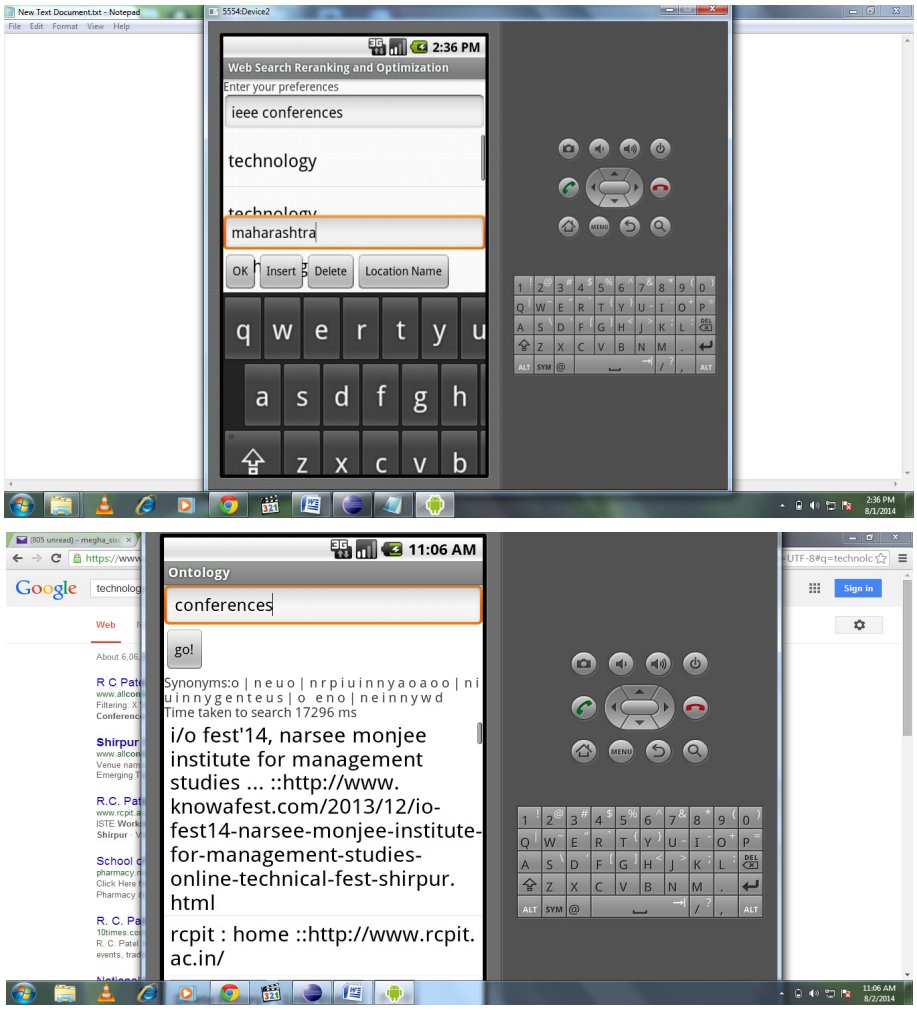


Figure 7. Results of synonym based query recommendation for another sample query

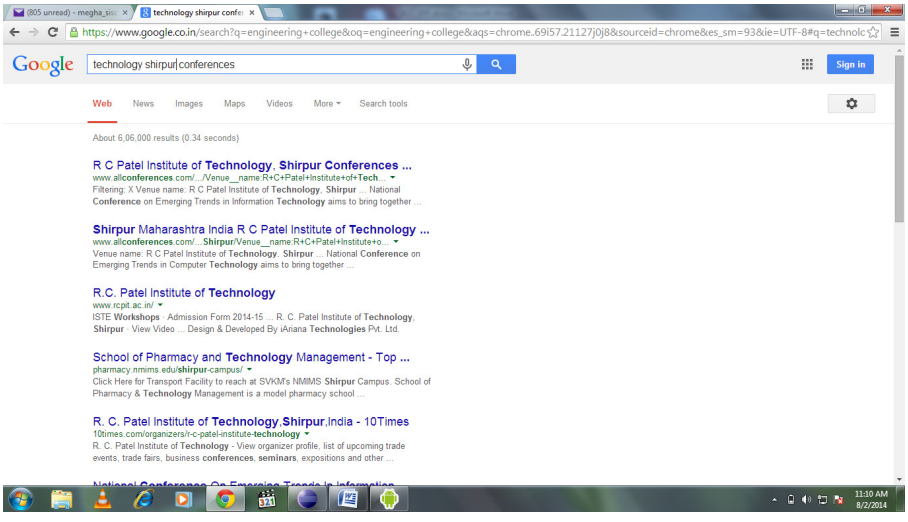


Figure 6. Results of Google search engine for the same sample query

6. CONCLUSION

Mostly many query recommendation systems try to use the previous queries which are similar in either manner with current query. But these methods lack to exactly understand users information need. In order to improve the performance the synonyms and location information is added with snippet information. Global and local scale snippet click models have been used with google search engine log data along with synonyms, which are retrieved for the query by online synonym service in addition with user preferences. By analyzing the results returned by search engine google and compare these results with synonym based query recommendation system. It has been observed that synonym based query recommendation is more efficient. In addition to that synonym based query recommendation performs better for low frequency query. In future, we hope to extend this approach to make use of correctly identified intent for query rewriting by fetching users current location automatically to improve the performance of searching.

REFERENCES

[1] Liu, Miao, Zhang, Ma, and L. Ru, (2011) "How do users describe their information need: Query recommendation based on snippet click model", Expert Systems with Applications, vol. 38(11), pp. 13847-13856.

[2] Baeza-Yates, Hurtado, and M. Mendoza, 2005 "Query recommendation using query logs in search engines", Current Trends in Database Technology-EDBT 2004 Workshops, Springer Berlin Heidelberg.

[3] Cucerzan, and R. White, 2007 "Query suggestion based on user landing pages", Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval, ACM.

[4] Xiaoyan, Bo, Junliang, and M. Xiangwu, 2008 "An effective method for chinese related queries recommendation", In Software Engineering, Artificial Intelligence, Networking, and Parallel/Distributed Computing, 2008, SNPD'08, Ninth ACIS International Conference on, pp. 381-386, IEEE.

[5] He, Jiang, Liao, Hoi, Chang, Lim, and H. Li, 2009 "Web query recommendation via sequential query prediction", In Data Engineering, 2009, ICDE'09, IEEE 25th International Conference on, pp. 1443-1454, IEEE.

[6] Zahera, El Hady, and W. El-Wahed, 2010 "Query Recommendation for Improving Search Engine Results", In World Congress on Engineering and Computer Science (WCECS), San Francisco, USA, vol. 1.

[7] Sumathi, Padmaja Valli, and T. Santhanam, 2010 "Automatic recommendation of web pages in web usage mining", International Journal on Computer science and Engineering (IJCSE), vol. 2, pp. 3046-3052.

[8] Goyal, and N. Mehala, 2011 "Concept based query recommendation", In Proceedings of the Ninth Australasian Data Mining Conference, vol. 121, pp. 69-78, Australian Computer Society, Inc.

[9] Wen, Nie, and H. Zhang, 2001 "Clustering user queries of search engine", In Proceeding of the 10th international conference on World Wide Web, 2013, pp. 162-168.

[10] Zaiane, and A. Strilets, 2002 "Finding similar queries to satisfy searches based on query traces", In Advances in Object-Oriented Information Systems, pp. 207-216, Springer Berlin Heidelberg.

[11] Baeza-Yates and B. Ribeiro-Neto, 1999 "Modern information retrieval", vol. 463, New York: ACM press.

[12] Eugene, Brill, and S. Dumais, 2006 "Improving web search ranking by incorporating user behavior information", In Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval, pp. 19-26, ACM.

INTENTIONAL BLANK

WEB TESTING APPLICATION WITH PHP AUTOMATED TOOL

Iulia Ștefan and Ioan Ivan

Department of Automation, Technical University, Cluj-Napoca, Romania
Iulia.Stefan@aut.utcluj.ro, ionut.ivan1@gmail.com

ABSTRACT

The web applications development has experienced an explosive growth in variety and complexity during the past decade. Most web-based applications are modelled as three tier architecture, the client side experience remaining virtually unchanged, while server-side is updated. However, client-side architecture can change with unexpected results. Consequently, testing procedures should support continue improvements to pursue the current trends and technology. This paper presents an automated tool for testing client-side component of web applications. The testing data is extracted using a crawler. Adopting several procedures, the general aspect of the page is analysed (CSS regression testing). All of the content is tested, including links, images, forms, and scripts. The resulted test cases are automatically created, leaving the user with the option to decide over their usage.

KEYWORDS

Automated testing, web application, PHP

1. INTRODUCTION

The development of the web applications has become an important area in the field of software engineering. Web applications have a core set of specific characteristics like modularity, by which different functionalities of the same product are written in an in-dependent manner. Their reusability for different tasks in the same application is a viable cost reduction solution. A notable advantage is that modules can be created using a variety of technologies deter-mining the performance improvements.

The techniques used for developing web applications have become more and more varied. The enhancement in mobile devices industry in the past few years has created the need for web applications to accommodate both regular clients, and mobile ones. Web application responsiveness has increased because data transfer has been made more and more efficient. The new technologies have driven the validated concepts to new heights and old technologies have been updated to keep up with the fast pace. JavaScript libraries are the best example, by their later evolution.

The diversity of development technologies has led to a growth in web application testing technology. In spite of this fact, testing has become more difficult. A complete solution can consist of several modules written in several languages, there is no single tool for testing the whole application. Web applications are made from static and dynamic components. The static

Natarajan Meghanathan et al. (Eds) : CSEN, ADCO - 2014
pp. 15–21, 2014. © CS & IT-CSCP 2014 DOI : 10.5121/csit.2014.4102

pages can be tested automatically with the use of crawlers or other spider-like tools. The dynamic content is difficult to be analysed in an automated approach.A short list of techniques is available. A record and replay tool is one solution. Such software tools record an initial testing scenario and then generate test scripts for automated regression testing. A major drawback of this technique is that if there is a change in the user interface, the generated scripts have a greater chance of failing. Another disadvantage is the cost associated to manual recording of scripts. However, it is relatively easy to deploy and it is autonomous, allowing the tester to deal with other tasks. In this paper, an automated tool that extracts information from the user interface and submits it to the user analysis is described. In section 2, a short description of similar tools is provided. In section 3, the aspects of testing activities are reviewed. Section 4 provides an insight related to the development of the present-ed tool. In section 5, experimental results are de-tailed, followed by section 6 with conclusions and last section, References.

2. GETTING STARTED

The domain of software testing represents an interesting research domain and application development.

Every element of a web application is considered an object [1]. The test cases are generated using as starting point the data flow chart.

The comparison between developing tests cases by programing or by capture-relay is emphasized by [2]. After several experiments, the conclusion evidenced the advantage of cost reduction in test maintenance by programing approach and the advantage of reduced time allocated when development was centered on the capture-relay approach. The tools used for functional web testing in this case were Selenium IDE and Selenium WebDriver.

The development of a new testing tool was enhanced as presented in [3] by using models. The test cases can be automatically or manually selected by request. In [4], the authors adopt a different approach. The Ajax events are used to design a state based-approach using the DOM as the template. Generated events determine the transition between states, and the interaction chain generates the test cases. An approached based on image analysis [5] detects defect images and broken links. Several testable objects are identified by the application decomposition. The image processing techniques could be an important aspect for automated CSS testing. The main concept is the usage of a static test copy of the web page when testing in order to separate the functionality from the shape.

The PhantomCSS tool is also presented. The tool generates an image of a portion of the page and the then establishes differences against the original.

3. TESTING TECHNIQUES

The main objective of testing activities is to detect and fix faults and demonstrate the product quality against stakeholder's requirements and application's specifications. Several stages are necessary: the test planning, the goal definition, the general strategy selection, the test execution, and the result's analysis. The testing process tends to start early during the development stage and continues after deployment. For the web applications, the requirements could be reduced in size and test models are not established.

There are two main approaches regarding software testing based on code visibility and are summarized further on.

3.1 White Box Testing

In white box testing, the source code is available in its entirety to the tester giving extra insight on the application and removing the not clarified particular piece of code. White box testing is most useful for internal unit verification. The tests are designed to highlight the application's behavior related to specific units.

However, the tools available for this approach are restricted to a handful of programming languages and technologies: java, C#, C++, PHP, python and several others.

If the frameworks are not maintained and updated constantly, the incompatibility with newer technologies is unavoidable.

3.2 Black Box Testing

The black-box testing approach (BBT) is also known as client-side analysis; there is no code available. The structure of the website becomes available by sending requests to the server and examining responses. This method cannot make a difference between static and dynamic pages without appropriate tools.

The simplest form of black box testing is to start running the software and make observations if expected and unexpected behavior is be easily distinguishable. Abnormal behavior, such as crashes are easy to spot. As soon as the cause has been determined to be part of the program, relevant information can be passed to a competent party for fixing.

Another, more advanced form of black box testing is the use of checklists. The checklists represent specifications about expected behavior based on applied known inputs. The term input stands for any action or resource provided at the beginning or during the runtime. If problems are discovered, specific actions are implemented in order to fix the issue.

The developed tool offers support for the black box testing techniques identifying test cases to be considered when creating tests. Regarding CSS verification, regression testing is to be used. Regression testing aims to uncover new faults in existing functional or non-functional parts of an application after changes such as patches, upgrades or configuration modifications have been made. A common method of regression testing is the rerunning of previous tests to check if the behavior of the application has changed in conflict with the existing specifications. This method can be used to test the correctness of a program as well as the quality of the output. A downside of regression testing is the cost. In theory, a complete test suite should be run after each update of the application, which is too resource-intensive to be applied [6]. In a real development and testing environment, a minimal set of tests is devised to check the specific functionalities that have been altered.

Other than client specifications, there is no need for discovering the entire application structure. The only situation where the server-side is involved is when the link response is checked. For the form and button analysis, the proposed tool is using pre-generated data (for form input).

4. DEVELOPMENT

A simple web application used for online order placement is used as software under test (SUT). The SUT's structure consists of several PHP files linked together. The content (images and text) is retrieved from a database. A common, simple style is used throughout the website. It is used because it offers a minimal structure, simple Ajax processing and a simple CSS file, which makes

it easier to analyze initial results. When all major issues have been resolved, larger, more complex websites should be tested.

Using a PHP page as an interface, the tester is able to choose between several options as shown in the Figure 2.

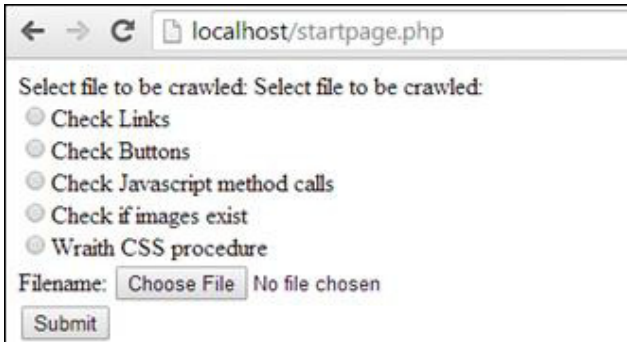


Figure 1. The user interface page for test option

The links and images are analyzed first. Using phpCrawl 0.81, an open source PHP library functions package, the target page is read and its content is saved for further use. From this content, links can be extracted. The application workflow is presented in the Table 1.

As shown in Figure 1, based on the server response when the crawler attempts to access them, links are classified as “good” (200 response), “bad” request (400 response), “forbidden” (403 response), not found (404 response) or not allowed (405 response). For the forbidden or not allowed responses, the user is prompted to check the rights for the called page. For the bad syntax response, the link is submitted for direct analysis. For the not found response, a mock page can be generated using information from the link itself.

The page only echoes any parameters that are passed to the page. This mock page is only a placeholder until the real page can be added to the server. The images are processed in a similar manner. Their existence is checked, and if they cannot be found, a temporary picture is inserted as a placeholder. A syntax check is performed and if there are any function callbacks in the <img> tag, the function name is searched for. If the function does not exist in the extracted content, the user is notified. An option is available to ignore any callbacks and only check for basic syntax. Forms and buttons are also extracted from the page content. To verify any validations that may be employed in the form, a correct set of inputs is given. If the response is good, the form is considered correct.

Forms and buttons are also extracted from the page content. To verify any validations that may be employed in the form, a correct set of inputs is given. If the response is good, the form is considered correct. This creates a term of comparison. Then, a wrong set of inputs is given. If there is an event such as a redirect, or a simple html required validation, the user is notified. A syntax check is also performed. The amount of time spent for CSS regression testing can be lengthy if the process is not automated. Verifying different browser resolutions can be time consuming in testing. By using automated tools, the testing time can be cut down dramatically, with fewer bugs making their way into ready to deploy products or live applications.

For the CSS verification, Image Difference Comparison is used. As the customer settles on a general design, a mock-up page can be created for later editing. This starting point can be compared to an existing template. Snapshots are taken from both the mock-up and the template at

the same state. These snapshots are then compared and differences are highlighted. For more detailed results, snapshots can be taken for each element in particular (i.e. a snapshot of the footer section).

Table 1. Application workflow

Steps	Operation	Description
1	Start page	Entry point of application
2	Acquire content	Extracts the content of page for analysis.
3	Content verification	Includes images and links.
4	Acquire template for aspect checking	Applies CSS regression testing to the page.
5	Compare empty page to template	Applies CSS regression testing to page as a whole.
6	Break down template and page into individual components and redo tests.	This ensures every element is correct.
7	Dynamic content testing	Ensures that content is generated properly.
8	Parse page for scripts and external functions	Find where the functions are being called, if any exist.
9	Find external functions by name	Compares function calls to a list of function names from external libraries (jQuery, prototype, etc.)
10	Find functions by name	Either finds function definitions in page or any external file.
11	Compare expected response with actual response	Can be done using image comparison techniques.
12	Check validators	Important for safety.

A more focused diff can be generated in this manner, but it may take away the more general view. This remains a user preference. Again, there is no need to know the whole structure of the site. This process can be repeated at any stage of the development process. For this issue, Phantom CSS is used because there is no current need to have a very complicated tool for analysis. It handles the image comparison section required by the CSS testing. As an alternative, Wraith front end regression testing tool can be employed. It is somewhat similar in functionality to Phantom CSS. The difference lies in the components. While both tools use PhantomJS as a headless browser to move through the site, Wraith uses a tool called ImageMagick for snapshots. Wraith also has the ability to create screenshots of different resolutions.

Other responses are more difficult to automatize because of either syntax or server issues. The Image Difference Comparison method had serious initial issues. Because it is based on image comparison, even the slightest modification compromised the results. For example, in a full page test suite, a 2 pixel wide padding added around the footer caused all the tests to fail. In such situations, individual component testing proved to be more reliable and more time efficient. Warnings were added to the user interface, but the choice is left to the user.

In the first runs, tests failed continuously because of the images in the site. This called for a different strategy. After the creation of the mock page, all visual content was replaced with a

blank space and then the page was compared with the template. The method yielded improved results, and false test failures were mostly removed.

5. EXPERIMENTAL RESULTS

The link analysis method has proven effective especially when the 404 response is returned. Mock pages are generated automatically if the user wishes to do so. After choosing the option from the user interface page as presented in Figure 1, the tester receives a response as shown in Figure 2.

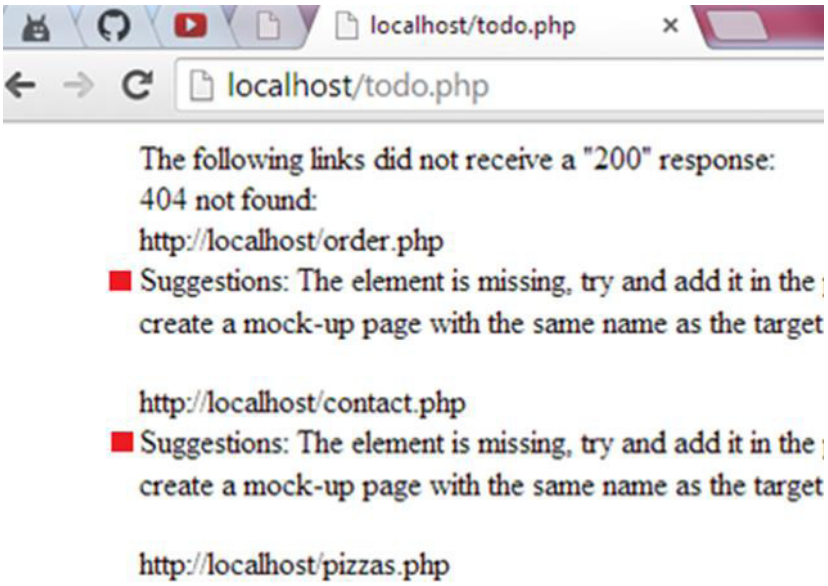


Figure 2. 404 Type error message with suggested options Web page

All links and images are obtained by the use of a crawler. They are all checked for syntax and existence. If a target of a link or an image can be replaced with a mock target, the user is to be notified.

All forms and buttons are tested. Form data is supplied by the user. Changes in the page are recorded and the user is notified. Buttons are tested for correct syntax. CSS testing generates a flat page and compares snapshots of it to a given template. Any differences are highlighted. Current work is largely focused on image processing and Ajax content analysis. The main goal now is to verify any function (i.e. JavaScript functions) for correct functionality.

The tool is under development stage and does not have a stable release. Functionalities may be added or removed to extend its usefulness.

6. CONCLUSION AND FUTURE WORK

An approach for testing web applications is presented in this paper. The front end of the application is decomposed and analyzed.

REFERENCES

[1] Kung, D., Liu, C. H. & Hsia, P.(2000) "A Model-Based Approach for Testing Web Applications", Proceedings of the Twelfth International Conference on Software Engineering and Knowledge Engineering, 6-8 July 2000, Chicago, USA, Chicago, Knowledge Systems Institute.

[2] Leotta, M., Clerissi, D., Ricca, F. & P. Tonella(2013) "Capture-Replay vs. Programmable Web Testing: an Empirical Assessment During Test Case Evolution", 20th Working Conference on Reverse Engineering, October 14-17, Koblenz, Germany, IEEE Computer Society.

[3] Boumiza, D.S. & Ben Azzouz, A. (2012) "Design and Development of a User Interface to Customize Web Testing Scenarios", Proceedings of The International Conference on Education and e-Learning Innovations, July 1-3, Sousse, Tunisia, New York, Institute of Electrical and Electronics Engineers (IEEE).

[4] Marchetto, A.,Tonella, P. & Ricca, F (2008) "State-Based Testing of Ajax Web Applications", 1st International Conference on Software Testing, Verification, and Validation, Lillehammer, Norway, April 9-11, 2008, Los Lillehammer, IEEE Computer Society.

[5] Torkey, F.A., Keshk, A., Hamza, T. & Ibrahim, A(2007) "A New Methodology for Web Testing", 2007 ITI 5th International Conference on Information and Communications Technology (ICICT 2007) - Media convergence: Moving to the next generation, New York: Institute of Electrical and Electronics Engineers (IEEE).

[6] Brooks, F (1995), "The Mythical Man Month: Essays on Software Engineering(2nd Edition)", Essex, Addison-Wesley.

AUTHORS

Eng. Iulia Ştefan, PhD. student, is a member of the Automation Department, Faculty of Automation and Computer Science, Tehnical University of Cluj-Napoca, Romania since 2006, being inv olved in several research projects covering web technologies, online platforms and software testing.

Eng. Ioan Ivan is a Master Degree student in Computer Science at the Technical University of Cluj-Napoca. His area of interest covers web technologies, testing, and more recently, the mobile development field.

INTENTIONAL BLANK

TRIBASIM : A NOVEL NETWORK ON CHIP SIMULATOR BASED ON SYSTEM C

Daniel Gakwaya[1], GaoYuJin[2], Jean Claude Gombaniro[3] and Jean Pierre Niyigena[4]

Department of Computer Science, Beijing Institute of Technology , Beijing,100081

[1]wayadn@yahoo.fr

[2]paulgyj@gmail.com

[3]Gombaniro002@yahoo.fr

[4]niyigelnx@yahoo.fr

ABSTRACT

In this paper, we develop a simulator for the Triplet Based (TriBA) Network On Chip processor architecture. TriBA(Triple-based Architecture) is a multiprocessor architecture whose basic idea is to bundle together the object programming basic philosophy and hardware multicore systems[1] .In TriBA ,nodes are connected in recursive triplets .TriBA network topology performance analysis have been carried out from different perspectives [2] and routing algorithms have been developed [3][4] but the architecture still lacks a simulator that the researcher can use to run simple and fast behavioral analysis on the architecture based on common parameters in the Network On Chip arena. We present TriBASim in this paper ,a simulator for TriBA ,based on system c[6] .TriBASim will lessen the burden on researchers on TriBA ,by giving them something to just plug in desired parameters and have nodes and topology set up ready for analysis.

KEYWORDS

Keywords: NOC ,triba , simulator, system c

1. INTRODUCTION

The last decade has seen Networks on chip emerge as a viable replacement for the traditional bus based interconnection system that has dominated in systems on chip for at least 3 decades. This is due the flexibility of design and most importantly the reduction in energy consumption for computing chips inside our electronic devices Networks on chip offer[5].

Networks on chip were introduced by a few pioneer papers that pointed out that future system on chip designs will be limited the quality of the interconnection system between computing modules[6,7,8]. They proposed a brand new idea that views the System on Chip as a micro-network of components. New designs would borrow ideas from the Data Networks research area and replace bus based interconnection systems with packet switched networks between modules within the System on Chip.

Although Networks on Chip have a lot of similarities with Data Networks ,there are differences one needs to consider .For instance NoCs are constrained to work within small distances inside the SoC while Data Networks can span kilometers of distance[6] .Also the links connection

Natarajan Meghanathan et al. (Eds) : CSEN, ADCO - 2014
pp. 23–30, 2014. © CS & IT-CSCP 2014 DOI : 10.5121/csit.2014.4103

structure is more predictable for NoCs than it is for Data Networks .This led to completely new designs, protocol stacks and routing algorithms new Networks on Chip would be built upon. It is also important to note that the micro-network of components way of thinking used in NoCs allows abstraction in Traffic Modeling[9].

Numerous network on chip architectures have been proposed in academia and industry, the topologies such as 2-D Mesh, Torus and Hypercube have been used in various network on chip designs. Along with these topologies, new routing algorithms, switching techniques and flow control mechanisms are selectively combined to meet the particular needs of the system on chip design[9].

TriBA is a network on chip architecture that enforces the concept of Object Oriented Design in the way SoCs are designed[10] .It is suitable for sophisticated embedded applications with multiple concurrent processing centers. This topology' s advantage over other 2D topologies such as hypercube topology is ease of realization and assembly [1]. Its nodes are connected in triplets , and higher order triba networks are recursively deduced from lower order ones. TriBASim is introduced in this paper, a simulator based on system c specifically designed to meet the daily needs of a researcher working on TriBA.

The rest of this paper is organized as follows: Section 2 explores already present NoC simulators and studies their intended use .Section 3 introduces TriBA and discusses the details relevant to our design; we delve into the design in section 4; Section 5 shows practical uses of the simulator .Future plans for TriBASim are addressed in Section 6 and Section 7 concludes the paper.

2. RELATED WORK

Numerous Network on Chip simulators have been developed before, targeting different areas in research and industry .Orion [11,12] was developed to run power and area analysis for Networks On Chips .Users input router and link components to build different network configurations and run their analysis .Power and area analysis for TriBASim was basen on On Orion power models .Noxim[13] NoC simulator is based on systemc ,and it can be used to evaluate the quality of a NoC in terms of delay throughput ,area and power consumption. Modified versions of Noxim have been used to run performance analysis using some popular topologies such as torus and twisted torus [14].

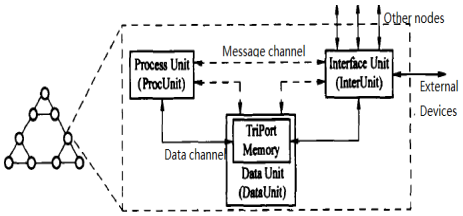
NIGRAM [15] is another Noc Simulator also based on systemc .It uses discrete events and is cycle accurate .It is very useful when testing routing algorithms on some regular topologies .One should also mention Nostrum[16] ,a project focusing on developing Network-on-Chip architecture. It addresses the communication issues from the physical to the application levels .These are the simulators that have been relevant to this research ,interested readers can refer to [17] to dig more and see a more detailed list .

3. TRIBA OVERVIEW

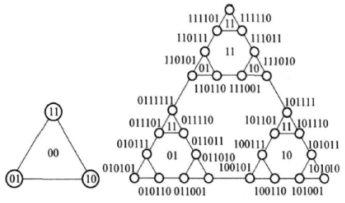
"A picture is worth a thousand words!" , Fig[1] and Fig[2] will be the basis for our description of TriBA . Fig[1] displays the low level architecture for a triBA node and Fig[2] emphasizes network aspects of a TriBA interconnection which is the focus of our design .We scratch the surface on the concepts used in our design and the interested reader is referred to more in depth references where appropriate . Just like common computer architectures out there, our

architecture is composed of computing modules, memory modules and the interconnection system to allow these two to communicate[18].

For triBA however special care was taken to separate computations from communication .It is composed of three submodules as shown in Fig[1] .ProcUnit carries out computations ,DataUnit is simply a chunk of read/write memory store our data and InterUnit ,the focus of our design, takes care of communications [1,18] .ProcUnit and DataUnit are abstracted away in our design to focus on network aspects of triBa and InterUnit is viewed as a node from here on .



Fig[1] TriBA Architecture



Fig[2] IDC132 addressed interconnected nodes

Each node is assigned an address .TriBA uses an addressing mechanism specifically designed for nodes in triplets ---IDC132 .It has impressive properties such as the reflexive symmetry of IDC132 addresses and the 120°rotation. These combined with the vertex distance computation help remarkably when computing the distance(hops) between nodes in our routing algorithms[19] .

Routing algorithms have been developed for TriBA , TDRA (Table Look up Deterministic Routing Algorithm) is one of them: when a node receives a message ,it has to decide if it is the recipient of the message or if it has to forward it to neighbouring nodes .When determining the route in TDRA , there is no need to store all the network information in the node ,and thus, the transmission overhead it might have generated is avoided[20] .

The algorithm uses two tables: a Channel Status Table (CST) that stores the working state of all the output ports of the node and a Route Table, that stores output port to be chosen for each destination node in the network, from the current node.

DDRA (Distributed Deterministic Routing Algorithm)[21] is another routing algorithms for TriBA .It has no routing table at all, the transfer of messages is carried out based on the inherent addressing properties of TriBA nodes. IDC 132 enforces locality, this allows the message to get directly to the destination node if it is local and only go across triplet boundaries when there is need to. IDC132 also allows telling the exact location of the node in the entire interconnection network just by looking at its address. The current version of TriBASim supports DDRA .Packet switching mechanisms were used in TriBASim and credit based flow control was implemented.

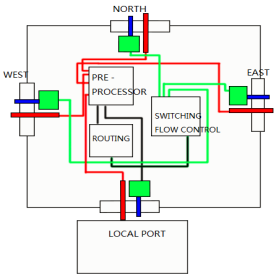
4. THE TRIBASIM ROUTER

TriBASim has been implemented in systemC , SystemC is a set of C++ classes and macros which provide an event-driven simulation interface in C++. These facilities enable a designer to simulate concurrent processes; each described using plain C++ syntax. SystemC processes can communicate in a simulated real-time environment, using signals of all the data types offered by C++, some additional ones offered by the SystemC library, as well as user defined. In certain

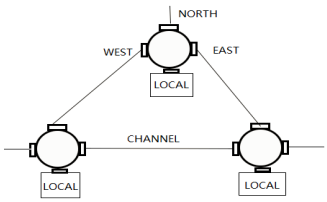
respects, SystemC deliberately mimics the hardware description languages VHDL and Verilog, but is more aptly described as a system-level modeling language[22].

ARCHITECTURE

Fig[3] shows a closer view to the InterUnit module of a typical node in TriBaSim. It comprises of 3 internal major sub modules: the pre-processor, the routing module and the switching and flow control module .We have 4 input output ports, 3 for communication with other nodes and one to interact with ProcUnit and DataUnit .The ports are color labeled for clarity .Red is for input ports which then pass the received data to the pre-processor sub module .Blue is for output ports and green squares represent our output buffers.



Fig[3] TriBASim Router Architecture



Fig[4] Node Ports interconnections

The ports are used to interface between noses .Links (channels) are used to connect nodes through ports .systemC provides convenience classes to implement ports and channels. sc_port< sc_fifo_in_if<sc_bv<64 > > > was used for input ports , sc_port< sc_fifo_out_if<sc_bv<64 > > > was used for output ports and sc_fifo <sc_bv<64>> was used for channels. Buffers implemented as fifos using the sc_fifo class have been designed to be on the output ports .The depths of the buffers can be set at the start of the simulation by passing appropriate parameters to TriBASim.

```
if( destAddr.3rdDoublet!=myAddr.3rdDoublet)
  if(11) sendNorth;exit;
  if(01) sendWest;exit;
  if(10) sendEast;exit;

if( destAddr.2ndDoublet!=myAddr.3rdDoublet)
  if(11) sendNorth;exit;
  if(01) sendWest;exit;
  if(10) sendEast;exit;

if( destAddr.firstDoublet!=myAddr.3rdDoublet)
  if(11) sendNorth;exit;
  if(01) sendWest;exit;
  if(10) sendEast;exit;

if(destAddr==myAddr)
  processMessage;
```

Fig[5]A Simplified Version of DDRA.

Upon reception of the packet the pre-processor checks whether the destination is the current node. The packet is passed to the local port if it is the case and passed to the routing module for destination port processing otherwise. Timings for sending and receiving overheads are also implemented in this sub module.

The routing module implements DDRA [21] .The algorithm enforces the principle of locality by sectioning IDC132 addresses into sections .This allows a level by level computation of the output port .A simplified version of the algorithm is shown in Fig[5] .

The information from the routing module is then passed to the switching and flow control module .Data is switched to the appropriate port through a simple virtual crossbar switch we have implemented. This module also manages our buffer space by making sure we write to the buffer when there is free space and read from it only when it is not empty .Our flow control is credit based.

We have followed the principle of incremental design; a tribaNode class was designed with addresses, buffers, ports and sub modules as data members and methods to implement node functionality such as sending and receiving data .Sub modules are themselves a set of C++ classes. With the node in place ,we designed a tribaTriplet class to take nodes and connects them in groups of three .The class only provides interface ports to connect to other triplets .The simulator can currently be configured to connect 3 ,9 and 27 nodes.

The latency computations involve sending and receiving overhead and the time of flight, these time values are based on experimental values .Combined with the packet transmission time which depends on the packet size and the link bandwidth, we can get the latency experienced when we send a packet through one link by the formula below. The channel bandwidth is set at the start of simulation when a user runs TriBASim .It is in orders of GBit/s.

$$Latency = Sending\ overhead + Time\ of\ flight + \frac{Packet\ size}{Bandwidth} + Receiving\ overhead$$

POWER AND AREA COMPUTATIONS

Orion has been used to do power and area analysis in our simulator ,Orion is a power-performance interconnection network simulator that is capable of providing detailed power characteristics, in addition to performance characteristics, to enable rapid power-performance tradeoffs at the architectural-level [23] .Orion power models are based on real characteristics of hardware composing the interconnect like buffers ,gates and wires .

Considering a flit traversing our router, the total flit power can be computed as follows:

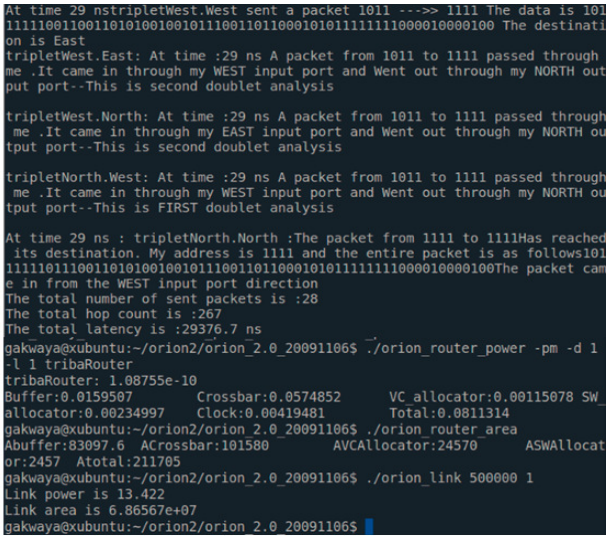
$$E_{flit} = E_{wrt} + E_{arb} + E_{xb} + E_{link}$$

Where E_wrt is the power dissipated when writing to the buffer, E_arb the power dissipated on arbitration, E_xb the power dissipated on switching and E_link the power dissipated on the link .From a user perspective, all we needed to specify was the parameters for the components of the interconnect and Orion provided the results based on the data they have collected. [23] Has more on the details of how power is modelled.

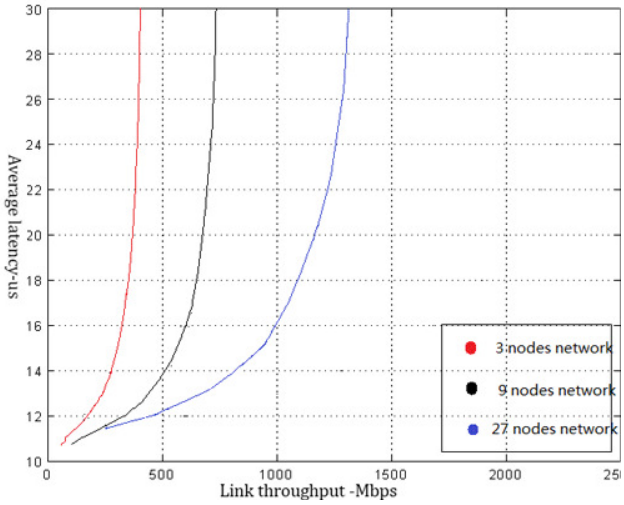
5. CASE STUDIES

The figures below show a set of simulations we run with triBASim .In Fig[6] we traced the path followed by a packet from source to destination logging port information on each intermediary node .Critical network information can be easily obtained by activating convenience methods on node and triplet classes .In Fig[7] we studied how average latency in the network varies per link throughput per number of nodes Results show that lower level triBA networks saturate earlier

than their higher level counterparts. We have used the same link area and power configuration on entire networks but networks with different configurations can also be studied.



Fig[6] Packet tracing Simulation within TriBASim



Fig[7] Delay-latency-node count analysis within TriBASim

6. FUTURE WORK

TriBASim can already run the common chores that Network On Chip simulators are supposed to run .We hope to add support for multiple routing algorithms other than DDRA .The simulations we have run are based on random traffic models .We hope to delve into studying the characteristics of the traffics patters for our in-house SoCs and incorporate them in future versions.

7. CONCLUSIONS

A new simulator for the Triplet Based NoC architecture has been suggested .We went through a broad overview of TriBA and displayed its basic characteristics and state of the art .Furthermore, we described the details for the design of our simulator and ended the paper with practical uses showing its usefulness to the triBA researcher and anyone interested in NoCs in general.

REFERENCES

[1] 一种新的非冯•诺依曼计算机体系结构 TriBA ,石峰, 计卫星, 乔保军, 刘滨,北京理工大学学报,Vol .26 No.10 Oct 2006 ; A New Non Von Neumann Architecture TriBA ,SHI Feng , J I Wei2xing , QIAO Bao2jun , L IU Bin , Transactions of Beijing Institute of Technology , Vol .26 No.10 Oct 2006 .

[2] TriBA 互联拓扑结构及其性能分析 ,刘彩霞，石 峰，乔保军，HAROON Ur Rashid，宋 红 ,北京理工大学学报 ，计 算 机 工 程 Vol .36 No .15 ; TriBA Interconnection Topology Structure and Its Performance Analysis ,LIU Cai-xia, SHI Feng, QIAO Bao-jun, HAROON Ur Rashid, SONG Hong ,Journal of Beijing Institute of Technology ,Computer Engineering , Vol .36 No .15.

[3] 基三分层网络中一种基于查表的确定路由算法 ,乔保军, 石峰, 计卫星, 刘 滨 , 北京理工大学学报，计算机应用 Vol . 26 No .9 ,Table lookup determined routing algorithm for triplet based hierarchical interconnection network QIAO Bao jun, SHI Feng, JI Wei xing, LIU Bin,Journal of Beijing Institute of Technology ,Computer Applications ,Vol .26 No .9

[4] 基三分层互连网络及其路由算法设计 ,计算机工程与设计,乔保军，石峰， 计卫星,北京理工大学学报 ，计算机工程与设计，Vol .28 No .18 ；Triplet-based hierarchical interconnection network and design of its routing algorithm,QIAO Bao-jun, SHI Feng, JIWei-xing ,Journal of Beijing Institute of Technology ,Computer Engineering and Design ,Vol .28 No .18.

[5] From “Bus” and “Crossbar” to “Network-On-Chip”,Arteris S.A.Copyright 2009 Arteris S.A. All rights reserved.

[6] A generic architecture for on-chip packet-switched interconnections ,Guerrier, P.Greiner, A;Univ. Pierre et Marie Curie, Paris, France Design, Automation and Test in Europe Conference and Exhibition 2000. Proceedings

[7] A Router Architecture for Networks on Silicon ,Edwin Rijpkema, Kees Goossens, and Paul Wielage ,Philips Research LaboratorieProf. Holstlaan 4, 5656 AA Eindhoven, The Netherlands, Proceedings of progress 2001, 2nd workshop on embedded systems

[8] Networks on Chip A New Paradigm for Systems on Chip Design ,Luca Benini ,Giovanni De Micheli ,IEEE ,System-on-Chip, 2005. Proceedings. 2005 International Symposium on System On Chip ,17-17 Nov.Page(s) 2 – 6

[9] Survey of Network on Chip (NoC) Architectures & Contributions,Ankur Agarwal,Cyril Iskander,Ravi Shankar,Journal of Engineering,Computing and Architecture ISSN 1934-7197 Volume 3,Issue 1,2009

[10] Locality Aware Optimal Task Scheduling Algorithm for TriBA A Novel Scalable Architecture,KHAN Haroon-U r-Rashid,SHI Feng(石峰), Jour nal of Beijing Institute of Technology , 2008, Vol. 17, No. 3

[11] ORION 2.0: A Fastand Accurate NoC Power and Area Model for Early-Stage Design Space Exploration,Andrew B. Kahng, Bin Li, Li-Shiuan Peh and Kambiz Samadi, Proceedings of Design Automation and Test in Europe (DATE), Nice, France, April 2009.

[12] Orion: A Power-Performance Simulator for Interconnection Networks ,Hang-Sheng Wang Xinping Zhu Li-Shiuan Peh Sharad Malik,Proceedings of MICRO 35, Istanbul, Turkey, November 2002

[13] http://www.noxim.org/

[14] Enhanced Noxim simulator for performance evaluation of network on chip topologies,Swaminathan, K.,Thakyal, D. ; Nambiar, S.G. ; Lakshminarayanan, G. ; Seok-Bum Ko,Engineering and Computational Sciences (RAECS), 2014 Recent Advances,978-1-4799-2290-1

[15] open source simulator for network on chip ,Monika Gupta1, S. R. Biradar2, B. P. Singh,International Journal of Computers & Technology ,Volume 4 No. 2, March-April, 2013, ISSN 2277-3061

[16] A High Level Power Model for the Nostrum NoC , Penolazzi, S, Jantsch, A. ,Digital System Design: Architectures, Methods and Tools, 2006. DSD 2006. 9th EUROMICRO Conference

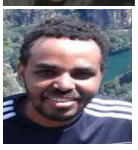
[17] http://networkonchip.wordpress.com/2011/02/22/simulators/
[18] A New Non Von Neumann Architecture TriBA ,SHI Feng , J I Wei2xing , QIAO Bao2jun , L IU Bin ,Transactions of Beijing Institute of Te,Vol. 26 No. 10 ,Oct . 2006
[19] 基三分层互连网络和 2D-Mesh 的比较 ,北京理工大学学报 ， 计算机科学 2007 Vol. 34 No.9 ;Comparison of the Triplet-based Hierarchical Interconnection Network and 2-D Mesh for Multi-core Processor ,Journal of Beijing Institute of Technology ,2007 ,Vol .34 No .9
[20] 基三分层网络中一种基于查表的确定路由算法 ,乔保军, 石峰, 计卫星, 刘 滨 , 北京理工大学学报， 计算机应用 Vol . 26 No .9 ,Table lookup determined routing algorithm for triplet based hierarchical interconnection network QIAO Bao jun, SHI Feng, JI Wei xing, LIU Bin,Journal of Beijing Institute of Technology ,Computer Applications ,Vol .26 No .9
[21] 基三分层互连网络及其路由算法设计 ,计算机工程与设计,乔保军， 石峰， 计卫星,北京理工大学学报 ， 计算机工程与设计， Vol .28 No .18 ；Triplet-based hierarchical interconnection network and design of its routing algorithm,QIAO Bao-jun, SHI Feng, JIWei-xing ,Journal of Beijing Institute of Technology ,Computer Engineering and Design ,Vol .28 No .18.
[22] SystemC: From the Ground Up, Second Edition , David C. Black (Author), Jack Donovan (Author), Bill Bunton (Author), Anna Keist (Author) Springer; 2nd edition (December 30, 2010),ISBN-10: 0387699570 ISBN-13: 978-0387699578
[23] Orion: A Power-Performance Simulator for Interconnection Networks ,Hang-Sheng Wang Xinping Zhu Li-Shiuan Peh Sharad Malik,Microarchitecture, 2002. (MICRO-35). Proceedings. 35th Annual IEEE/ACM International Symposiu,0-7695-1859-1

AUTHORS

Daniel GAKWAYA
A student at BEIJING INSTITUTE OF TECHNOLOGY currently pursuing his master's degree,School of Computer Science ,Department of Advanced Embedded Computing . His research interests lie in Network Optimizations and Computer Graphics.

Gao Yu Jin
Associate Professor, BEIJING INSTITUTE OF TECHNOLOGY , School of Computer Science, Department of Advanced Embedded Computing ., His research interest include Embedded Multicore Processors.

Jean Claude GOMBANIRO
Master's student at the School of Computer Science Of BEIJING INSTITUTE OF TECHNOLOGY, Department of Natural Languages Processing His research interests lie in Big Data Processing and Language recognition Algorithms.



Jean Pierre NIYIGENA
Master's student at the School of Computer Science Of BEIJING INSTITUTE OF TECHNOLOGY, Department of Data Networks .His research interests lie in Geolocation Algorithms.

STUDY OF FACTORS AFFECTING CUSTOMER BEHAVIOUR USING BIG DATA TECHNOLOGY

Prabin Sahoo[1], Dr. Nilay Yajnik[2]

[1,2]SVKM'S NMIMS, Deemed to be University, Mumbai
[1]prabins@hotmail.com, [2]nilayy@nmims.edu

ABSTRACT

Big data technology is getting momentum recently. There are several articles, books, blogs and discussion points to various facets of big data technology. The study in this paper focuses on big data as concept, and insights into 3 Vs such as Volume, Velocity and Variety and demonstrates their significance with respect to factors that can be processed using big data for studying customer behaviour for online users.

KEYWORDS

Big data, customer behaviour, customer activities, Volume, Varieties, Quantitative research

1. INTRODUCTION

The internet is the phenomenal innovation of the information system era. With billions of internet users across the globe generate huge amount of data like never before. There are millions of web sites which cater to various demands of its users. E-commerce, search engines, online shopping, banking, trading etc. are all accessible from any parts of the world. This has brought a new dimension to the user behaviour pattern. For example, there are multiple airline sites through which one can book an air ticket to any destination. While these facilities make the consumer life simpler, but creates a competitive environment for the service providers. Since there are multiple options available for customers, it is very difficult to sustain the customer base as with a slightest inconvenience can force customer to switch to a different service provider. Therefore, it is utmost important for the service providers to understand the customer demand, choices, preferences etc. and provide them high quality of services. So the challenge is how to find out these factors? Traditionally, in a buyer and supplier relationship, the customer visits a retail store; the supplier interacts with the customer and may request the buyer to fill a survey from. The information collected can be fed into a database, and can be processed to find out the liking and disliking factors. In such scenario, the supplier depends solely on the response rather the mercy of the consumer if he or she likes to fill the survey form with all integrity. For online shopping, there are sites which ask for filling an online survey, but that is again if the buyer wants to fill it or not. Therefore, a mechanism is required through which information about the customer can be processed to understand the behaviour of customer. Since each activity on the internet is recorded, so when a customer visits a site can be tracked. However, the log generated out of each visit is enormous as it accumulates over a period of time. The logs contain unstructured data which require huge amount of efforts for extraction, parsing and finally loading into a database. The entire process is very time consuming and therefore the traditional system fails. Big data is the technology which solves the problem of processing huge volume of data. It has primarily 3

Natarajan Meghanathan et al. (Eds) : CSEN, ADCO - 2014
pp. 31–39, 2014. © CS & IT-CSCP 2014 DOI : 10.5121/csit.2014.4104

characteristics such as Volume, Varieties, and Velocity. This paper focuses on these characteristics to find out its significance on the behavioural attributes of online customers. A survey has been conducted from the big data users and a model has been developed to process users' response with established theory.

2. LITERATURE REVIEW

There are theories on studying customer behaviours (Chung-Hoon Park and Young-Gul Kim, 2003); (Susan M. Keaveney, Madhavan Parthasarathy,2001); (Limayem, Moez ; Dept. of Inf. Syst., City Univ. of Hong Kong, Kowloon, China ; Khalifa, M. ; Frini, A., 2000). These studies discusses about the factors/attributes of online customer.

Information satisfaction (Chung-Hoon Park and Young-Gul Kim, 2003) is closely related to product information quality, service information quality and user interface quality. This theory is useful for the big data research under consideration, as the purpose is to establish the outcome from big data analysis to behavioural factors.

Customer switching behaviour (Susan M. Keaveney, Madhavan Parthasarathy, 2001) theory has observed that online service usages are more with continuers than the switchers. This theory is useful for big data as it can identify the customers who are using frequently the online services. Further the theory also indicates that propensity of risk taking behaviour in buying a new product is more in case of service continuers. This will help in introducing new product/services on the most frequent product/service line of business.

Big data technology is based on 3 key concepts i.e. volume, variety and velocity (Philip Carter, 2011; Ramesh Nair, Andy Narayanan, 2012). It can handle huge data volume with structured, unstructured format. The logs from web are mix of unstructured or semi structured and structured in nature which can be processed using big data technology.

Volume: Volume is the amount of data that can be processed. Though volume represents huge volume in case of big data, but it is a relative term to be defined precisely.

Velocity: The demand to analyse the data in real time (Philip Carter, 2011; Ramesh Nair, Andy Narayanan, 2012), speed of data in and out (Wikipedia, retrieved 2014)

Variety: Data can be of structured, semi structured and unstructured in nature. Structured data can have delimiter to separate various columns. For example, "customer name, customer id, address, item, quantity" represent a structured format where the fields are separated by comma. Unstructured data can be of various types such as email, blog, twitter logs whereas semi structured data can be a combination of both i.e. structured and unstructured.

3. MODEL, CONCEPTS

The proposed model figure 2.0 uses theory, survey (Primary data), statistical model and big data. The theory is based on quality of service/product and the service usages. Quantitative research methodologies have been used in building the model.

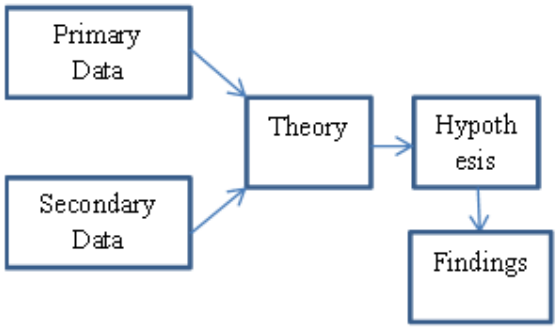


Figure 1. Quantitative research to establish Big Data factors for behaviour prediction

3.1. Quantitative research

Quantitative research is required to establish numerically that big data is useful in predicting the behaviour of customers. Survey has been conducted with the big data users to gather big data usages for e-commerce through questionnaire sessions. Participants were selected from Big Data users from e-commerce industries. Using statistical analysis hypotheses have been evaluated using IBM SPSS©.

3.2. Dependent Variable vs. Independent Variable

In table 1.0 the dependent and independent variables have been shown. The goal is to evaluate the impact of independent variables on the dependent variables.

Table 1. Independent vs. Dependent variables

Independent Variable	**Dependent Variable**
Transaction information	i) Service Usages ii) Quality of Service
Feedback	
Browsing information	

3.3. Theory

3.3.1 Quality of product/service

Quality of service or product (Valarie A. Zeithaml, A. Parasuraman, Arvind Malhotra, 2002) has been discussed as an important aspect in the web based services. For example, the transaction on a particular service is less because of less demand of the service or may be that the web service quality is not up to the mark. Therefore, to understand the root cause it is important to analyse the quality of service. If the quality is good, then the numbers of transactions decide the demand of the service or the product under study. Quality of service is further supported by the theory of quality of product information (Chung-Hoon Park and Young-Gul Kim, 2003).

3.3.2 Service Usages

Online service usages (Susan M. Keaveney, Madhavan Parthasarathy, 2001) theory indicates that the online service usage is one of the factors that determine the customer switching behaviour.

This is crucial in favour of online usage and therefore use of big data in studying customer behaviour. With higher usages of online services, high volume of logs will be produced which can be processed using big data to find out the customer access pattern on a service or product.

3.4 Model

In this model key factors such as blogs, surfing patterns, and transaction information are being fed into big data. The data format is unstructured and the volume is huge which cannot be processed through traditional data processing tools. Using Big Data these factors are tested against the theory of i) Quality of service and ii) Service usages.

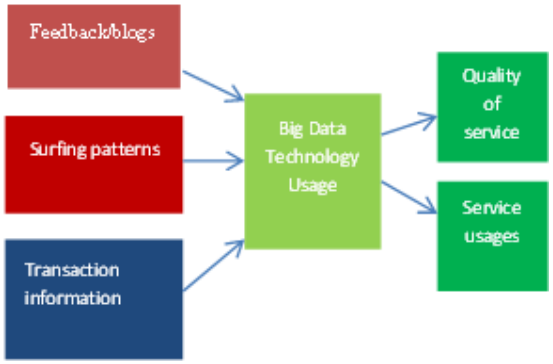


Figure 2. Model for studying key factors of customer behaviour using big data

3.4 Concepts

3.4.1 Feedback through blogs:

The social media has enabled a new medium through which customers express their opinion about a product or service. If a service/product does not meet the user expectations, users do not hesitate to express it over blogs. The data out of blogs are unstructured and the volume is huge.

Surfing patterns:

Understanding surfing patterns is one of the major factor for improving the quality of service. In big data context, big data should be able to identify the demands of a particular service. For example, if a customer is searching a travel destination and how frequently and how many customers are looking for the particular destination can predict the preference of customer. Service demand is discussed (A. Dan et al., 2004).

3.4.2 Transactions:

Online transactions are very common. Customers buy and sell using online web sites. Enormous amount of data is generated out of these transactions. Big data has the capabilities to process these data.

3.4.3 One Sample T Test:

One sample t test has been chosen in this model as the full population information is not available and to make sure the sample selected comes from a particular population. Big Data is evolving

and it is new in the industry. It is having its own class. Therefore one sample T test is justifiable to determine that the sample is selected from a population of known mean (μ).

3.4.4 Customer behaviour prediction:

Using big data technology if the quality of service and service usages are identified, then determining the behavioural aspect of customer can be established.

4. HYPOTHESIS

H. Big Data enables studying Quality of services by its ability to process blogs which is unstructured in nature.

H1. Big Data enables studying Quality of services by enabling processing of huge volume of data.
H2. Big Data enables studying Quality of services by enabling processing of huge volume of data in less time.
H3. Big Data enables studying Quality of services by enabling processing of varieties of data

H. Big Data enables studying Service Usages by its ability to process browsing (surfing) pattern

H1. Big Data enables studying service usages by enabling processing of huge volume of data.
H2. Big Data enables studying service usages by enabling processing of huge volume of data in less time.
H3. Big Data enables studying service usages by enabling processing of varieties of data

H. Big Data enables studying by its ability to process transactional patterns

H1. Big Data enables studying transactional patterns by enabling processing of huge volume of data.
H2. Big Data enables transactional patterns by enabling processing of huge volume of data in less time.
H3. Big Data enables transactional patterns by enabling processing of varieties of data

5. SURVEY

QUESTIONNAIRE:

a) I use big data for processing blogs to study customers' feedback on product/services.
   1) Strongly Disagree 2) Disagree 3) somewhat agree 4) Agree 5) Strongly agree
b) I use big data because it helps in processing huge volume of data which is not possible using traditional system
   1) Strongly Disagree 2) Disagree 3) somewhat agree 4) Agree 5) Strongly agree
c) I use big data because it helps in processing data faster than traditional approach.
   1) Strongly Disagree 2) Disagree 3) somewhat agree 4) Agree 5) Strongly agree
d) I use big data because it helps in processing varieties of data such unstructured, structured, semi structured.
   1) Strongly Disagree 2) Disagree 3) somewhat agree 4) Agree 5) Strongly agree
e) I use big data to study quality of services
   1) Strongly Disagree 2) Disagree 3) somewhat agree 4) Agree 5) Strongly agree

f) I use big data to study service usages
   1) Strongly Disagree 2) Disagree 3) somewhat agree 4) Agree 5) Strongly agree
g) I use big data to study customers' browsing behaviour
   1) Strongly Disagree 2) Disagree 3) somewhat agree 4) Agree 5) Strongly agree
h) I use big data to study transactional data which provides insights about service usages.
   1) Strongly Disagree 2) Disagree 3) somewhat agree 4) Agree 5) Strongly agree

6. FINDINGS

6.1 BLOGS -> QUALITY OF SERVICE

One-Sample Statistics

	N	Mean	Std. Deviation	Std. Error Mean
blogs	48	4.48	.545	.079

One-Sample Test

	Test Value = 4					
					95% Confidence Interval of the Difference	
	t	df	Sig. (2-tailed)	Mean Difference	Lower	Upper
blogs	6.087	47	.000	.479	.32	.64

Figure 3. One sample Test

Figure 3 shows that blogs processing has mean of 4.4 which implies that more respondents use big data to process blogs to study customers' feedback on product/services, this is further supported statistically that with 95% confidence interval blogs are processed by big data. Therefore the null hypothesis that big data is used to process blogs is accepted.

Paired Samples Correlations

		N	Correlation	Sig.
Pair 1	blogs & qos	48	.377	.008

Figure 4. Paired sample correlations

blogs * qos Crosstabulation

Count

		qos			Total
		2	4	5	
blogs	3	0	1	0	1
	4	1	16	6	23
	5	0	9	15	24
Total		1	26	21	48

Symmetric Measures

		Value	Asymp. Std. Error[a]	Approx. T[b]	Approx. Sig.
Ordinal by Ordinal	Gamma	.676	.164	3.076	.002
N of Valid Cases		48			

a. Not assuming the null hypothesis.

b. Using the asymptotic standard error assuming the null hypothesis.

Figure 5. Paired sample correlation and Gamma Testing

Figure 4 and figure 5 show that blogs and qos are related at a level of significance 0.05 using paired sample correlations and Gamma testing. This implies that with 95% confidence interval blogs analysis determines quality of services. This is because customers write their opinion, experience on a product or service using social media through blogs.

6.2 BROWSING PATTERNS -> SERVICE USAGES

Symmetric Measures

		Value	Asymp. Std. Error[a]	Approx. T[b]	Approx. Sig.
Ordinal by Ordinal	Gamma	.045	.265	.169	.866
N of Valid Cases		48			

a. Not assuming the null hypothesis.

b. Using the asymptotic standard error assuming the null hypothesis.

Figure 6. Gamma Testing on browsing pattern vs. service usages

Figure 6 shows that with 95% confidence interval browsing patterns do not influence service usages. Therefore the null hypothesis such as browsing pattern is significantly related to service usages is rejected.

6.3 TRANSCATIONS -> SERVICE USAGES

Symmetric Measures

		Value	Asymp. Std. Error[a]	Approx. T[b]	Approx. Sig.
Ordinal by Ordinal	Gamma	.353	.222	1.529	.126
N of Valid Cases		47			

a. Not assuming the null hypothesis.

b. Using the asymptotic standard error assuming the null hypothesis.

Figure 7. Gamma Testing on transaction information vs. service usages

Figure 7 shows that with 95% confidence interval transaction information do not influence service usages. Therefore the null hypothesis such as transactional data is significantly related to service usages is rejected.

One-Sample Statistics

	N	Mean	Std. Deviation	Std. Error Mean
transacations	47	3.81	.900	.131

One-Sample Test

	Test Value = 4					
					95% Confidence Interval of the Difference	
	t	df	Sig. (2-tailed)	Mean Difference	Lower	Upper
transacations	-1.458	46	.152	-.191	-.46	.07

Figure 8. One sample test of transaction

Figure 8 shows that with 95% confidence interval transaction data have no significance for big data processing. The mean of 3.8 indicates transactions are somewhat processed using big data which implies that transactional data used occasionally.

6.4 BROWSING PATTERNS -> QUALITY OF SERVICE/PRODUCT

Symmetric Measures

		Value	Asymp. Std. Error[a]	Approx. T[b]	Approx. Sig.
Ordinal by Ordinal	Gamma	1.000	.000	20.993	.000
N of Valid Cases		48			

a. Not assuming the null hypothesis.

b. Using the asymptotic standard error assuming the null hypothesis.

Figure 9. Gamma test Browsing pattern vs. quality of service/product

Figure 9 shows that with 95% confidence interval browsing pattern is significant for big data processing to determine quality of service/product.

6.5 VOLUME, VELOCITY, VARIETIES

One-Sample Test

	Test Value = 4					
					95% Confidence Interval of the Difference	
	t	df	Sig. (2-tailed)	Mean Difference	Lower	Upper
volume	-3.517	47	.001	-.417	-.66	-.18
velocity	-.405	47	.688	-.042	-.25	.17
varieties	.000	47	1.000	.000	-.21	.21

Figure 10. One-Sample Test. (volume, velocity, varieties)

Figure 10 shows that volume is having significance in big data processing with 95% confidence interval. Varieties and Velocity are not statistically significant with 95% confidence interval as big data is mostly used for unstructured data processing and velocity from users' perspective is

not that significant probably due to the fact that big data mostly used in batch processing which is slow compared to online data processing, basically batch processing would take hours vs. online processing would take seconds.

7. CONCLUSIONS

Big data plays a significant role in processing online data from internet. Data collected from internet source, social media logs are huge which is limited by the processing power of traditional tools. From the statistical findings it is evident that big data is meant for processing huge volume of data thus capable of eliminating the limitation of traditional tools. Further unstructured data from internet and social media logs such as blogs, browsing patterns found to be mostly used for big data processing. Velocity, Varieties were not found to be as important as volume. This may be because Velocity is implicit to processing huge volume of data. For example, processing 100 TB data can be processed using traditional tool in several days that implicitly tells that huge volume is a challenge because the processing time is more. Logs from social media, internet surfing are primarily unstructured in nature and there is evidence that blogs, browsing behaviour are significant in big data processing, therefore Varieties has not been significantly felt. But since big data is a data processing framework, varieties of data format can be processed in this. For online customers, big data can process to find quality of services/product from blogs, and browsing patterns. Transactional data was not found that significant for online users using big data.

ACKNOWLEDGEMENT

My sincere thanks to Dr. Nilay Yajnik, Head of Information Technology, SVKM's NMIMS, MUMBAI for his guidance.

REFERENCES

[1] Chung-Hoon Park and Young-Gul Kim (2003), "Identifying key factors affecting consumer purchase behavior in an online shopping context", International Journal of Retail & Distribution Management, Volume 31. Number 1, 2003 . pp. 16-29 # MCB UP Limited. ISSN 0959-0552

[2] Susan M. Keaveney, Madhavan Parthasarathy (2001) "Customer Switching Behavior in Online Services" ,Academy of Marketing Science, 29, 4; ProQuest CentralLIMAYEM MOEZ , KONG KOWLOON, KHALIFA, M. , FRINI, A. (2000), WHAT MAKES CONSUMERS BUY FROM INTERNET? A LONGITUDINAL STUDY OF ONLINE SHOPPING, SYSTEMS, MAN AND CYBERNETICS, PART A: SYSTEMS AND HUMANS, IEEE TRANSACTIONS ON (VOLUME:30 , ISSUE: 4)

[3] Valarie A. Zeithaml,A. Parasuraman,Arvind Malhotra (2002), "Service Quality Delivery Through Web Sites: A Critical" Review of Extant Knowledge, Journal of the Academy of Marketing Science. Volume 30, No. 4, pages 358-371

[4] Philip Carter. (2011), "Big Data analytics: Future architectures, Skills and roadmaps for the CiO", http://www.sas.com/resources/asset/BigDataAnalytics-FutureArchitectures-Skills-RoadmapsfortheCIO.pdf, retrieved 2012.

[5] Ramesh Nair,Andy Narayanan (2012), "Benefitting from Big Data Leveraging Unstructured Data Capabilities for Competitive Advantage", Data and TechnologyPerspective,booz&co.

[6] T. Ramayah, Jasman J. Ma'ruf, Muhamad Jantan, Osman Mohamad (2002), "TECHNOLOGY ACCEPTANCE MODEL: IS IT APPLICABLE TO USERS AND NON USERS OF INTERNET BANKING", The proceedings of The International Seminar, Indonesia-Malaysia, 14-15th October 2002

[7] Sara Abbaspour Asadollah, Thiam Kian Chiew, "Web Service Response Time Monitoring: Architecture and Vaidation", Q.Zhou(Ed.):ICTMF 2011, CCIS 164, © Springer-Verlag Berlin Heidelberg 2011

[8] A. Dan et al., "Web services on demand: WSLA-driven Automated management", © 2004, IBM SYSTEMS JOURNAL, VOL 43, NO 1, 2004

[9] Wikipedia., "Big Data", http://en.wikipedia.org/wiki/Big_data, retrieved July 2014

INTENTIONAL BLANK

MEASURING THE EFECTIVENESS OF TEST CASE PRIORITIZATION TECHNIQUES BASED ON WEIGHT FACTORS

Thillaikarasi Muthusamy and Dr. Seetharaman.K

[1]Department of computer science,
Annamalai University, Annamalai Nagar, Tamilnadu, India
selva.thillai69@gmail.com
kseethadde@yahoo.com

ABSTRACT

Test case prioritization schedule test cases in an order that increases the success in achieving some performance target. The most important target is ,at what rate the fault is detected. Test cases should run in an order that increases the opportunity of fault exposure and as well detecting the most rigorous faults, most primitively in its testing life cycle. Test case prioritization techniques have proved to be advantageous for improving regression testing activities. whereas code coverage based prioritization are being studied by most scholars, hitherto test case prioritization techniques based on requirements in cost effective manner has not been analyzed. Here we suggest to put forth a model for system level test case prioritization from software requirement specification and to develop user fulfillment with quality software that can also be cost effective.Thus improving the rate of severe fault detection. The projected model priorities the system test cases, based on six factors. They are customer allotted priority, developer observed code implementation complexity, changes in requirements, fault impact of requirements, completeness and Traceability. The anticipated prioritization techniques is experimented with two set of industrial projects. The results realistically show that proposed prioritization techniques improves the rate of fault detection .

KEYWORDS

Regression Testing, Test case prioritization, Fault severity, Rate of fault detection.

1. INTRODUCTION

Software regression testing is an activity which includes enhancements, error corrections, optimization and deletion of existing features. These modifications may cause the system to work improperly. Hence, Regression Testing becomes essential in software testing process. this leads to regression testing in which all the tests in the accessible programmes or suite should be re-executed. Thus incurring excess cost, time and resources. Test case prioritization is an important

Natarajan Meghanathan et al. (Eds) : CSEN, ADCO - 2014
pp. 41–51, 2014. © CS & IT-CSCP 2014 DOI : 10.5121/csit.2014.4105

technique adopted in regression testing. Prioritize the test cases depending on business impact, importance & frequently used functionalities. Selection of test cases based on priority will significantly reduce the regression test suite. Here we suggest a new approach for test case prioritization for prior fault detection in the regression testing process.

Here, we have projected a new approach to test case prioritization for quick fault detection based on practical influences. We have implemented the proposed technique in a banking application project and effectiveness is calculated by using APFD metric.

2. TECHNIQUES REVISITED

This segment narrates about the test case prioritization techniques to be used in our empirical study are as follows:

Test case prioritization is an important regression testing technique, which approaches typically and restructure existing test cases for regression testing accordingly to achieve targets.

Badhera et al.[1] presented a technique to execute the modified lines of code with minimum number of test cases. The test case prioritization technique organizes the test case in a test suite in an order such that fewer lines of code need to be executed. Thus faster code coverage is attained which will lead to early detection of faults. Bixin Li et al.(2012) proposed an automatic test case selection for regression testing of composite service, based on extensible BPEL flow graph.

B. Jiang et al. [2] projected an ART-based prioritization method which uses the algorithm and accepts the test suite as input, produces the output in prioritized order. The basic idea is about building the candidate set of test cases, which in turn picks one test case from the candidate set until all test cases have been selected. Here two functions are used in this algorithm for calculating the distance between a pair of test cases and also to select a test case from the candidate set. Calculation of distance is determined by code coverage data. Then we find a candidate test case which is related with the distance test cases that has been prioritized earlier.

Dr. ArvinderKaur and ShubhraGoyal [3] developed a new genetic algorithm and prioritize regression test suite within a time constrained environment on the basis of entire fault coverage. This algorithm is automated and the results are analyzed with help of Average Percentage of Faults Detected (APFD).

Hong Mei et al. [4] proposed a new approach for prioritizing test cases in the absence of coverage information which is widely used in java programs under the JUnit framework. A new approach called JUPTA which operates in the absence of coverage information and analyzes the static call graphs of JUnit test cases. Further, it estimates the ability of each test case to achieve code coverage and schedules the test cases in an order based on those estimates.

H.Do et al. [5] presented the importance of time constraints on test case prioritization and discovered that constraints which alters the performance of technique. Further, conducted three set of experiments which reveals the time constraints. The outcome show that the time constraint factor play a significant role in determining the cost effectiveness and cost benefit trade-offs

among the techniques. Next experiment reproduces the first experiment, calculating several threats to validate numbers of faults present. Third experiment manipulates the number of faults present in programs to examine the effect of inaccuracy on prioritization and exhibits the relative cost-effectiveness of prioritization techniques.

Park et al. [6] introduced a cost awareness model for the test case prioritization and fault severities which revealed in the previous test execution. As well as it does not significantly change form one outcome to another. Mohamed A Shameem et al. (2013) presented a metric for assessing the rate of fault detection. This algorithm identifies the faults in prior and the effectiveness of prioritized test cases are compared with the non prioritized cases by Average Percentage Of Fault Detection (APFD).

M. Yoon et al. [7] proposed a method to prioritize new test cases by estimating the requirements of risk exposure value and also analyzing risk objects. Further it calculates the relevant test cases and thereby determining the test case priority through the evaluated values. Moreover, we demonstrate the effectiveness of our technique through empirical studies in terms of Average Percentage Of Fault Detected (APFD) and fault severity.

R. Abreu et al. [8] projected a Spectrum-based multiple fault localization method to find out the fault location apparently. R. Bryce et al. (2011) suggested a model which describes prioritization criteria for GUI and web applications in an event driven software. The ultimate purpose is to evolve the model and to develop a unified theory about testing of EDS.

R. Krishnamoorthi and S. A. Mary [9] presented a model that prioritizes the system test cases based on six factors: customer priority, changes in requirement, implementation complexity, usability, application flow and fault impact. This prioritization technique is examined in three phases with student projects and two sets of industrial projects. Here results were found to improve the rate of severe fault detection

S. Raju and G.V. Uma [10] initiated a cluster-based test case prioritization technique. Here, the test cases are clustered based on their dynamic runtime behavior. Significantly researchers reduced the required number of pair-wise comparisons. Researchers presented a value-driven approach to system-level test case prioritization which prioritizes the requirements for test. Here, prioritization of test cases is based on four factors: rate of fault detection, requirements volatility, fault impact and implementation complexity.

The rest of this paper is organized as follows. Section three discusses about the proposed work. Section four discusses about the experimental results and analysis. Section five discusses about the test cases to be prioritized. Finally, section six consists of conclusion. References are given in last section.

3. PROPOSED WORK

This section, briefly discusses about the prioritization factors.

3.1 Prioritization Weight Factors

It deals with, computation of prioritization factors such as (1) customer allotted priority , (2) developer observed code execution complexity, (3) changes in requirements, (4) fault impact (5) completeness and (6) traceability which is essential for prioritizing the test cases since they are used in the prioritization algorithm. Weights are assigned to each test case in the software testing according to the factors. Then, test cases are prioritized based on the weights assigned.

3.1.1 Customer-Allotted Priority (CP)

It determines the requirements of the customer and the value are assigned by the customers. The values vary from 1 to 20, where 20 are used to identify the highest customer priority. So, improving customer's fulfillment imposes, initial testing of the highest priority needs of the customer. Greater effort should be taken in identifying faults and their impacts on the execution path of program as these faults results in repeated failures. It has been proved that customer-Allotted value and satisfaction can be improved by fixing on customer needs for development.

3.1.2 Developer-observed Code Implementation Complexity(IC)

It is an individual measure of complexity expected by the development team to implement the requirements. First every necessity is evaluated by assigning a value from 1 to 20. Based on the implementation complexity, the higher complexity is implied by a larger value. Large number of faults that occurs in a requirement has high implementation complexity.

3.1.3 Changes in Requirements (RC)

It is a degree assigned by the developer in the range of 1 to 20 which indicates that the requirement is changed as many times during the development cycle with respect to its origin. The volatility values for all the needs are expressed on a 20-point scale where the need is altered more than 20 times. The number of changes for any requirement 'i' is divided to the highest number of changes which in turn yields the change in requirement *Ri* where the requirement is 'i'. If the ith requirement is changed M times and N is the maximum number of requirements, then the requirement change *Ri* can be calculated as follows:

$$Ri = (M/N) \times 10 \qquad (1)$$

The errors in the requirement level are approximated to 50% of all faults detected in the project. The change in requirements is the major factor that features the failure of the project.

3.1.4 Fault Impact of Requirements (FI)

It allows the development team to differentiate the requirements that had customer reported failures. Developers can recognize requirements that are expected to be error free by using the prior data collected from older versions since system evolves to several versions. The number of in-house failures and field failures determine the fault impact of requirements. It is a measure for released product. It is proved that field failures are more likely to be fault prone modules than modules that are not fault prone.

3.1.5 Completeness (CT)

This part indicates requirement based function to be executed, the rate of success, the limitations and any limitation which manipulate the expected solution. (boundary constraints). The consumer assigns value from 1 to 20. When the condition is selected for reuse by scrutinizing the completeness of each requirement into consideration, customer satisfaction can be enhanced.

3.1.6 Traceability (TR)

Relation between requirement and assessment can be calibrated by means of Traceability. If the test cases are not concerned to individual requirement, the common problem reported is scarcity of traceability. Hence poor traceability leads to failure and going beyond the desired limit of the project. It is executed by undergoing précised way rather than a conventional process. Most of the minor cases for software failures are identified due to lack of traceability. Requirement traceability is defined as ability to monitor life of requirement in either ways i.e. from the inception through construction, specification, subsequent execution and usage through continuous advancement and recurrence in any of the stages. The evaluator allots value in the range from 1 to 20. After assessing individual requirement for the concerned traceability, the standard of software can be improved by opting the traceability into consideration is chosen for subsequent usage.

3.2 Proposed Prioritization Algorithm:

Values for all the 6 factors are assigned for each test case and analyzed continuously during the software development process. We can compute weighted prioritization value (WPV) for each test case i shown in Eqn(2)

$$WPV=\sum_{i=1}^{10}(PFvalue_i * PF\ weight_i) \quad (2)$$

Where, WPV is weight prioritization value for each test case are calculated from 10 factors.

PF $value_i$ is a value assigned to each test case.

PF $weight_i$ is a weight assigned for each factor.

The computation of WPV for a requirement is used to compute the Weighted Priority (WP) for its associated test cases. Let there be n total requirements for a product and test case j maps to i requirements. Weighted Priority (WP) is calculated in Eqn(3) as

$$WP_{j=}\ (\sum_{x=1}^{i} PFV_x / \sum_{y=1}^{n} PFV_y) \quad (3)$$

By calculating these values we can prioritize the test cases based on WPV and WP for each and every test case in the test suite. **Figure 1** shows, which explains the overview for the proposed prioritization approach which comprises of prioritization factor values for each test case normalized to 20 values and we can prioritize those test cases based on weighted priority value then produces the prioritized test suite.

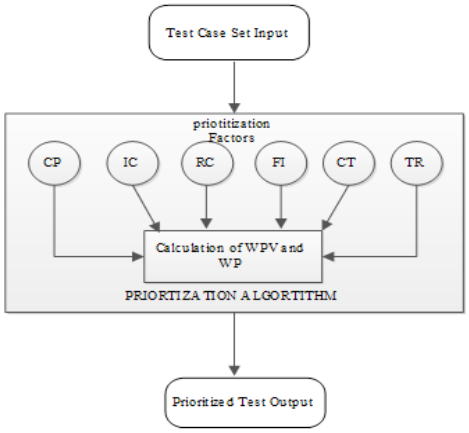


Figure 1. Overview of the implementation of proposed technique

Now we introduce the proposed technique in an algorithmic form here under: This algorithm calculates WPV (weighted priority value) and WP (Weighted Priority) for every test cases which takes into the account of un-prioritized test input. Then any sorting algorithm like quick sort or heap sort can be implemented to sort the WP values in descending order.

3.2.1. Algorithm

Input: Test Case Set (denoted as TS)
Output: Prioritized Test Suite (denoted as PS)

General Process:
Begin
For each test case t in TS
    Calculate WPV for t
End for
While TS is not empty do
    Calculate WP in TS
End While
Sort t in descending order based Weightage
Add t to PS
Return PS
End

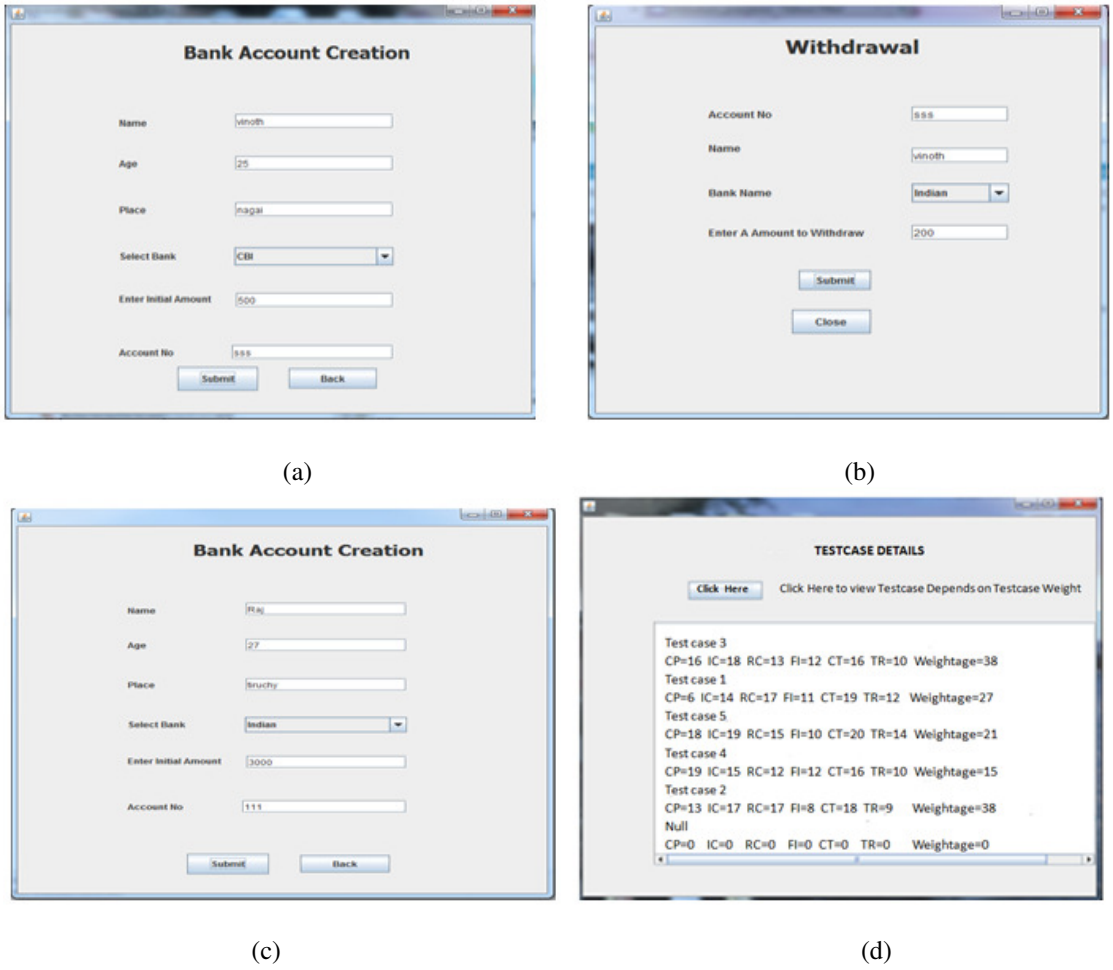


Figure 2. The samples of (a).Requirement for entering account number (The field must be in integer), (b). the sample screen for withdrawal operation, (c). The fault occurs during the bank account creation for the same account number, (d). Final screen for proposed prioritization technique

4. EXPERIMENTAL RESULTS AND ANALYSIS

The test case prioritization system is proposed in this paper which was implemented in the platform of java (JDK 1.6). Here we can use bank application system for regression testing and the results during the process are described as follows: We can create test cases for banking application to check their functionalities. Fig 1 shows that the initial screen obtained for regression testing. When user enters the details it satisfies certain constraints and data must be saved in the database with regard to the operations of the banking applications. Test cases are generated for every wrong details entered by the user, if the requirements for the specific operations are not satisfied, sufficient number of test cases are generated by our proposed system.

After entering the account details for a particular user account, the account number must be unique i.e., the field should be in integer and this can be described in **Figure 2a.** During withdrawal operation, the requirement for account number should be an integer for a definite bank and the test case is generated during this operation can be described in **Figure 2b**. In **Figure**

2c, the field account number is already stored and it should be unique so that the major fault is occurred and the test case are generated and shown. The above figure describes the final output after regression testing. After executing the possible test conditions for each requirement in the banking application, test case are generated. In view of the above we can prioritize the generated test cases using the factor values. Then, we can sort the test cases based on test case weightage and the results are exhibited in the **Figure 2d.**

5. DISCUSSIONS

Here we can evaluate the effectiveness of the proposed prioritization technique by means of APFD metric and the results are compared with random ordered execution. The test suite has been developed for banking application project which consists of 5 test cases and it covers a total of 5 faults. The regression test suite *T* contains 5 test cases with default ordering {T1, T2, T3, T4, and T5} and the number of faults occurs during the regression testing {F1, F2, F3, F4, and F5}. The test case results are shown in the **Table 1**.

Table 1. Fault Detected By Test Suites In Bank Project

Testcases/ Faults	T1	T2	T3	T4	T5
F1				x	
F2		x	x	x	
F3		x	x	x	
F4	x		x		x
F5		x	x		x
No.of faults	1	3	4	3	2

5.1 APFD Metric

The test case prioritization techniques is evaluated by metric of Average Percentage of Fault Detected (APFD*)*. Let *T* be a test suite containing n test cases, *F* be a set of *m* faults revealed by *T*, and TF_i be the first test case index in ordering *T* that reveals fault i. The following equation shows the APFD value for ordering *T'*

$$\text{APFD} = 1 - \frac{TF1 + TF2 + \ldots + TFm}{nm} + \left(\frac{1}{2n}\right) \qquad (4)$$

Researchers have used various prioritization techniques to measure APFD values and found that it produces statistically significant results. The APFD measures that the average number of faults are detected in a given test suite. The

APFD values ranges from 0 to 100 and percentage of fault are detected by plotting the area under the curve towards the percentage of test case executed.

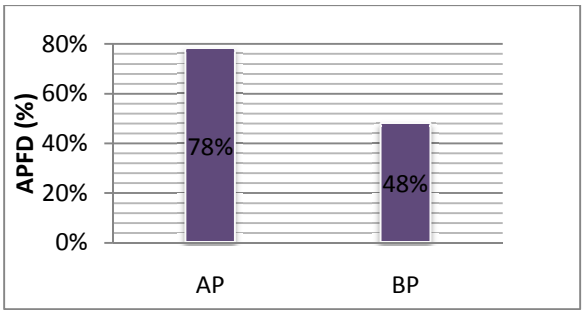
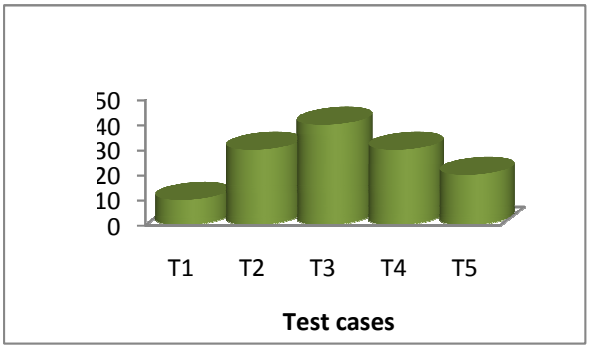


Figure 3. APFD metric for test cases

In this paper, APFD metric is used and the proposed test sequence is {T3, T1, T5, T4, T2}. Then the APFD metric after prioritization is APFD(T,P) is 0.74 and the APFD metric before prioritization is APFD(T,P) is 0.45 as per the above formula. **Figure 3** Shows that the APFD metric compares both prioritized and non-prioritized test suite.



Figrure 4. Fault identified by each test case.

The above figure shows that the test case 5 detects more number of faults and it is shown in **Figure 4**. In the prioritized test suite, more number of faults can be identified when compared with the random execution of test sequence and the same is shown in the **Figure 5**.

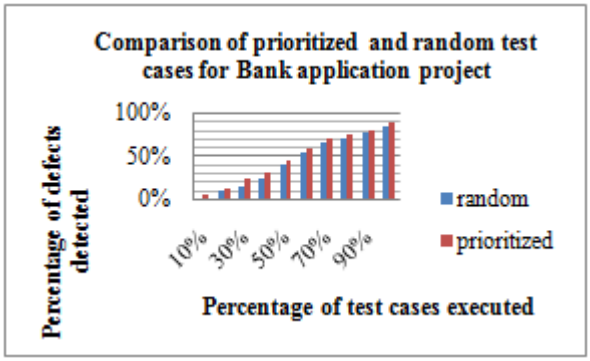


Figure 5. TSFD is higher for prioritized test case which reveals more defects.

Thus the prioritized test cases provides better fault detection than the non – prioritized test cases. Further test case prioritization technique will reduce the processing time of the project by prioritizing the most important test cases.

6. CONCLUSIONS

Here, we have proposed a new prioritization technique for prioritizing system level test cases to improve the rate of fault detection in regression testing. Futher, new practical set of weight factor are used in the test case prioritization process. The new set are tested in the regression test cases. The APFD metric is used to validate the prioritization algorithm. Experimental Results shows that this technique leads to improve the rate of fault detection in comparison with random ordered test cases. Also it reserves the large number of high priority test with least time during a prioritization process.

REFERENCES

[1] Badhera, Usha; Purohit G.N. Biswas, Debarupa. 2012.Test Case Prioritization Algorithm Based Upon Modified Code Coverage In regression Testing.International Journal Of Software Engineering &Applications.Vol. 3 Issue 6, pp.29-34.

[2] BixinLi , Dong Qiu , Hareton Leung , Di Wang.2012. Automatic test case selection for regression testing of composite service based on extensible BPEL flow graph. Journal of Systems and Software. Vol:85 n.6, pp.1300-1324.

[3] B. Jiang, Z. Zhang, W.K Chan, T.H Tse, Adaptive random test case prioritization, in: Proceedings of the 24th IEEE/ACM International Conference on Automated Software Engineering (ASE 2009), IEEE Computer Soceity press, Los Alamitos, CA, 2009, pp.233-244.

[4] Dr. ArvinderKaur and ShubhraGoyal. 2011. A Genetic Algorithm for Fault based Regression Test Case Prioritization. International Journal of Computer Applications. Vol: 32(8).pp:30-37.

[5] Hong Mei, Dan Hao, LingmingZhang,Lu Zhang, Ji Zhou, and Gregg Rothermel. 2012.A Static Approach to Prioritizing JUnit Test Cases.IEEE Transactions On Software Engineering, Vol. 38, No. 6.

[6] H. Do, S. Mirarab, L. Tahvildari, and G. Rothermel.2010. The Effects of Time Constraints on Test Case Prioritization:ASeries of Controlled Experiments. IEEE Trans. Software Eng.Vol:36. no. 5.pp:593-617.

[7] H. Park, H. Ryu, J. Baik.2008. Historical value-based approach for cost-cognizant test case prioritization to improve the effectiveness of regression testing, in: Proc. of the 2nd Int'l Conf. Secure System Integration and Reliability Improvement. pp. 39–46.

[8] Mohamed A Shameem and N Kanagavalli.2013. Dependency Detection for Regression Testing using Test Case Prioritization Techniques. International Journal of Computer ApplicationsVol 65(14): pp:20-25.

[9] M. Yoon, E. Lee, M. Song and B. Choi.2012. A Test Case Prioritization through Correlation of Requirement and Risk. Journal of Software Engineering and Applications. Vol. 5 No. 10. pp. 823-835. doi: 10.4236/jsea.2012.510095.

[10] R. Abreu, P. Zoeteweij, A.J.C. van Gemund.2009. Spectrum-based multiple faultlocalization, in: Proceedings of the 24th IEEE/ACM International Conference on Automated Software Engineering (ASE), pp. 88–99.

[11] R. Bryce, S. Sampath, and A. Memon.2011. Developing a Single Model and Test Prioritization Strategies for Event-Driven Software. IEEE Trans. Software Eng. Vol. 37. no. 1. pp. 48-64.

[12] R. Krishnamoorthi and S. A. Mary.2009.Factor Oriented Requirement Coverage Based System Test Case Prioritization of New and Regression Test Cases. Information and Software Technology. Vol. 51.No. 4. pp. 799-808.

[13] S. Raju and G.V. Uma.2012. An Efficient method to Achieve Effective Test Case Prioritization in Regression Testing using Prioritization Factors. Asian Journal of Information Technology. Vol:11.issue:5.pp:169-180. DOI: 10.3923/ajit.2012.169.180

AUTHORS

Short Biography

M.Thillaikarasi, working as a Assistant Professor in Department of Computer Science and Engineering, Annamalai University, Chidambaram, TamilNadu. She has finished B.E,M.E and pursuing Ph.D in the field of software engineering from Annamalai University She has teaching experience of 8 years. She has published 2 International journals

Dr.K.Seetharaman, working as a Associate Professor in DDE - Engineering Wing(computer), Annamalai University, Chidambaram, TamilNadu. He has finised M.Sc.,PGDCA.,M.S.,Ph.D., He has teaching experience of 10 years and Research experience of 3 Years. He has published 8 International journals and 3 National journals.

INTENTIONAL BLANK

PRIORITY BASED RSA CRYPTOGRAPHIC TECHNIQUE

Meenakshi Shankar[1] and Akshaya.P[2]

[1]Department of Electrical and Electronics Engineering,
Sri Venkateswara College of Engineering, Sriperumbudur, India
meenakshishankar93@gmail.com
[2]Department of Information Technology,
Sri Venkateswara College of Engineering, Sriperumbudur, India
akshya.frenz@gmail.com

ABSTRACT

The RSA algorithm is one of the most commonly used efficient cryptographic algorithms. It provides the required amount of confidentiality, data integrity and privacy. This paper integrates the RSA Algorithm with round-robin priority scheduling scheme in order to extend the level of security and reduce the effectiveness of intrusion. It aims at obtaining minimal overhead, increased throughput and privacy. In this method the user uses the RSA algorithm and generates the encrypted messages that are sorted priority-wise and then sent. The receiver, on receiving the messages decrypts them using the RSA algorithm according to their priority. This method reduces the risk of man-in-middle attacks and timing attacks as the encrypted and decrypted messages are further jumbled based on their priority. It also reduces the power monitoring attack risk if a very small amount of information is exchanged. It raises the bar on the standards of information security, ensuring more efficiency.

KEYWORDS

RSA Algorithm, Cryptography, Priority Scheduling, Encryption & Decryption, Information Security.

1. INTRODUCTION

Message passing in a confidential manner is the key feature of any successful cryptographic technique. Cryptography plays a major role in data protection and authenticity in applications running in a system connected to a network. It allows people to communicate or transfer data electronically without worries of deceit and deception (confidentially)in addition to ensuring the integrity of the message and authenticity of the sender. There is a need for cryptographic algorithms because of the exponential increase in electronic transfer of data in several fields such as, e-commerce, banking, finance, etc. [1].

Cryptography is the science of devising methods that allow information to be sent in a secure form in such a way that the only person able to retrieve this information is the intended recipient [2]. Cryptanalysis is the science of analysing and breaking secure communication. Classical cryptanalysis involves an interesting combination of analytical reasoning, application of mathematical tools, pattern finding, patience, determination, and luck. Cryptanalysts are also called attackers. Cryptology embraces both cryptography and cryptanalysis [3].

Natarajan Meghanathan et al. (Eds) : CSEN, ADCO - 2014
pp. 53–61, 2014. © CS & IT-CSCP 2014 DOI : 10.5121/csit.2014.41006

Cryptography is broadly divided into two categories depending upon the Key; which is defined as the rules used to convert an original text into encrypted text: - Symmetric Key Cryptography and Asymmetric Key Cryptography [4]. In symmetric key cryptography, the encryption and decryption are done using the same key (symmetric key). In asymmetric cryptography, encryption and decryption are done using different keys.

The development of cryptographic techniques has resulted in a large number of ways to securely communicate with a higher degree of efficiency and privacy.

This paper proposes the implementation of RSA algorithm with encryption according to priority and cypher text transfer in parts, alternatively using round-robin technique.

2. CRYPTOGRAPHY

Cryptographic algorithms are classified based on the number of keys used as

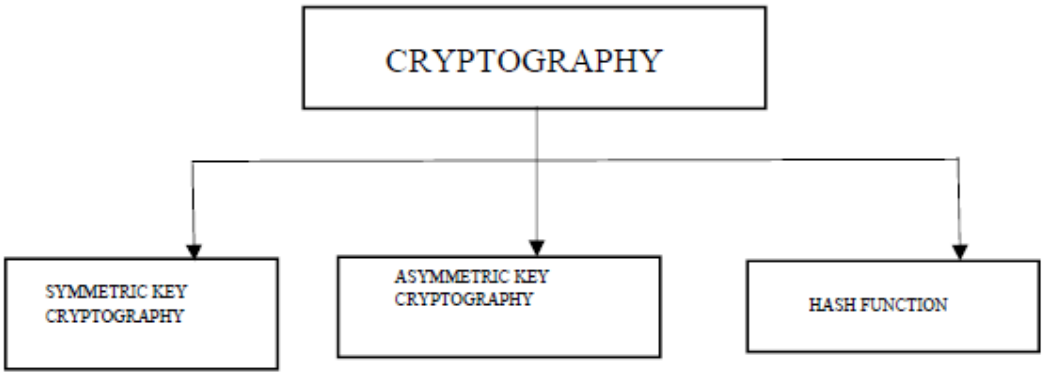


Figure 2. Types of Cryptography

2.1. Secret-Key Cryptography

In secret key crypto there is only one key. It is used for both encryption and decryption. A key refers to any code that yields plain text when applied to cypher text.This key is shared by both sender and receiver. If the key is disclosed the secrecy of the information is compromised. The key is known to both the sender and the receiver, hence does not protect the sender from the receiver forging a message & claiming is sent by sender. Lengthy keys are used to increase the security and to decrease the chances of identifying the key through brute force. It is relatively fast as it uses the same key for encryption and decryption [5]. However, more damage if can occur if the key is compromised. When someone gets their hands on a symmetric key, they can decrypt everything that was encrypted with that key. Since symmetric encryption is used for two-way communication, both sender and receiver end data gets compromised.

2.2. Public-Key Cryptography

Public-key/ two-key/ asymmetric cryptography involves the use of two keys: a public-key, which may be known to everyone, used to encrypt messages and verify signatures and a private-key, known only to the recipient, used to decrypt messages and sign (create signatures). It is called asymmetric cryptography because the key used to encrypt messages or verify signatures cannot be used to decrypt messages or create signatures [5]. Asymmetric key cyphers increase the security and convenience as private keys never have to be transmitted or revealed to anyone. Public key encryption is slow compared to symmetric encryption. It is difficult to encrypt bulk

messages. Interference by a third party results in a type of attack called man-in-middle attack. Damages due loss of private key are mostly irreparable.

Digital signature is a mechanism by which a message is authenticated, proving that amessage is definitely coming from a given sender, much like a signature on a paper document.

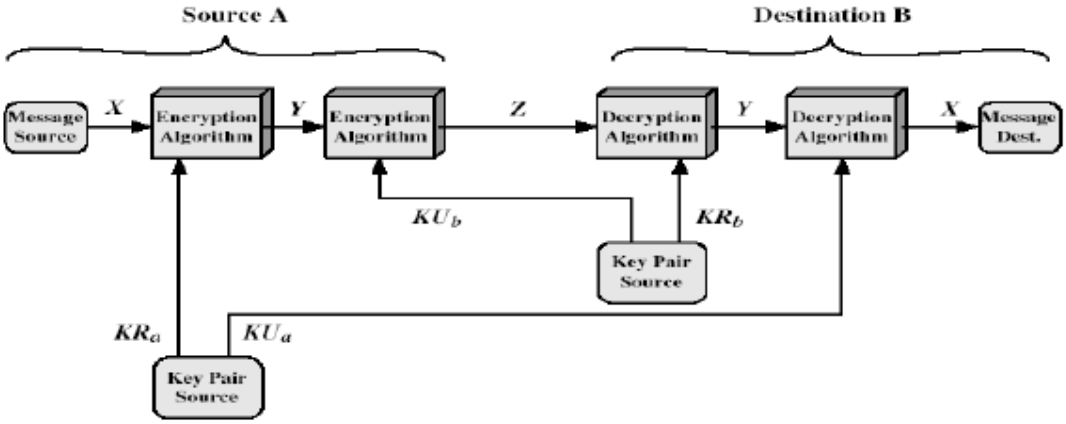


Figure 2. Public Key Cryptosystems: Secrecy and Authentication

2.3. Hash Function

The Hash Function uses a mathematical transformation to irreversibly "encrypt" information. This algorithm does not use keys for encryption and decryption of data. It rather uses a fixed-length hash value which is computed based on some plaintext that makes it impossible for either the contents or the length of the plaintext to be recovered. These algorithms are typically used to provide a digital fingerprint of a file's contents, often used to ensure that the file has not been altered by an intruder or virus. Hash functions are also commonly employed by many operating systems to encrypt passwords to provide some amount of integrity to a file [6].

3. RSA CRYPTOGRAPHIC ALGORITHM

RSA stands forRon Rivest, Adi Shamir and Leonard Adlemanat MIT who first proposed a description of the algorithm publically in 1977. It is a form of asymmetric cryptography.A user of RSA creates and then publishes a public key based on the two large prime numbers, along with an auxiliary value. The prime numbers must be kept secret. Anyone can use the public key to encrypt a message, but with currently published methods, if the public key is large enough, only someone with knowledge of the prime numbers can feasibly decode the message[7].

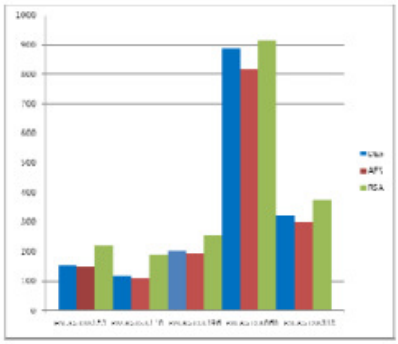
According to the patent issued by theDerwent World Patents Index, RSA algorithm is described as: The system includes a communications channel coupled to at least one terminal having an encoding device and to at least one terminal having a decoding device. A message-to-be-transferred is enciphered to cipher text at the encoding terminal by encoding the message as a number M in a predetermined set. That number is then raised to a first predetermined power (associated with the intended receiver) and finally computed. The remainder or residue, C, is... computed when the exponentiated number is divided by the product of two predetermined prime numbers (associated with the intended receiver).

International Journal of Science and Research (IJSR), India Online ISSN: 2319-7064

S.NO	Algor	Pack Size (KB)	Encrypt Time (Sec)	Decrypt Time (Sec)	Buff Size
1	DES	153	3.0	1	157
	AES		1.6	1.1	152
	RSA		7.3	4.9	222
2	DES	118	3.2	1.2	121
	AES		1.7	1.2	110
	RSA		10.0	5.0	188
3	DES	196	2.0	1.4	201
	AES		1.7	1.24	200
	RSA		8.5	5.9	257
4	DES	868	4.0	1.8	888
	AES		2.0	1.2	889
	RSA		8.2	5.1	934
5	**DES**	312	3.0	1.6	319
	AES		1.8	1.3	300
	RSA		7.8	5.1	416

algorithm show very minor difference in time taken for encryption and decryption process.

Figure 5. Comparative analysis of Buffer Size among DES, AES and RSA algorithm



By analyzing Figure 5 , it shows buffer size usages by AES, DES and RSA algorithm and noticed that RSA algorithm buffer size usages are highest for all sizes of document file.

Figure 3. Comparative analysis of RSA

A cryptographically strong random number generator, which has been properly seeded with adequate entropy, must be used to generate the primes p and q. An analysis comparing millions of public keys gathered from the Internet was carried out in early 2012 by Arjen K. Lenstra, James P. Hughes, Maxime Augier, Joppe W. Bos, Thorsten Kleinjung and Christophe Wachter. They were able to factor 0.2% of the keys using only Euclid's algorithm [8][9].

4. PRIORITY SCHEDULING

CPU Scheduling [14] [15] is the basis of multi programming operating system. By switching the CPU among processes, the operating system can make the computer more productive. Whenever the CPU becomes idle, the operating system must select one of the processes in the ready queue to be executed. This selection process is carried out by the Short-term Scheduler are CPU Scheduler. It selects from all the processes in memory that are ready to execute and allocate the CPU to one of them. The ready queue can be implemented using one of the scheduling algorithms

Scheduling is done in terms of priority in priority scheduling. Priorities are generally some fixed range of numbers. However there is no general agreement on whether the smallest number has the highest or lowest priority. Some systems use low numbers to represent low priority; others use low numbers for high priority. This difference can lead to confusion.

Priorities can be defined either internally or externally. Internally defined priorities use some measurable quantity or quantities to compute the priority of a process. External priorities are set by criteria that are external to the Operating System.

Priority schedulingcan be either preemptive or nonpreemptive. When a process arrives at the ready queue, its priority compared with the priority of the currently running process. A preemptive priority-scheduling algorithm will preempt the CPU if the priority of the newly

arrived process is higher than the priority of the currently running process. A nonpreemptive priority-scheduling algorithm will simply put the new process at the head of the ready queue.

A major problem with priority-scheduling algorithms is indefinite blocking or starvation. A solution to this problem is aging. It is the technique of gradually increasing the priority of processes that wait in the system for a long time [15].

5. ROUND-ROBIN SCHEDULING

The round-robin scheduling algorithm is designed especially for time sharing systems. It is similar to First Come First Serve scheduling, but preemption is added to switch between processes. A small unit of time called a time quantum or time slice is defined. It is generally from 10 to 100 milliseconds. The ready queue is treated as a circular queue. The CPU scheduler goes around the ready queue, allocating the CPU to each process for a time interval of up to one time quantum.

To implement RR Scheduling the ready queue is kept as a FIFO queue of processes.New processes are added to the tail of the ready queue, sets a timer to interrupt after one time quantum and dispatches the process.

One of two things will then happen. The process may have a CPU burst of less than one quantum In that case, the process itself will release the CPU voluntarily. Otherwise, the timer will go off and will cause an interrupt in the operating system. A context switch will be executed and the process will be put at the tail of the ready queue[10][14][15].

6. METHODOLOGY

The RSA Algorithm is used to create a private- public key pair. It is a type of asymmetric key cryptography.

6.1. RSA Algorithm

The steps for implementation of RSA algorithm are given below

1. Get two integers, p and q from the user.
2. Check if p and q are prime. If prime, continue the process, else exit the code.
3. Calculate (p-1)*(q-1) and name it as $\phi(n)$.
4. Calculate n=p*q.
5. Get an input e to act as private key, under the condition that $1<e<\phi(n)$ and gcd(e, ϕ (n)) =1.(gcd-greatest common divisor)
6. Compute the value of d such that $1 < d <\phi(n)$ and $e.d \equiv 1 \pmod{\phi(n)}$.

NOTE:
The public key is (n, e) and the private key is (n, d).
The values of p, q and $\phi(n)$ are private.
‘e’ is the public or encryption exponent.
‘d’ is the private or decryption exponent.

Encryption:
The cypher text C is found by the equation '$C = M^e \bmod n$' where M is the original message.

Decryption:
The message M can be found form the cypher text C by the equation 'M = C^d mod n.

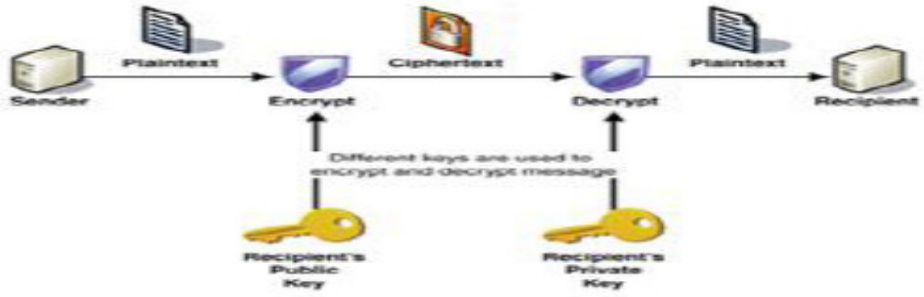


Figure 4. Communication using RSA

6.2. Procedure

This paper proposes the RSA algorithm with some variations in its implementation that will enhance the security of the information transfer. The proposed procedure is given below:

1. The input-prime numbers-(p,q) are obtained from the user.

2. n and ϕ(n) are calculated.

3. All the co-prime numbers from 1 to ϕ(n) are listed out and the user is allowed to choose 'e' from the given values, in addition to any data required for the normal implementation of the RSA algorithm.

4. The private key is obtained by calculating'd'.

5. The message that has to be encrypted is obtained from the user along with the priorities for various parts. The input message is split into low priority, medium priority and high priority parts by the user.

6. The messages are encrypted ($C = M^e$ mod n) and sent to the receiver in parts using round-robin technique. The receiver decrypts the split messages and joins them using the proposed decryption algorithm which is essentially the reverse of the encryption algorithm and uses the RSA algorithm's decryption technique ($M=C^d$ mod n), thus obtaining the message.

7. A software is proposed to be provided for the implementation of this technique. The back end is provided using java code

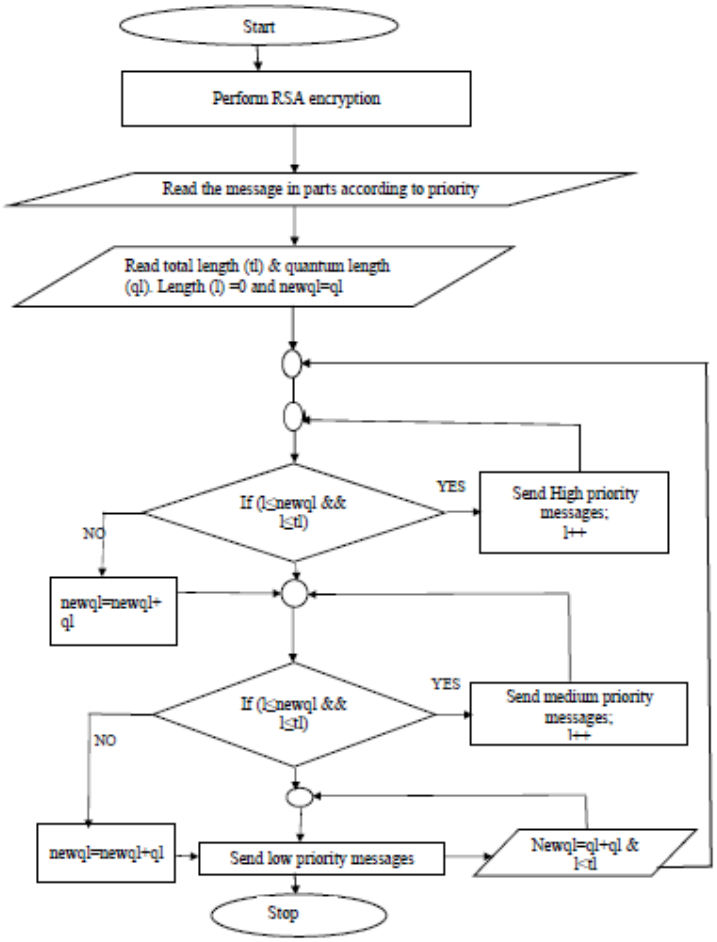


Figure 5. Flowchart for proposed algorithm

7. RELATED WORK

The concepts of dual RSA involving message digest 5 and RSA algorithm to integrate these algorithms effectively and work efficiently in order to provide high security for communication. While the dual RSA mechanism provides a better security, it is takes a longer time to encrypt and decrypt the blocks. The performance is similar to that of the original RSA algorithm, however the encrypting standard of dual RSA is twice as much as that of RSA. Also, dual RSA has better computation, memory and storage capacity. The RSA -dual algorithm achieves decryption 0.25 times faster than the original RSA. The main disadvantage is the time taken for the entire process.

Dual hybrid protection using RSA and DES algorithm through bluetoothidentifies the ways in which file transfer through Bluetooth. A safe combination of security protocols that are the best for wireless transfer of data via Bluetooth have been used, avoiding third party security attacks much as possible. The Digital Signature Algorithm and RSA combination hybrid protocol involves the usage of this combination to provide a stable and secure communication network using the C language [11] [12] [13].

Many cases deal with the security constraints in cloud computing. Basic fiesel network is described, analyzed and used along with the AES algorithm of encryption and decryption. This combination is utilized efficiently to increase the security for cloud computing.

The basic knowledge and information highlighting the concepts on cryptography and internet security along with the scheduling concepts have played a major role in the formation of the concept discussed in this paper. The main idea is to provide heightened security, integrity and immunity against attacks.

7. CONCLUSION

This paper presents an effective method that combines techniques that can be used to successfully communicate secretively in a network. The proposed algorithm reduces the effectiveness of intrusion and brute-force attacks as only a part of the message will be available even if the intruder interrupts any message and decrypts it. Also decrypting part of a message is not very easy. It uses RSA Algorithm, one of the most effective and commonly used cryptographic algorithms and adds more steps to it to reduce attacks. Side channel attacks will not be very effective on this technique as the power levels and leakages that are used to identify the algorithm used will vary from that of RSA algorithm. If an intruder is identified then the sending can be stopped and so, he will not receive the whole message as the messages are sent in parts. This therefore reduces the effectiveness of the man-in-middle attack. Thus, if combined with effective methods to prevent side channel and man-in-middle attacks this algorithm will prove to be very effective. This can also function effectively function as a software that can be used to encrypt/decrypt messages.

REFERENCES

[1] Nentawe Y. Goshwe, (2013). Data Encryption and Decryption Using RSA Algorithm in a Network Environment. IJCSNS International Journal of Computer Science and Network Security, VOL.13 No.7.

[2] P. Gutmann, (2004). Cryptographic Security Architecture: Design andVerification‖. Springer-Verlag.

[3] Ayushi, (2010). A Symmetric Key Cryptographic Algorithm. International Journal of Computer Applications (0975 - 8887), VOL.1 No.15.

[4] SuyashVerma, RajnishChoubey, RoopaliSoni, (2012).An Efficient Developed New Symmetric Key Cryptography Algorithm for Information Security. International Journal of Emerging Technology and Advanced Engineering, ISSN 2250-2459, Volume 2, Issue 7.

[5] Willian Stallings, Cryptography and Network Security: principles and Practice. Tsinghua University Press, 2002.6

[6] Afolabi, A.O and E.R. Adagunodo, (2012). Implementation of an improved data encryption algorithm in a web based learning system. International Journal of research and reviews in Computer Science. Vol. 3, No. 1.

[7] Rivest R, Shamir A, Adleman L, (1978), A Method for Obtaining Digital Signatures and Public-Key Cryptosystems. Communications of the ACM 21 (2): 120–126.

[8] Markoff, John (February 14, 2012). Flaw Found in an Online Encryption Method. New York Times.

[9] Ron was wrong, Whit is right,Arjen K. Lenstra, James P. Hughes,Maxime Augier, Joppe W. Bos, Thorsten Kleinjung, and Christophe Wachter, EPFL IC LACAL, Station 14, CH-1015 Lausanne, Switzerland, Self, Palo Alto, CA, USA.

[10] Abraham Silberschatz, Peter Baer Galvin, Greg Gagne, (2012), Operating System Concepts.Wiley India Pvt Ltd, Sixth Edition.

[11] http://en.wikipedia.org/wiki/Google_App_Engine.

[12] Wayne A. Jansen, (2011), Cloud Hooks: Security and Privacy Issues in Cloud Computing‖, 44th Hawaii International Conference on System Sciences.

[13] G. Jai Arul Jose1, C. Sajeev2, (2011) Implementation of Data Security in Cloud, International Journal of P2P Network Trends and Technology.

[14] Borst, S.C., Kumaran, K., Ramanan, K., Whiting, P.A. (2002). Queueingmodels for user-level performance of Proportional Fair scheduling.

[15] Shakkottai, S., Stolyar, A.L. (2001). Scheduling algorithms for a mixtureof real-time and non-real time data in HDR. In: Teletraffic Engineeringin the Internet Era, Proc. ITC-17, Salvador da Bahia, eds. J.M. de Souza N.L.S. da F

AUTHORS

Meenakshi Shankar and Akshaya.P are final year Electrical and Electronic Engineering and Information Technology students respectively in Sri Venkateswara College of Engineering, Sriperumbudur, India.

They completed their schooling in 2011 from D.A.V Girls Senior Secondary School, Gopalapuram, Chennai. They are interested in Information Security and Cryptography

AUTHOR INDEX