Computer Science & Information Technology 65

David C. Wyld
Jan Zizka (Eds)

Computer Science & Information Technology

Sixth International Conference on Information Technology Convergence and Services (ITCS 2017)
Sydney, Australia, February 25~26, 2017

AIRCC Publishing Corporation

Volume Editors

David C. Wyld,
Southeastern Louisiana University, USA
E-mail: David.Wyld@selu.edu

Jan Zizka,
Mendel University in Brno, Czech Republic
E-mail: zizka.jan@gmail.com

ISSN: 2231 - 5403
ISBN: 978-1-921987-63-2
DOI : 10.5121/csit.2017.70301 - 10.5121/csit.2017.70309

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

Preface

The Sixth International Conference on Information Technology Convergence and Services (ITCS 2017) was held in Sydney, Australia, during February 25~26, 2017. The Sixth International Conference on Signal & Image Processing (SIP 2017), The Fourth International Conference on Foundations of Computer Science & Technology (CST-2017), The Fourth International Conference on Artificial Intelligence & Applications (ARIA-2017) and The Sixth International Conference on Natural Language Processing (NLP 2017) was collocated with The Seventh International Conference on Computer Science, Engineering and Applications (ITCS 2017). 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 ITCS-2017, SIP-2017, CST-2017, ARIA-2017, NLP-2017 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, ITCS-2017, SIP-2017, CST-2017, ARIA-2017, NLP-2017 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 ITCS-2017, SIP-2017, CST-2017, ARIA-2017, NLP-2017.

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.

David C. Wyld
Jan Zizka

Organization

General Chair

Natarajan Meghanathan,	Jackson State University, USA
Brajesh Kumar Kaushik,	Indian Institute of Technology - Roorkee, India

Program Committee Members

Aalya Alajaji	Prince Sultan University, Saudi Arabia
Abraham Gizaw	Hawasa University, Ethiopia
Abrar Abdelhaq	Yarmouk University, Jordan
Addi AIT-MLOUK	Cadi Ayyad university,Morocco
Adnan Rawashdeh	Yarmouk University, Jordan
Ahmed Korichi	University of Ouargla, Algeria
Amiya Kumar TRIPATHY	Edith Cowan University, Australia
Atallah M, AL-Shatnawi	Al al-Byte University, Jordan
Che-Rung Lee	National Tsing Hua University, Taiwan
Dabin Ding	University of Central Missouri, United States
Dmitry Zaitsev	International Humanitarian University, Ukrain
Doina Bein	The Pennsylvania State University, USA
Elaheh Pourabbas	National Research Council, Italy
Emad Al-Shawakfa	Yarmouk University, Jordan
Emilio Jimenez Macias	University of La Rioja, Spain
Erritali Mohammed	Sultan Moulay Slimane University, Morocco
Eyad M. Hassan ALazam	Yarmouk University, Jordan
Fatma A. Omara	Cairo University, Egypt
Feng Yan	University of Nevada, USA
Fernando Tello Gamarra	Federal University of Santa Maria, Brazil
Francisco Prieto Castrillo	Massachusetts Institute of Technology, USA
Gammoudi Aymen	University of Tunis, Tunisia
Gelenbe	Imperial College, UK
Ghasem Mirjalily	Yazd University, Iran
Goreti Marreiros	Polytechnic of Porto, Portugal
Hamed Al-Rubaiee	University of Bedfordshire, United Kingdom
Hannaneh Hajishirzi	University of Washington, USA
Hayet Mouss	Batna Univeristy, Algeria
Hongyang Sun	Ens De Lyon University, France
Hongzhi	Harbin Institute of Technology, China
Houcine Hassan	Univeridad Politecnica de Valencia, Spain
Isa Maleki	Islamic Azad University, Iran
Jae Kwang Lee	Hannam University, South Korea
James Smith	Webscale Networks, USA
Jang-Eui Hong	Chungbuk National University, South Korea
John Tass	University of Patras, Greece
Jun Zhang	South China University of Technology, China

Kayhan Erciyes	Izmir University, Turkey
Kheireddine abainia	USTHB university, Algeria
Kishore Rajagopalan	Prairie Research Institute, US
Lee Beng Yong	Universiti Teknologi MARA, Malaysia
Liyakathunisa Syed	Prince Sultan University, Saudi Arabia
Mahdi Mazinani	IAU Shahreqods, Iran
Mahdi Salarian	University of Illinois, USA
Mario Henrique Souza Pardo	University of São Paulo, Brazil
Masnida Hussin	University Putra Malaysia, Malaysia
Masoumeh Javanbakht	Hakim Sabzevari University,Iran
Md Forhad Rabbi	Curtin University, Australia
Mohamed AMROUNE	Larbi Tebessi university, Algeria
Mohamed Tounsi	Prince Sultan University,Saudi Arabia
Mohammed Al-Sarem	Taibah University, KSA
Mohammed Ghazi Al-Zamel	Yarmouk University,Jordan
Mostafa Ghobaei Arani	Islamic Azad University, Iran
Mudassir Khan	King Khalid University, Saudi Arabia
Nahlah M. Ameen Shatnawi	Yarmouk University, Jordan
Nasrin Akhter	University Putra Malaysia, Malaysia
Neda Firoz	Ewing Christian College, India
Nicolas H. Younan	Mississippi State University, USA
Nishant Doshi	Marwadi Education Foundation, India
Nor Asilah Wati Abdul Hamid	Universiti Putra Malaysia, Malaysia
Noura Taleb	Badji Mokhtar University, Algeria
Partap Singh	IMS College, Roorkee
Rocio Maciel Arellano	University of Guadalajara, México
Shoeib Faraj	Institute of Higher Education of Miaad, Iran
Siuly Siuly	Victoria University, Australia
Soumaya Chaffar	Prince Sultan University, Saudi Arabia
Suleyman Al-Showarah	Isra University, Jordan
Sunil Vadera	University of Salford, UK
Tak-Lam Wong	The Education University of Hong Kong, China
Thiago Pinheiro	Federal University of Amapá, Brazil
Truong Huu Tram	National University of Singapore, Singapore
Valeria Cesario Times	Federal University of Pernambuco, Brazil
Varun Soundararajan	Googleplex, USA
Wen-Lian Hsu	Academia Sinica, Taiwan
Wonjun Lee	The University of Texas at San Antonio, USA
Xiaofeng Wang	Bohai University, China
Yang Wang	Shenzhen Institutes of Advanced Technology, China
Yu Sun	California State Polytechnic University, USA
Yun Tian	Eastern Washington University, USA
Yu-Sheng Su	National Central University, Taiwan
Zaid Hussain	Kuwait University, Kuwait
Zhang jianhong	North China University of technology, China
ZhenChun Huang	Tsinghua University, P.R.China

Technically Sponsored by

Computer Science & Information Technology Community (CSITC)

Database Management Systems Community (DBMSC)

Information Technology Management Community (ITMC)

Organized By

Academy & Industry Research Collaboration Center (AIRCC)

TABLE OF CONTENTS

Sixth International Conference on Information Technology Convergence and Services (ITCS 2017)

Sixth International Conference on Signal & Image Processing (SIP 2017)

Fourth International Conference on Foundations of Computer Science & Technology (CST-2017)

Fourth International Conference on Artificial Intelligence & Applications (ARIA-2017)

Sixth International Conference on Natural Language Processing (NLP 2017)

FACILITATING VIDEO SOCIAL MEDIA SEARCH USING SOCIAL-DRIVEN TAGS COMPUTING

Wei-Feng Tung and Yan-Jun Huang

Department of Information Management,
Fu-Jen Catholic University, Taipei, Taiwan

ABSTRACT

Online video search or stream live on social media has become tremendous widespread and speedy increased continuously in recent years. Most of the videos shared on social media are aimed at the more number of views from audiences. What and how many videos the users shared all around the world have created a great amount and varied videos and the other data into Internet cloud's database and even can be viewed as a kind of big data of digital contents. This research is to present how to implement a social-driven tags computing (SDT) which can be used to facilitate online video search on social media platforms.

KEYWORDS

SDT, Tags Computing, Video Search, Social Media, Social-Driven

1. INTRODUCTION

The video social media and social networks have widely and deeply used no matter on websites or on mobile phones Apps (i.e., YouTube, Instagram, Vine, Dubsmash, Snapchat…etc.) Social media have also tremendous changed the ways to communicate and share things to the other people. Furthermore, social media enable the users exchange, interact, and share the many resources on their communities of social networks. For example, video shared media enables their uses all over the world upload or search all of online videos' resources. As for the users, sponsors, and advertisers, how to promote the videos is their motivations and goals.

In order to increase the effective videos' search via the social media, there are some new methodologies that can facilitate the users to share and browse these video resources effectively. Online users can initiatively input one or more tags (keywords) when they upload their videos. Tagging is also to allow the users actively add one or more tags resources come from users' thoughts. The relevant tags can be determined from the existing tags' databases. Some of these tags (or keywords) might be added by the authors. These tags created just increase the possibility of being searched, and a way to share the authors' opinion toward the resource. Most of the time, these social tags are created according to the users' perceptions toward the resources instead of written by the scholars or authorities of the resource sorting system.

David C. Wyld et al. (Eds) : ITCS, SIP, CST, ARIA, NLP - 2017
pp. 01– 09, 2017. © CS & IT-CSCP 2017 DOI : 10.5121/csit.2017.70301

Majority social media platforms allow any user to state tags. Thus, the users would judge the resources according to their personal experience. Sometimes, users just do not have any idea in mind, not sure which tags are suitable for the resources. Concerning folksonomy, it can be collected a group of users under cooperation and sharing condition in public, adding tags or marks to provide meaning to certain resource. In this research, a social-driven tags computing (SDT) can be represented and used to help the users for their tagging and further facilitate the video searching.

2. LITERATURE REVIEW

Social tagging is the practice of generating electronic tags by users rather than specialists as a way to classify and describe content. Comparing with the information based on scholars or experts, social-driven tags computing (SDT) is a kind of new tagging model, which is also a user-generated classification. The reason why social media include tagging function is to help the users classify their video resources, and the increase of spam tags would destroy the good will of SDT function. Hence, the paper still adopts and implements the SDT computing, and expects the improvements of tagging mechanism on social media.

2.1 Tagging

Most of tagging websites include bookmark, photos, index, video, and blog [1]. Basically, if there is no word on the resource such as photos and videos, what we need most importantly would be users or resource authors' tag for sorting the resources. Since resources like video and photos normally lack word descriptions, it would be too hard to classify tags for the users. If we add a great amount of tags from users or authors' perceptive it would give us a hand for resources indexing and or searching [2]. Even if the resources from blog are mainly formed with words, sometimes these blogs would cause the problems such as too much content or the meaning of words with diverse meaning.

Annotating tags can be defined as a tagging behavior likes keyword for describing the Internet resource. Basically, a definition of tags is similar to keyword indexing. Tagging also possesses the function of content resource classifications [3]. Tagging is the first level analysis, and classification is the second level analysis-paralysis activity [4]. It would be a magnificent task to form a sorting framework, and then reorganize the tags with the framework and model. In comparison with traditional sorting system, bookmark, tagging is relatively easy for users to learn and use, it would not increase the burden of cognitive, and it's easier for maintaining.

Generally speaking, SDT is the organizing model which combines the public's tags or keywords to form the main topic/theme for classification. Every Internet user has his or her own information management model, including personal bookmark, tags, index, email file document etc. Some are sorted with the set level or classification framework; some are sorted with the keywords that are qualified to be recognized, still other even without sorting in advance [5].

2.2 Folksonomy

Folksonomy is the combination of "Taxonomy" and "Folk," which means classification and people [6]. One of the meanings of Folksonomy is a group of users under cooperation and sharing condition in public, adding tags or marks to provide meaning to certain resource. It does not have concept of level, but it has the trait of clustering, meaning that once the resource is tagged more and more times, it could create new definition to the resource and replace the definition laid by experts. Folksonomy does not request the people who classify the resource with professional knowledge, what's more, it encourage the users to sort the resource freely so that the loose classification structure could become convergent gradually and form the definition that could be accepted by the public and scholars.

This kind of sorting mechanism is called Folksonomy. Folksonomy's meaning is close to Social-driven tags computing (SDT). This concept was created by Thomas Vander [6]. Folksonomy is a different way of classification from the traditional systematic classification system. It is conducted by the public, which would come up with the tags or comments toward the resources that are closer to the opinions or feedbacks given by the users.

3. SOCIAL-DRIVEN TAGS COMPUTING

The research proposes a social-driven tags computing (SDT) which can provide online users an enhanced list of tags from the existing tags' database as well as video search. In terms of search engine optimization (SEO), majority online video web sites adopt various advanced or innovative recommendation technologies that can efficiently help their users to share their videos and tag their videos' metadata as shown in Figure one.

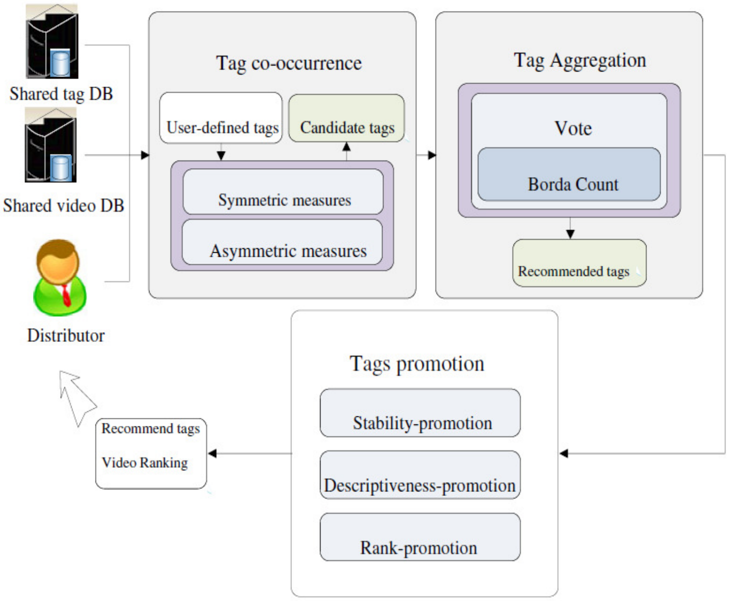


Figure 1. Social-driven tags computing (SDT) framework

3.1 Social-Driven Tags Computing

The first step of this research is to estimate a similarity measurement. Similarity in common use has twofold: symmetric measurement and asymmetric measurement.

Symmetric similarity measurement: Jaccard coefficient can be used to measure the co-occurrence value between tag t_i and tag t_j to measure the degree of similarity shown in Equation 1 [7].

$$J(t_i, t_j) := \frac{|t_i \cap t_j|}{|t_i \cup t_j|} \tag{1}$$

Jaccard coefficient $\mathrm{J}(\mathrm{t_i}, \mathrm{t_j})$ indicates that the interaction of tag $\mathrm{t_i}$ and tag $\mathrm{t_j}$, divide the union of tag $\mathrm{t_i}$ and tag$\mathrm{t_j}$. Jaccard coefficient is applied to measure the similarity of relative tags to determine the tag candidate in this VPA.

Asymmetric similarity measurement: The count of single tag can be normalized to further assess the tag co-occurrence value as follows in Equation 2.

$$\mathrm{P}(\mathrm{t_j}|\mathrm{t_i}) := \frac{|\mathrm{t_i} \cap \mathrm{t_j}|}{|\mathrm{t_i}|} \tag{2}$$

The probability of tag $\mathrm{t_i}$and tag $\mathrm{t_j}$ represents simultaneously while tag $\mathrm{t_i}$ appears.

3.1.1 Tags Aggregation

The tagging candidates can be determined when the co-occurrence of tags was assessed. In the next procedure, the candidate tags can be integrated into a candidate tags list. The first step indicates that these candidate tags need to sort by aggregation using vote and sum. The second step indicates the filtered tags and recommended tag list can be generated through 'Borda count' [8].

While similarity measurements are done with calculation, the tagging aggregation function is the second procedure and the third promotion function for the determination of ranking objects. In the next voting process, the recommended tags can be decided by Borda count. The tags derived from the different candidate tags are compared to the other sets of tags. These selected tags can be voted 1 or 0 and further determined for recommended tags in the next process.

$$vote(u, c) = \begin{cases} 1 & if\ c \in C_u \\ 0 & otherwise \end{cases} \tag{3}$$

If the recommended tags that are selected from the candidate tags are determined, the scores of recommendation can be rated by the counts of voting (u, c) in the voting process.

$$score(\ c\ ) :\ = \sum_{u \in U} vote(\ u, c\ ) \tag{4}$$

3.1.2 Tags Promotions

Most of tags are annotated in the shared tag archive; these tags are usually identified as the unstable tags for recommendation. On the contrary, some tags would be useful to describe the shared object more so than others. However, the tag promotion functions are threefold: stability-promotion, descriptiveness-promotion, and ranking-promotion.

3.2 Integrated Tags Computing

The tag prompt approach is to facilitate the further determinations of ranking scores from the candidates for recommended tag.

Stability-promotion: In order to promote those tags for which the statistics are more stable, the frequency of usages of tags can be measured to represent the levels of stability shown in Equation 5.

$$stability(u) \coloneqq \frac{k_s}{k_s - abs(k_s - \log(|u|))} \tag{5}$$

where $|u|$ represents the frequency of tag u, k_s is a parameter for training.

Descriptiveness-promotion: If the descriptiveness-promotion is high frequency, the shared tags can increase the high frequency for the recommendation of shared objects shown in Equation 6.

$$descriptive(c) \coloneqq \frac{k_d}{k_d + abs(k_d - \log(|c|))} \tag{6}$$

where k_d is a pre-defined training parameter; c is one set of candidate tags.

Ranking-promotion: The co-occurrence provides a good evaluation that can estimate the relationships among the shared tags (u) and change to the ranking(r) for candidate tags $(c \in C_u)$ shown in Equation 7.

$$rank(u, c) := \frac{k_r}{k_r + (r - 1)} \tag{7}$$

$where\ k_r$ is a damping parameter.

According to the three different promotion-functions, a holistic promotion value can be estimated by multiplication shown in Equation 8.

$$promotion(u, c) \coloneqq rank(u, c) \cdot stability(u) \cdot descriptive(c) \tag{8}$$

Based on the aggregation methods of Vote and Sum, the score can be computed by vote and promotion.

$$score(c) \coloneqq \sum_{u \in U} vote(u, c) \cdot promotion(u, c) \tag{9}$$

$$score(c) \coloneqq \sum_{u \in U} vote(u, c) \cdot rank(u, c) \cdot stability(u) \cdot descriptive(c) \tag{10}$$

Where score is the voting results; the three promotion functions use the multiplication of $rank\ (u, c)$,$satability\ (u)$, and $descrriptiv$e (c).

In terms of another mode, it can combine Sum and promotion function.

$$score(c) \coloneqq \sum_{u \in U} (P(c|u) \quad , if\ c \in C_u) \cdot promotion(u, c) \tag{11}$$

$$score(c) \coloneqq \sum_{u \in U} (P(c|u) \quad , if\ \in C_u) \cdot rank(u, c) \cdot stability(u) \cdot descriptive(c) \tag{12}$$

Where score (c) is the sum of voting and promotion functions, $promotion(u, c)$ is a multiplication of $rank(u, c)$, $stability(u), and\ descriptive(c)$.

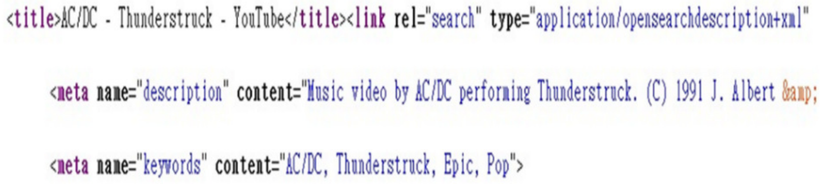
Different combinations of vote, sum, promotion function, and no-promotion function can be used to focus on the different types of shared videos and tags archive [9].

In terms of the recommendation technology, the vote-promotion algorithm (VPA) estimates the ranking scores based on vote value, stability value, descriptive value, and rank value for the results of video-tag relationship prediction. VPA is capable of measuring the degrees of relevance in a numerous collection of tags from the shared video archives.

The algorithm of tagging computing can help the distributors predict a ranking list of recommended tags and videos based on the other relative tags. Figure 6 shows that the recommended tags can be analyzed and determined when users post the initial two tags. This distributor can obtain 6 recommended tags (i.e., Billie Jean, Michael Jackson, Singer, soul, live, and mj) if he posts two tags 'Thriller' and 'Moonwalk'.

3.3 System Process of SDT Computing

The proposed social-driven tags computing (SDT) adopts a 'Crawler' system to search for the relative tags on the video sharing websites. All the names and tags of the shared videos are stored in a video database shown in Figure 2 as resources for video query.



```
<title>AC/DC - Thunderstruck - YouTube</title><link rel="search" type="application/opensearchdescription+xml"

    <meta name="description" content="Music video by AC/DC performing Thunderstruck. (C) 1991 J. Albert &amp;

    <meta name="keywords" content="AC/DC, Thunderstruck, Epic, Pop">
```

Figure 2. Source codes of AC/DC video web pages on Youtube

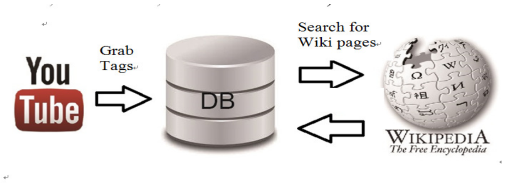
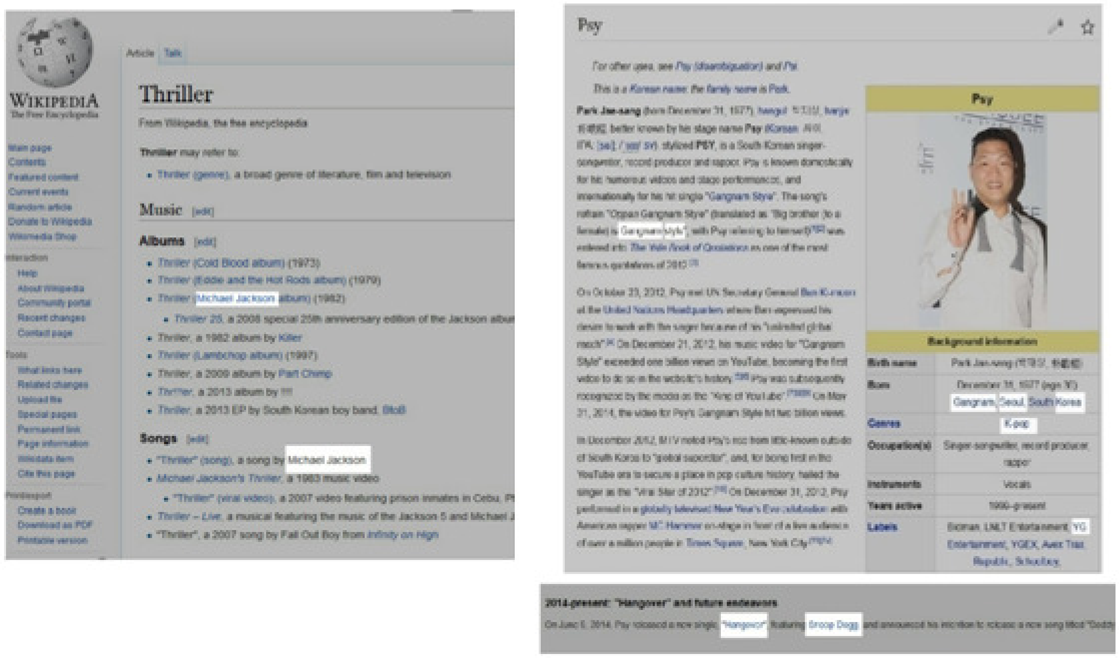


Figure 3. Tag resources from Wikipedia

Wikipedia can be used to refer to the determination process and further adjust the sequence of tags according to the relative tags or terms form searched tags estimations. For example, ‘Thriller’ tags can be changed its relative list of tags.



Thriller

From Wikipedia, the free encyclopedia

Thriller may refer to:

- Thriller (genre), a broad genre of literature, film and television

Music [edit]

Albums [edit]

- Thriller (Cold Blood album) (1973)
- Thriller (Eddie and the Hot Rods album) (1979)
- Thriller (Michael Jackson album) (1982)
  - Thriller 25, a 2008 special 25th anniversary edition of the Jackson albu
- Thriller, a 1982 album by Killer
- Thriller (Lambchop album) (1997)
- Thriller, a 2009 album by Part Chimp
- Thriller, a 2013 album by !!!
- Thriller, a 2013 EP by South Korean boy band, BtoB

Songs [edit]

- "Thriller" (song), a song by Michael Jackson
- Michael Jackson's Thriller, a 1983 music video
  - "Thriller" (viral video), a 2007 video featuring prison inmates in Cebu, Ph
- Thriller – Live, a musical featuring the music of the Jackson 5 and Michael J
- "Thriller", a 2007 song by Fall Out Boy from Infinity on High

Psy

For other uses, see Psy (disambiguation) and Psi.

This is a Korean name; the family name is Park.

Park Jae-sang (born December 31, 1977; hangul: 박재상; hanja: 朴載相), better known by his stage name **Psy** (Korean: 싸이 IPA: [s͈ai] / ˈsaɪ/ SY), stylized **PSY**, is a South Korean singer-songwriter, record producer and rapper. Psy is known domestically for his humorous videos and stage performances, and internationally for his hit single "Gangnam Style". The song's refrain "Oppan Gangnam Style" (translated as "Big brother (to a female) is Gangnam Style", with Psy referring to himself)[5] was entered into The Yale Book of Quotations as one of the most famous quotations of 2012.[7]

On October 23, 2012, Psy met UN Secretary General Ban Ki-moon at the United Nations Headquarters where Ban expressed his desire to work with the singer because of his "unlimited global reach".[8] On December 21, 2012, his music video for "Gangnam Style" exceeded one billion views on YouTube, becoming the first video to do so in the website's history.[9] Psy was subsequently recognized by the media as the "King of YouTube".[10][11] On May 31, 2014, the video for Psy's Gangnam Style hit two billion views.

In December 2012, MTV noted Psy's rise from little-known outside of South Korea to "global superstar", and, for being first in the YouTube era to secure a place in pop culture history, hailed the singer as the "Viral Star of 2012".[12] On December 31, 2012, Psy performed in a globally televised New Year's Eve celebration with American rapper MC Hammer on-stage in front of a live audience of over a million people in Times Square, New York City.[13][14]

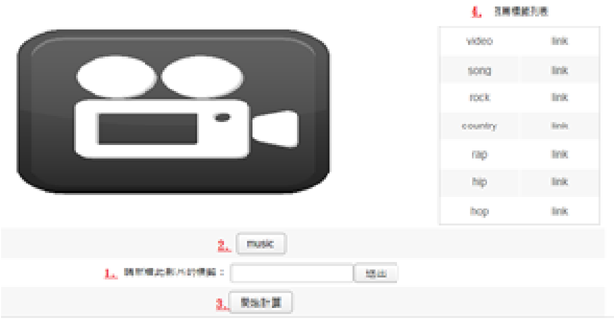
Psy	
Background information	
Birth name	Park Jae-sang (박재상, 朴載相)
Born	December 31, 1977 (age 37) Gangnam, Seoul, South Korea
Genres	K-pop
Occupation(s)	Singer-songwriter, record producer, rapper
Instruments	Vocals
Years active	1999–present
Labels	Sidus, LNLT Entertainment, YG Entertainment, YGEX, Avex Trax, Republic, Schoolboy

2014-present: "Hangover" and future endeavors

On June 8, 2014, Psy released a new single "Hangover" featuring Snoop Dogg and announced his intention to release a new song titled "Daddy

Figure 4. Wikipedia data facilitate the tag determination

When a user intends to upload a video and needs to provide the tags at the same time, the tags system developed by the research can then generate a list of recommended tags from the video database and wikipedia as shown in Figure 3 and 4. Determined by SDT, the weights are also represented for ranking (Fig. 5)



4. [illegible]

video	link
song	link
rock	link
country	link
rap	link
hip	link
hop	link

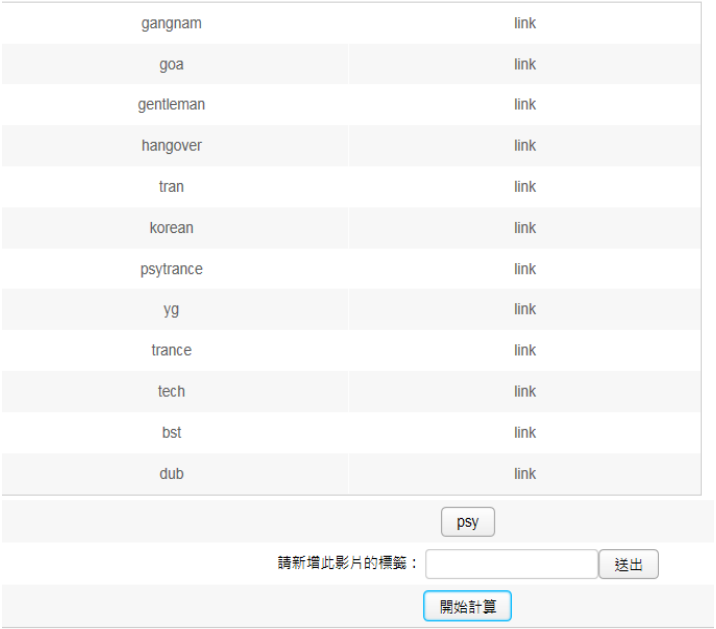
2. music

1. 請[illegible]： [illegible]

3. [illegible]

Figure 5. SDT Computing User Interface

SDT estimates the ‘Jaccard coefficient’ to calculate the co-occurrence values to provide the candidate tags based on the particular tags from the user queries as the follows figure 6.



gangnam	link
goa	link
gentleman	link
hangover	link
tran	link
korean	link
psytrance	link
yg	link
trance	link
tech	link
bst	link
dub	link

Figure 6. Tags recommended for Video Search

To improve the efficiency ranking of search, more detailed tags should be given higher weight than general tags as shown in Table 1. To adjust the weights of tags by the computing of ranking, the promotion estimations can be facilitated for video search (Fig. 7)

Table 1. Tags Estimations by Promotion

Tags	**Parachutes**	**Stories**	**Ghost**	**Magic**	**Paradise**
Ranking Promotion	0.9523	1	0.9090	0.7820	0.8
Stable Promotion	0.6356	0.6356	0.6356	0.6356	0.6356
Describe Promotion	0.7153	0.6705	0.5670	0.7820	0.5916
Usage Frequency	4	31	58	19	49

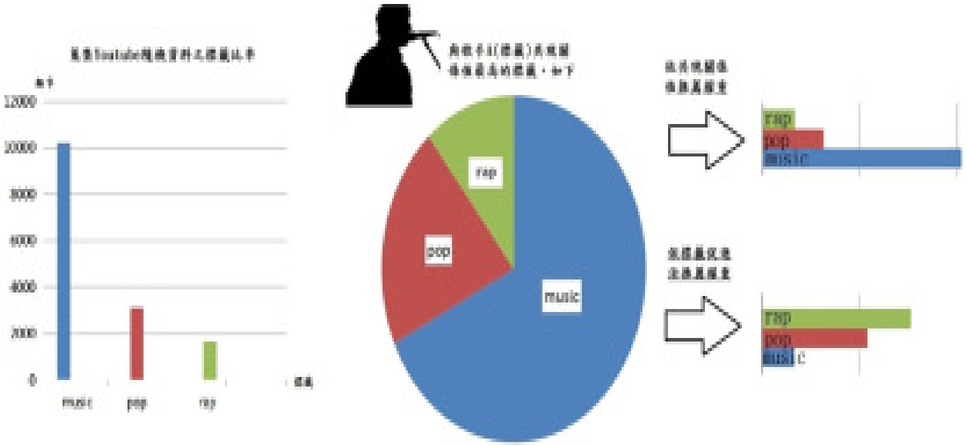


Figure 7. The weights of tags can be adjusted by SDT computing

4. CONCLUSION

Social-driven tags computing is to facilitate online video search and even sharing. As most of the videos shared on social media are aimed at the more number of views from audiences, the users (distributors) want to annotate some valuable tags. What and how many videos the users make decision by themselves, but this research can help the users to choice the other recommended tags for the specific video resources. The SDT methodology includes the co-occurrence estimations and tags voting as well as promotions like stability, descriptive, and rank. Those algorithms are able to determine the valuable tags according to the existing tags' databases in the social media.

REFERENCES

[1] Hammond, T., et al. , "Social bookmarking tool", D-Lib Magazine Vol. 11, no. 4, 2005.

[2] Trant, J. and Wyman, B., "Investigating social tagging and folksonomy in art museums with Steve.Museum", Proceedings of the Www Collaborative Web Tagging Workshop, 2006.

[3] Marlow, C., Naaman, M., and Boyd, D. "HT06, tagging paper, taxonomy, Flickr, academic article, to read", Conference on Hypertext. 2006.

[4] Sinha, R.(2005, September). A cognitive analysis of tagging. [online] Available: http://www.rashmisinha.com/archives/05_09/tagging-cognitive.html.

[5] Jones, W., Dumais, S., and Bruce, H. "Once found, what then? A study of "keeping" behaviors in the personal use of Web information". Proceedings of the American Society for Information Science & Technology, p.p. 391-402. 2002.

[6] Thomas Vander Wal. Explaining and showing broad and narrow folksonomies. 2005.

[7] Loet Leydesdorff, "Regional Development in the Knowledge-Based Economy: The Construction of Advantage". Journal of Technology Transfer. Special Issue, pp. 1–15, 2006.

[8] Singurbjornsson, B. and Zwol, R. V., "Flickr tag recommendation based on collective knowledge", Proceeding of the 17th International Conference on World Wide Web, 2008, pp.327–336.

[9] Christian Wartena, Rogier Brussee, and Martin Wibbels, "Using Tag Co-occurrence for Recommendation", International Conference on Intelligent Systems Design & Applications, 2009, pp.273-278.

INTENTIONAL BLANK

““USATESTDOWN” A PROPOSAL OF A USABILITY TESTING GUIDE FOR MOBILE APPLICATIONS FOCUSED ON PERSONS WITH DOWN SYNDROME.”

Doris Cáliz[1], Loïc Martínez [1], Richart Cáliz[2].

[1]Department ETSIINF, DLSIIS, Madrid Polytechnic University, Campus de Montegancedo 28660, Boadilla del Monte, Madrid, Spain
[2]Department of Computer Sciences FIS Group, National Polytechnic University, Ladrón de Guevara E11-25 y Andalucía Quito, Ecuador

ABSTRACT

The usability testing of mobile applications involving persons with Down syndrome is an issue that has not be comprehensively investigated and there is no single proposal that takes on board all the issues that could be taken into account[1]. This study aims to propose a practical guide ¨USATESTDOWN¨ to measure and evaluate the usability of mobile applications focusing on Down syndrome users and their primary limitations. The study starts with an analysis of existing methodologies and tools to evaluate usability and integrates concepts related to inspection and inquiry methods into a proposal. The proposal includes the opinions of experts and representative users; their limitations, the applicability during the development process and the accessibility. This guide is based on the literature review and the author's experience in several workshops where persons with Down syndrome used mobile devices.

KEYWORDS

Usability Testing, Mobile Applications, Cognitive Disability, Down Syndrome, Human Computer Interaction (HCI), Mobile Devices.

1. INTRODUCTION

Down syndrome (DS) is a genetic disorder with a worldwide incidence close to one in every 700 births but the risk varies with the mother's age. Persons with DS have impaired cognitive processing, language learning and physical abilities, as well as different personal and social characteristics [11]. Because Persons with DS have special characteristics, they need high levels of usability of the products they use. A usability testing methodology suitable for participants including persons with DS needs to be well designed taking on count their special skills [12].
The International Organization for Standardizations (ISO) bases usability on three main attributes: effectiveness, efficiency and satisfaction. Systems with good usability are easy to learn, efficient, not prone to errors and generate user satisfaction [10].

David C. Wyld et al. (Eds) : ITCS, SIP, CST, ARIA, NLP - 2017
pp. 11– 26, 2017. © CS & IT-CSCP 2017 DOI : 10.5121/csit.2017.70302

Testing products with representative users is a key factor for user-centred design. When such representative users are persons with disabilities the user testing process becomes a challenge and in this case evaluation methods based on heuristics and inspection could not attend the final user needs [3].

Persons Persons with Down syndrome have many difficulties to use the mouse and the keyboard because they have fingers shorter than usual [4]. Multi-touch technology helps to solve this problems when people use devices, such as mouse, keyboard or joystick, and enables users to take advantage of the direct manipulation interaction and the benefits of direct touch [9]. There are a big range of functional abilities in individuals with Down syndrome, related to the extent of impairment in the sensory and motor channels [5], memory, and cognition and communication skills [6]. These sensory and motor issues would need to be taken into consideration when researchers want to evaluate a mobile application in individuals with Down syndrome.

The authors have performed a detailed research on articles related with this topic and they have not found a guide to support the usability testing process for mobile applications focused on persons with Down syndrome [7]. After that, they they have evaluated the use of a tool called "Gestures" by a group of 100 persons with DS. The goal was to analyse the skills of these persons to perform basic gestures [8]. The authors found that DS children 5 to 10 year-old are able to perform most of the evaluated multi-touch gestures with success rates close to 100 per cent. This research study is designed to be a preliminary investigation of how users with Down syndrome could potentially utilize touch-screens gestures tasks to obtain a sense of some of the potential challenges to effective use of tablet computers for this population and to investigate how usability testing involving Persons with Down syndrome could be effectively performed. [9].The result of combining the literature review and the research experience in several workshops is the guide to perform usability testing when the participants are persons with DS. This guide is called "USATESTDOWN".

Where is USATESTDOWN

In the Human Computer Interaction area one of the most commonly used design philosophies to create high quality products for users is the User-Centred Design (UCD) approach [2] UCD refers to the philosophy that the intended user of a product should always be in the centred of the design process throughout all phases of the design [3]. Usability testing, according to Dumas & Redish [4], aims to achieve the following five goals, to: Improve the product's usability, Involve real users in the testing, give the users real tasks to accomplish, Enable testers to observe and record the actions of the participants, Enable testers analyse the data obtained and make changes accordingly. USATESTDOWN is inside Evaluate the Design Requirements as we can see in Fig 1.

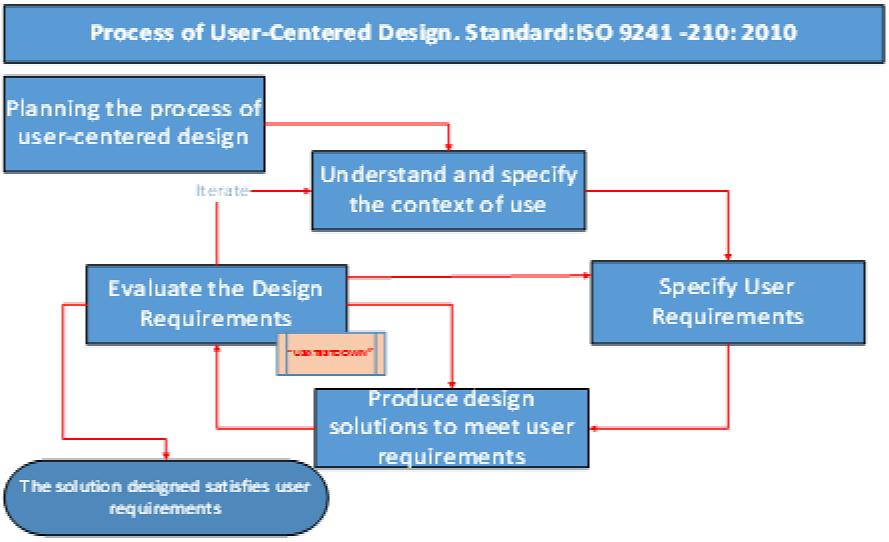

Process of User-Centered Design. Standard:ISO 9241 -210: 2010
Planning the process of user-centered design
Understand and specify the context of use
Iterate
Specify User Requirements
Evaluate the Design Requirements
Produce design solutions to meet user requirements
The solution designed satisfies user requirements

Figure 1: Process of User Centred Design.

2. RELATED WORK

2.1 Usability evaluation methods

There are three types of usability evaluation methods: observational, analytical and inquiry evaluation methods [5]. Evaluation methods that collect data by observing users' experiences with a product are called observational evaluation methods. Usability testing, user-oriented view and user performance testing are types of observational evaluation methods [6]. Methods that do not collect data from users' experiences but rely on the opinion of experts are called inspection or analytical evaluation methods. These methods have a product-oriented view. Examples of analytical evaluation methods are Heuristic Evaluation [7] Cognitive Walkthrough and Semiotic Inspection [8]. Inquiry methods have a user-oriented view. Inquiry methods tend to identify broad usability problems or opinions about a product as a whole such as user Satisfaction Questionnaires and Focus Groups [9].

Usability Evaluation methods for mobile applications focused on persons with Down syndrome

While there is some related research, it is incomplete. Devan does not consider mobile or touch screen devices. The author used an application called JECRIPE, this application works in a PC. Additionally it is not a Usability Testing Guide [10]. Kumin and Lazar did a usability evaluation to understand potential interface improvements and they suggest different tips to evaluate usability but it is not a Usability Test guide [11]. AR BACA SindD is a usability evaluation framework for an augmented reality framework for learners with DS but they did a specific evaluation in AR Systems but it is not a Usability test guide [12]. Adebesin, Kotzé show the important role of two evaluation methods in the usability [13]. The authors did not speak about touch screen, usability guide etc. (MEL-SindD) discusses the usability assessment of the courseware but it is not focus on mobile applications [14].

2.2 Working Method Overview

The guide reproduces the usual usability testing steps. The guide provides recommendations taking into account the needs of people with DS in the usability testing process.

In general, the working method has four main phases, as shown in Fig. 2. The process is iterative.

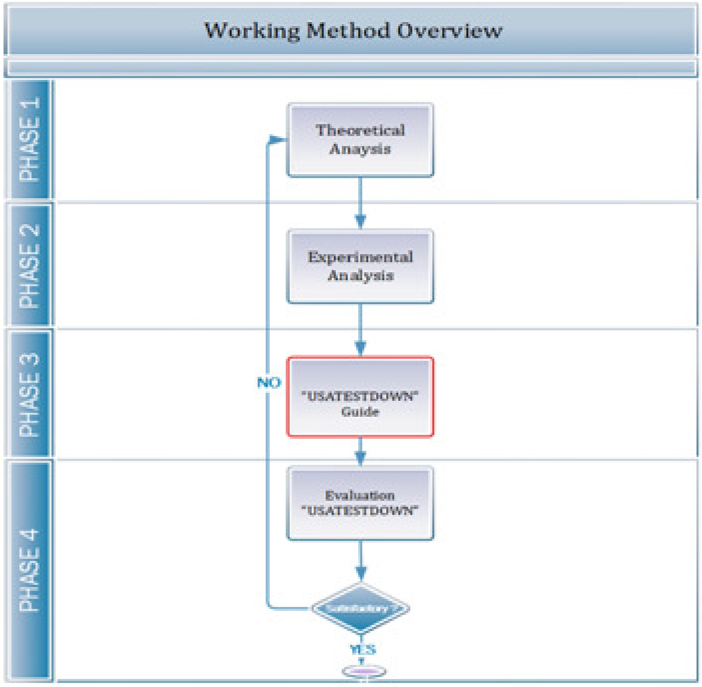


Figure 2: Working Method Overview

- **Theoretical analysis.** A state of the art on usability testing involving persons with Down syndrome.

- **Experimental analysis.** There have been made experiments on usability testing with persons with Down syndrome.

- **The guideline "USATESTDOWN"**: This phase consists on the preparation of a guideline to perform usability testing involving persons with Down syndrome. The contents of the guideline, called "USATESTDOWN" are based on the results of phases 1 and 2. The development of the guideline will be iterative. Observational evaluation has been chosen as the method to be used in the usability testing.

- **Evaluation of "USATESTDOWN":** The USATESTDOWN guideline will be evaluated with a set of experiments involving persons with Down syndrome in usability testing of mobile applications. The results of the evaluation will be used to improve the guideline.

2.3 Usability Testing Previous Contributions for Mobile Applications Focused on Persons with Down Syndrome

The most common method for evaluating how usable a product or system is usability testing, which involves testing prototypes with real users [4]. Participating users are given a set number of tasks that they have to perform using a prototype or a full system. Data on the effectiveness, efficiency and satisfaction of users are collected during testing. Generally, the usability process is divided into the following steps: 1. Recruit participants, 2. Establish the tasks, 3. Write the instructions, 4. Define the test plan, 5. Run the pilot test, 6. Refine the test plan, 7. Run the test session, 8. Analyse the collected objective, and 9. Report results.

We found 5 articles related with our topic after a Literature Review research. We used the definition of the main steps of usability testing [15] to analyse the contributions of each author on each usability testing step. We took the authors contributions in each point [7], [8], [9], [10], [11]. But is important notice they contribute only with the steps 1, 2, 3, 5, 7, 8, the steps 4, 6, 9 were deleted because there are not contributions. We had the results in table 1. We can see there are several empty spaces, meaning that there are not contribution in those specific steps.

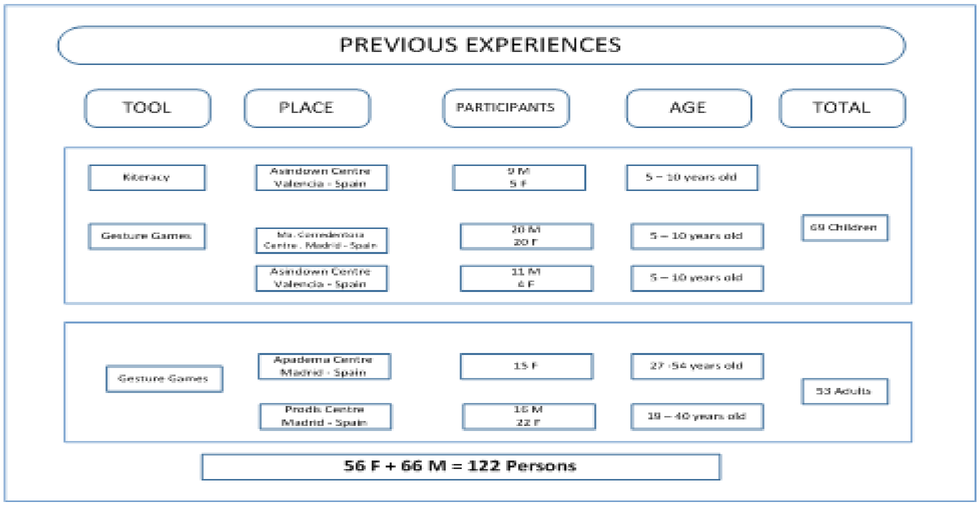
Table 1 Previous Contributions of Usability Testing

Paper	**1. Participants**	**2. Tasks**	**3. Instructions**	**5. Pilot testing**	**7. Testing**	**8. Analyse**
[10]2013	X	X				X
[13]2010	X	X		X.	X	
[12]2011	X	X				
[11]2012		X		X		
[14]2009	X	X	X		X	

There is not a Guide to evaluate Usability in mobile applications focused on Down syndrome person. Consequently, the authors proposed the need to develop guidelines on the usability testing process in mobile applications involving participants with Down syndrome.

2.4 Collected Experience

USATESTDOWN is a guide to support usability testing of mobile applications when the participants are persons with DS. It has been developed by combining information collected from a literature review [15] and experience acquired during four workshops with approximately 100 people with DS [16][17]. We performed several workshops in different Special Dow Syndrome Centre in Spain (Asindown [16], Maria Corredentora [17], Apadema [18], Prodis [19]) as we show in the figure 1. We evaluated 122 persons, 69 children and 53 adults with Down syndrome to determinate the skills, behave and how they interact with mobile devices.



PREVIOUS EXPERIENCES

TOOL	PLACE	PARTICIPANTS	AGE	TOTAL
Kiteracy	Asindown Centre Valencia - Spain	9 M 5 F	5 – 10 years old	
Gesture Games	Ma. Corredentora Centre , Madrid - Spain	20 M 20 F	5 – 10 years old	69 Children
	Asindown Centre Valencia - Spain	11 M 4 F	5 – 10 years old	
Gesture Games	Apadema Centre Madrid - Spain	15 F	27 -54 years old	53 Adults
	Prodis Centre Madrid - Spain	16 M 22 F	19 – 40 years old	

56 F + 66 M = 122 Persons

Figure 3: Collected Experience

3. ¨USATESTDOWN¨ GUIDE PROPOSAL

USATESTDOWN is a guide to help usability tests of mobile applications focused on users with Down syndrome. Applying the usability testing guide USATESTDOWN, the evaluators can easily manage the usability test with applications on mobile devices for persons with Down syndrome in the different workshops following the different steps that the guide proposes. We describe the 9 steps of USATESTDOWN recommend for the authors, such as: [23], [24], [25], [24], and [26]. We describe this guide with specific activities in order to evaluate Mobile applications software focused on persons with Down syndrome. The flow of the process was adjusted to account for the reality of the persons with Down syndrome. We showed in Fig, the USATESTDOWN Guide process.

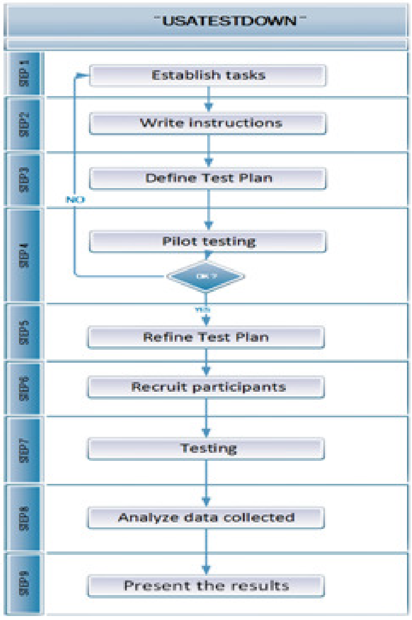


Figure 4: Process of Usability Testing as defined in USATESTDOWN

After the USATESTDOWN figure scheme we have 9 tables from number II until X with the following information:

- **Definition:** According with the authors [18], [19], [20] what we should expect in this point.
- **Biography Research:** A collection about what the authors propose in this step [7] [8] [9] [10] [11].
- **Usatestdown:** Contribution of the proposal Guide in order to the experience with the previous workshops realized by the authors
- **Documents:** Documents to support the step adapted specially to people with DS.

1. Establish the tasks.

Table 2: Usatestdown: Establish the tasks

USATESTDOWN	
Step:	**Establish the tasks**
Definition	This step consists of defining the tasks that the participants will complete in the usability tests should be defined. These tasks will be identified in pre-development phases to identify which of them will form a part of the evaluation that will include tasks that appear in certain usability specifications, as well as other representative tasks.
Biography Research	Holds a 30-minute training session [10], takes 20-minute videos per child and uses the DEVAN method to work directly with children with DS. On the other hand, [13] evaluates a literacy portal in Africa using the following tasks: submission of evaluation criteria, submission of document stating procedure to be followed, submission of document on interfaces and applications for evaluation, signature of anonymity and confidentiality forms. In the research by [13], the experts identify critical usability problems in the early stages of the development cycle and divide the evaluation into two phases: acceptance testing and usability. [14] divides the tasks used in the evaluation into several phases: PHASE 1. Identify user needs, iteratively engage students in testing, and collect data from teachers and parents of students with DS, PHASE 2. Conduct the usability evaluation, and PHASE 3. Collect the data from specialist teachers and parents and hold the scheduled interviews.
Usatestdown	• To establish the tasks, it should be considered that they are increased according to a gradual completion. This is to say that the completed levels will be driven by each individual with Down’s Syndrome due to the fact that they have different levels of abilities and what might be simple for one may be more complex for others. Therefore, it is necessary that they determine to which level they will complete. • The tasks should not be very complex on a difficult scale of 1 to 3, where 1 is esaier , the thask dhould take 1, • The session should be done in 10 minute sessions for each person, because they will get tired easlily. This point will allow the individual to evaluate the application with curiosity without getting overwhelmed or bored. • Do not limit the time. The participants will stress and become confused if they have a time limit for the task. We could see on the sesions that the participants were getting afraid with the topic task limited with time. They could feel nervous.

2. Write the instructions

Table 3: Usatestdown : Write the instructions

USATESTDOWN	
Step :	***Write the instructions***
Definition	Specify the instructions given to the users (oral, written, or both) to complete each task
Biography Research:	[14] describe the instructions for identifying the needs of users, which are collect data, interview students' paediatrician and primary school teachers, interact socially with students; identify the learning needs. Understand the problems through conversations with parents; interview specialists, teachers and parents as informers on the background of students and the research.
Usatestdown	• Before completing the test, it is necessary to give a presentation to the people who will participate with the evaluation. They should interact with the aplication on a training way. The Facilitator should explain the project objectives and he must to ask if the participant would like to participate even if the Tutor recomeded this participant. The willingness of the participants to participate is very important because the results depend upon it. • Task scenarios for this usability test will be based on the tool and taking on count the participants number • Observation method needs facilitator to record all children action, behaviours and facial expressions while observing children playing the game. In order to guide facilitator, an observation checklist is needed to analyse the participants behavoiur.

3. Define the test plan

Table 4: Usatestdown: Define the test plan

USATESTDOWN	
Step :	***Define the test plan***
Definition	It is necessary to specify the protocol with alternative activities, such as, welcome, interview preview, completing the tasks by observing the user, satisfaction questionnaire, personal interview to collect qualitative information, etc. It is recommended to write an introductory commentary to express a welcome to the users. It is necessary, as part of these instructions, to collect the data needed by the users to complete the tasks.
Biography Research:	NO CONTRIBUTIONS
Usatestdown	• It is necessary to complete a demographic survey including name, age, gender, experience with mobile devices. It must to have only general information, even it could use a fake name because family, tutors and participants are so reserved wit information that it would allow them to be identifier. Never take last names. Don't push the people to answer if they don't want, it may generate a bad atmosphere to work. This is like the test that was used in the workshops and can be found in Annex 1. It should be completed by the evaluator. • It is important to prepare a user satisfaction survey with a scale of no greater than 3 categories and, if possible, with graphics of faces (happy, neutral, sad). We tried wit 5 answers but it was confusing to the participants. Aditionally we had a meeting with the students psicologist supervisors and they also recommended only 3 levels. • Avoid providing documents with long text to the people who will participate in the test Generally, people with DS have vision problems and it is taxing for them to read and speak. It is recommended that the instructions be given verbally and in a graphic form that is simple, allowing them to understand the information. In the first workshop session , we wrote test to explain to the participants the steps that they should follow but they had probles to read or pronunce and to understand. It si better to avoid • Write short questions only with 3 answers (not, may be, yes), in this specific order

	because if they find fist the option YES they will not read the rest of te options. If you want to ask about quantity answer posibility is (few, many, too many). It means only 3 answers. • Write questions focused questions on the applications and that do not evaluate the affinity of the participant . The participants are likely to provide friendly answers whether or not they like the application because they tend to be friendly people. People with Down Syndrome are so friendly and they like to make friends, this is their normal behaviour, we had seeral cases with participants whoes answer the questionary saying the aplication was esy to understand, but when we analized the results with the log aplicaction the participants could not get success on the task or the success level was so low. Aditionally when we asked Did you like the aplication ?' No body said NO, it showed us that it was not the rally truth. The wanted just to be friendly. • The sessions will be facilitated and observed by only one facilitator because on the workshops the participants were shy when the see many new people in the room. • People need to be encouraged to participate and facilitator should stress the value of the child's input and show appreciation and gratitude. This is a task that the facilitator should do. • It is recommended that, at the moment of introducing the tasks, the educators of children with DS are present. This is very necessary in an initial demographics test. • Apply at least 2 evaluation methods, it means oservational method to analise the user behaviour and the tool should have a log or a way to evaluate if the task was completed susscesfully or not, aditionally you shoul use a satisfaction cuestionary, in this case the proposal recommend the SUS Questionary adapted to persons with Down Syndrom. • Establish objective metrics with a completion time for the task, error rate, etc. Performing the workshops we could notise the time is not a good parameter to take on count because when we said people with down syndrome participants that they should do the task on a specific time, inmediatly they were scared becase they thought is an evaluation about how smart they are and it is obviosly a big problem even to people without down syndrome . • The participants should have pre training about the aplication, it mens tutor should teach to the participants how the aplication work and the should interact par minutes with the aplication before the real test. • Establish subjective metrics such as sucess, frustration, satisfaction, etc. while they are using the application. Success: Defined as the completion of a task done correctly and without help. Satisfaction: When the user gets the correct result easily, when the user shows happiness when interacting with the application, etc. Frustration: When the user has problems answering a tutor's questions, when the user gives an incorrect answer, when the user doesn't understand the process, etc. • Schedule a break halfway through the test session and remind the participants that they can stop at anytime. • To evaluate Success, frustration, satisfaction you could use the videos recorder during the workshop or sesion or at te same time the tutor could take on count , you should measure of every tasks , how many times the participant showed this rection. • Hold a meeting prior to executing the test because it is very important to break the ice with the participant so that they feel safe and trusting of the process at the workshop, This is a way to make friendly and relaxin the work enviroment • Do not use the technique of "thinking out loud" because the majority of the participants have difficulty expressing themselves
Documents	The documents should be printed -UsatestDown Demographic Questionary (Annexes) -UsatestDown SUS Questionary adapted to people with Down Syndrome focused in the whole process -UsatestDown SUS Questionary to Tutors focused in the whole process (Annexes)

4. Run the pilot test

Table 5: Usatestdown: Run the pilot test

USATESTDOWN	
Step :	***Run the pilot test***
Definition	Execute the test protocol using the welcome, the written instructions, completing the observations, measuring times, completing the interviews, etc. in order to analyze if the proposed process functions as expected. In the case that it is not, it should be writen as the protocol describe.
Biography Research:	F. Adebesin et al. [13] conducted a pilot test aimed at understanding how applications work. L. Kumin and J. Lazar,[11] believe formal data collection to be important for the pilot test. This should be followed by a second session during which they suggest modifying the list of tasks, adding a warm-up task.
Usatestdown	• Record in video the interaction of the person with the mobile device. It is recommended and very important in order to qualitatively evaluate their interaction with it when we review the videos. Be careful not to film faces and to obtain authorization in order to complete this point. • Ask for permission if it is necessary to film faces,but evaluator should be really polite because this is a sensitive topic. Some times is necessary the fathers autorization to record participants faces . • It is recommended that the pilot test is done through a small samples but in this case it means only one person because the evaluator will notice with the second participant the mistake will be the same. This will allow a definition of the first process errors without needing to involve all of the participants in the process, repeating the same error. • During the entire evaluation process or user participation, it is necessary that the tutors or professors with whom they are interacting on a daily basis are present and provide a sense of support as we could see in the sessions we performed. • The facilitator will sit next to the child during the session and his/her role is to fill in the observation form while interacting with the child to make them feel at ease. We rocomend the participant doesnot write himself the questionary. They use to have problems to write or read. • There are two questionnaires that need to be assessed which are demographic questionnaire and post task questionnaire but we propose specifics questionaies adapted to participants with Down Syndrome. We designed this questionaries with Special Psicologist whoes work every day with those participants , taking as base the SUS questionarie. • Post task questionnaire will be conducted right after each test session with the help from the Tutor or Parent. Facilitador should be close just in case the participant have a dubt. • Take note of the times when the participant asks for help. • Use simple words when directing the participants. When you explain to the participants the tasks, process, objectives etc, you should use a esy vocabulary and you must to speak • Speak slowly slow and some times is necessary repeat the same idea because the participats have a problem to concentrate their attention. • Schedule a break halfway through the test session and remind the participants that they can stop at anytime.
Documents	The documents should be applied -UsatestDown Demographic Questionary (Annexes) -UsatestDown SUS Questionary adapted to people with Down Syndrome focused in the whole process -UsatestDown SUS Questionary to Tutors focused in the whole process (Annexes)

5. Refine the test plan after analysing the results of the pilot tests.

Table 6: Usatestdown: Refine the test plan after analyzing the results of the pilot tests.

USATESTDOWN	
Step :	***Refine the test plan after analysing the results of the pilot tests.***
Definition	Once analyzing the results of the pilot test, modifications may be made to the protocol, instructions, task data, task sequencing, interview questions, etc., if necessary.
Biography Research:	NO CONTRIBUTIONS
Usatestdown	If an error is encountered in the test pilot, it is necessary to make an immediate change to the plan and execute a second session. The appropriate corrections should be taken.

6. Recruit participants

Table 7: Usatestdown: Recruit participants

USATESTDOWN	
Step :	***Recruit participants***
Definition	Process to determine the type and number of participants needed for the usability tests.
Biography Research:	From the analysis of the research with regard to the recruitment of participants, we find that [10] take four children aged from 6 to 12 years with DS, [13] use five usability experts and six learners, [12] use from three to five interface design and learning content experts, and [20] work with two paediatricians, primary school teachers and 11 children with DS. This illustrates the importance of working with on average 10 paediatricians, interface and learning content evaluators and people with DS. The activities specified by [11] are validate the criteria for recruiting participants, like computer experience. Nielsen's study showed that a group of five users with different background, mixed gender and aged five to six years old , they were able to find about 80% of the findings in a system.
Usatestdown	• The first step in recruiting young participants is sending information about the study to the places whoes are working with the particpants profile that we want to work, it means, . • Tutor should recommend the particcipants proffile because the mental age is different than the bilogical age. We should not only take on count the participants age, we should analize what are the especial skils that every participant have. In our case at the beginning we found participants with low mental dishabilities and another so extrem. It makes not homogeneous group because we were not evaluating really the aplication, we were evaluating just the cognitive disabilities. • The facilitator should ask the participant if he/she wants to colaborate because the participant should be voluntier. We had a case with a participant who behaved on a rude way . We asked him if he want to participate and he did not want. It is the best way to evaluate because some times they feel pushed to contribute. • Don't push them to finish fast.

7. Run the test session

Table 8: Usatestdown: Run the test session

USATESTDOWN	
Step :	***Run the test session***
Definition	This is the essential part of the evaluation because it is here that the usability evaluation is completed. (1) Welcome; (2) Ask the participants to carry out the tasks; (3a) If performance is measured, measure the times, (3b) If performance is not measured, interrupt the user to clarify their decisions; (4) Note the number of errors and other objective data; (5) Distribute a satisfaction questionnaire and complete a personal interview.
Biography Research:	[14] collect the data iteratively from people with DS in Phase 1. Another aim is identify the suitability of the teaching material for the learning problems that students are set. [13] describe the testing steps: execute evaluation, write report, submit report to immediate evaluator, okay report, and compile evaluation reports.
Usatestdown	• Do not complete the final test on the same day as the pilot testing because the users will be tired and confused if the first pilot process failed. • It is recommended to execute the complete test from the beginning, including the changes that were made to the test plan after the pilot. • Consider the reactions of the people being evaluated for each of the tasks that they complete. It is very important to determine their satisfaction level and the improvements that could be made in the next version. • Solve all of the questions that the user has during the process. • After completing the usability test session with a down syndrome participant, facilitator needs to ask the child to answer post task questionnaire. • Take note of the times when the participant asks for help. • Don't push the people to participate if they don't want, it may generate a bad atmosphere to work. • Don't push the people to answer if they don't want, it may generate a bad atmosphere to work. • Use simple words when directing the participants. When you explain to the participants the tasks, process, objectives etc, you should use a esy vocabulary and you must to speak • Speak slowly slow and some times is necessary repeat the same idea. • Take note of the times when the participant asks for help. • Schedule a break halfway through the test session and remind the participants that they can stop at anytime.
Documents	-UsatestDown Demographic Questionary (Annexes) -UsatestDown SUS Questionary adapted to people with Down Syndrome focused in the whole process -UsatestDown SUS Questionary to Tutors focused in the whole process (Annexes)

8. Analyse the collected information

Table 9: Usatestdown: Analyze the collected information

USATESTDOWN	
Step :	***Analyse the collected information***
Definition	Analyze the objective data (times, errors, etc.), the more subjective data (satisfaction questionnaire and interviews), and all of the data that contributes to understanding the behavior of the evaluated people from the usability test. The objective is to identify usability problems and propose improvements.
Biography Research:	The DEVAN method is based on the structured analysis of video material captured during user tests and was developed to detect usability problems in task-based products for adults. When used for evaluation with children, this method can be adjusted for the detection of usability and fun problems [21].
Usatestdown	• It is important to analyze the data with all of the parameters collected from the people who participated in the usability test. This involves qualitative content (logs) as well as quantitative (user reactions). • Conduct an analysis of the part that appeared qualitatively in the evaluation and the quantitative data results. It is not always the same result. • The tool should take the time automatically. • The tool should help the evaluation. • Success, Satisfaction and Frustration Rate per Task and Document • This includes; the people feeling, fun, ease of use and their satisfaction level towards the game • Data are collected while observing the child performing the task scenarios.

9. Report results to the development team or management.

Table 10: Usatestdown: Report results to the development team or management.

USATESTDOWN	
Step:	***Report results to the development team or management.***
Definition	Prepare a presentation or report to explain the usability problems that were encountered and how they can be improved.
Biography Research:	NO CONTRIBUTIONS
Usatestdown	• The results presentation should be done with all of the members of the group, with a clear document, and with the respective backups. • -In the case that the results were not satisfactory, improvements to the system should be made and it should be executed again, following the USA TESTDOWN guide. • Apply the Ethical Issues in Recruiting Participants, it means follow the rules that each centre have to manage the participant information with ethical process. We should safe in a private and confidential way the collected information • In the Maria Corredentora [22] case, we did a inform with the most important points because they use this information to improve the way to teach.

4. CONCLUSIONS AND FUTURE WORK

In general, we can see that it was necessary to adapt the SUS questionnaire for the persons with Down syndrome because it is a complex survey for participants. SUS was modified to evaluate the USATESTDOWN process, the guide, which was designed with the expert tutors who work with the participants daily.

In general, it is clear that the guide is viable and can be successfully used and modified to the needs of persons with Down syndrome, with this as an example of a real-world success. It was also evaluated by the expert tutors as part of this process, which was a great help and supported the adaptation of the guide. The participation of the expert tutors is very important as their experience greatly contributed to the implementation of the test, following the guide. Additionally, it is critically important to include the expert tutors with the interaction with the participants'

It is necessary also the previous interactions with the application to create a comfortable and familiar environment so the participants feel safe and trust the process as they are asked questions or doing a task. We recommend that times are not as strict and participants are able to work with as much flexibility as possible. The time parameter set by the first participant to force the second participant to complete the task in the same amount of time was not always produce the same cognitive or memory coefficients.

We suggest an evaluation stage where devices are given back to them to determine how much time is necessary for them to work independently from the tutors and then independent of the application and able to do the activity without the support of a tutor or the application.

REFERENCES

[1] J. Jadán-Guerrero, L. Guerrero, G. López, D. Cáliz, and J. Bravo, "Creating TUIs Using RFID Sensors—A Case Study Based on the Literacy Process of Children with Down Syndrome," Sensors, vol. 15, no. 7, pp. 14845–14863, 2015.

[2] L. M. Normand, "No Title."

[3] A. Brandão, E. Passos, C. Vasconcelos, A. Conci, E. Clua, P. T. Mourão, and M. Cordeiro, "Stimulating imitation of children with Down syndrome using a game approach," VIII Brazilian Symp. Games Digit. Entertain., pp. 97–100, 2009.

[4] J. R. Joseph S. Dumas, A Practical Guide to Usability Testing. .

[5] A. Lepistö, "Usability evaluation involving participants with cognitive disabilities," Proc. third Nord. Conf. Humancomputer Interact. Nord. 04, pp. 305–308, 2004.

[6] J. Nielsen and M. Kaufmann, "Usability Engineering," p. 340, 1993.

[7] J. Nielsen, "Heuristic Evaluation," Usability Insp. Methods, pp. 25–62, 1994.

[8] A. Brand??o, D. G. Trevisan, L. Brand??o, B. Moreira, G. Nascimento, C. N. Vasconcelos, E. Clua, and P. T. Mour??o, "Semiotic inspection of a game for children with Down syndrome," Proc. - 2010 Brazilian Symp. Games Digit. Entertain. SBGames 2010, no. August 2016, pp. 199–210, 2011.

[9] R. Pal, “On the Lewis-Nielsen model for thermal/electrical conductivity of composites,” Compos. Part A Appl. Sci. Manuf., vol. 39, no. 5, pp. 718–726, 2008.

[10] I. Macedo and D. G. Trevisan, “A Method to Evaluate Disabled User Interaction : A Case Study with Down Syndrome Children,” Univers. Access Human-Computer Interact. Des. Methods, Tools, Interact. Tech. eInclusion, pp. 50–58, 2013.

[11] L. Kumin and J. Lazar, “A Usability Evaluation of Workplace-Related Tasks on a Multi-Touch Tablet Computer by Adults with Down Syndrome,” J. Usability …, vol. 7, no. 4, pp. 118–142, 2012.

[12] R. Ramli and H. B. Zaman, “Designing usability evaluation methodology framework of Augmented Reality basic reading courseware (AR BACA SindD) for Down Syndrome learner,” Proc. 2011 Int. Conf. Electr. Eng. Informatics, ICEEI 2011, no. July, 2011.

[13] F. Adebesin, P. Kotzé, and H. Gelderblom, “The complementary role of two evaluation methods in the usability and accessibility evaluation of a non-standard system,” Proc. 2010 Annu. Res. Conf. South African Inst. Comput. Sci. Inf. Technol. - SAICSIT ’10, pp. 1–11, 2010.

[14] R. L. Yussof and H. Badioze Zaman, “Usability evaluation of multimedia courseware (MEL-SindD),” Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 5857 LNCS, pp. 337–343, 2009.

[15] D. Cáliz, L. Martínez, X. Alamán, C. Terán, and R. Cáliz, “‘ U Sability Testing in Mobile Applications Involving People With Down Syndrome : a Literature Review ,’” ICAIT 2016 Conf., 2016.

[16] Asindown ORG, “Asindown Valencia,” 2017. [Online]. Available: http://asindown.online/. [Accessed: 10-Jan-2017].

[17] M. Corredentora, “Ma. Corredentora,” 2017. [Online]. Available: http://mariacorredentora.org/wpmcorg/. [Accessed: 10-Jan-2017].

[18] A. ORG, “Apadema,” 1971. [Online]. Available: http://www.apdema.org/tag/madrid/.

[19] “Fundación PRODIS,” 2016. [Online]. Available: http://www.fundacionprodis.org/. [Accessed: 14-Jun-2016].

[20] R. L. Yussof and T. N. S. T. Paris, “Reading Activities Using the Scaffolding in MEL-SindD for Down Syndrome Children,” Procedia - Soc. Behav. Sci., vol. 35, no. December 2011, pp. 121–128, 2012.

[21] J. Marco, E. Cerezo, and S. Baldassarri, “Bringing tabletop technology to all: Evaluating a tangible farm game with kindergarten and special needs children,” Pers. Ubiquitous Comput., vol. 17, no. 8, pp. 1577–1591, 2013.

[22] “Maria Corredentora Centre,” 2016. [Online]. Available: http://mariacorredentora.org/wpmcorg/ . [Accessed: 11-Nov-2016].

AUTHORS

Ing. MSc. Doris Cruz Caliz Ramos.

- Computer Sciences Engineering
- Master in Management of Information Technology and Communications National
- Polytechnic School Ecuador. 2008 - 2012
- International Leadership Training. Germany. 2011 - 2012
- PHD Student in Polytechnic School Madrid. 2013- 2017
- Academic Visitor in Middlesex University London. 2015 – 2016
- Academic Visitor Karlsruhe Institute of Technology (KIT). Pervasive Computing Systems - TecO

Doctor. Loic Antonio Martinez Normand

- Professor Department ETSIINF, DLSIIS, Madrid Polytechnic University. 2008 – Today.
- Researcher in Group Investigation on Technology Informatics and Communications: CETTICO.
- President Sidar Foundation. 2002 – Today

Ing.MSc. Richarth Harold Caliz Ramos.

- Master in Management of Information Technology and Communications MSc, final mark: cum laude. National Polytechnic School (EPN), Quito, Ecuador (Fall 2008-Winter 2010)
- Telecommunications and Electronics Enineering, final mark: cum laude. National Polytechnic School (EPN), Quito, Ecuador (Fall 1995-Winter 2002)

A WIND POWER PREDICTION METHOD BASED ON BAYESIAN FUSION

Jianqi An[1,2,3], Zhangbing Chen[1,2], Min Wu[1,2], Takao Terano[3],
Min Ding[1,2], and Hua Xie[1,2]

[1]School of Automation, China University of Geosciences,
Wuhan, 430074, China
[2]Hubei key Laboratory of Advanced Control and Intelligent Automation for
Complex Systems, Wuhan, 430074, China
[3]Department of Computer Science, School of Computing, Tokyo Institute of
Technology, Yokohama, 226-8502, Japan

ABSTRACT

Wind power prediction (WPP) is of great importance to the safety of the power grid and the effectiveness of power generations dispatching. However, the accuracy of WPP obtained by single numerical weather prediction (NWP) is difficult to satisfy the demands of the power system. In this research, we proposed a WPP method based on Bayesian fusion and multi-source NWPs. First, the statistic characteristics of the forecasted wind speed of each-source NWP was analysed, pre-processed and transformed. Then, a fusion method based on Bayesian method was designed to forecast the wind speed by using the multi-source NWPs, which is more accurate than any original forecasted wind speed of each-source NWP. Finally, the neural network method was employed to predict the wind power with the wind speed forecasted by Bayesian method. The experimental results demonstrate that the accuracy of the forecasted wind speed and wind power prediction is improved significantly.

KEYWORDS

Wind Power Prediction, Numerical Weather Prediction, Bayesian Fusion, Wind Speed Prediction

1. INTRODUCTION

Wind power is a renewable, clean, plentiful and widely distributed energy, and does not produce greenhouse gas during operation [1-3]. WPP is of great importance to the safety of the power system and the effectiveness of power generations dispatching, which has become one of the most attractive research fields [4-6]. As wind power is major determined by the wind speed, the most key factor to predict accurate wind power is to forecast accurate wind speed. However, the wind speed is easily affected by many factors, such as air pressure, cloud chart, temperature, etc. These factors make the wind speed fluctuant and hard to be forecasted [7]. Therefore, the accurate forecasting of the wind speed is of great significance to achieve the effective prediction of wind power.

There are two main types of WPP approaches: physical approaches and statistical approaches. The physical approaches means that the wind power is predicted by the mathematical model of the generation process which is very difficult to build and need many hard-measured parameters.

David C. Wyld et al. (Eds) : ITCS, SIP, CST, ARIA, NLP - 2017
pp. 27– 37, 2017. © CS & IT-CSCP 2017 DOI : 10.5121/csit.2017.70303

The statistical approaches to predict wind power use the weather forecast, combining with the operating condition, to obtain the value of the wind power based on the historical values relationship between the power and NWP. WPP is considered at different time scales: long-term prediction (several days or weeks, even mouths ahead), short-term prediction (1-2 days ahead) and very short-term prediction (several minutes to hours ahead). The difficulty of short-term prediction of wind power is much more than that of the very short-term prediction; while the required accuracy of short-term prediction is much higher than that of the long-term prediction.

Many literatures have demonstrated some good results about the short-term wind power prediction. In [8], a wind-forecasting method based on intrinsic time-scale decomposition (ITD) and least squares support vector machine (LS-SVM) was proposed. In [9], a time series model of WPP at different time scales was studied. In this study, with the step of the prediction increasing, the error would rise up, since the time series model used its own previous forecast value. In [10], a wind power prediction model based on empirical mode (EMD) and support vector machine (SVM) was proposed to reduce the influence of non-linear and non-stationary characteristics on wind power prediction. In [11], GUO Pengfei et al. presented a short-term wind power prediction method based on genetic algorithm to optimize RBF neural network. The RBF neural network was optimized by genetic algorithm to get the best value of the weights, base width and centre of the basis function of RBF neural network. In [12], a method based on sliding window weighted recursive least squares for wind power prediction was proposed. The historical data in the method was weighted so that the model would be able to adapt the varied environment. In [13], the constraint factor particle swarm optimization was used to optimize the parameters of the autoregressive model to accurately predict the wind power.

Overall, scholars have proposed many methods for WPP, but most of the researches only use the single NWP data to predict the wind power. The NWP data cannot always accurately predict the actual weather, so using only one NWP can hardly ensure the accuracy of wind power prediction. In this work, a new WPP method based on Bayesian method and multi-source NWPs was proposed. In this method, three independent NWPs given by different weather forecast companies, called multi-source NWPs, are employed and fused using Bayesian method to predict wind power. The three independent NWPs have different characteristics; each one has its own advantages and disadvantages. This fusion method can fully use the complementarity of the three NWPs, which can effectively improve the statistical accuracy of WPP.

2. PROBLEM DESCRIPTION AND SCHEME OF WPP

It is generally believed that the power of wind generation is directly proportional to the wind speed. In the meantime, the temperature, humidity, pressure and other weather factors have a certain impact on the power of wind generation. But based on our research, in some wind farms, the impact on the wind power caused by these factors is limited. We established a relational model between the actual measured wind speed and the actual wind power by using a neural network. The accuracy of the model is over 95%, which means that the wind power can be predicted by introducing the wind speed as a single input. Therefore, the key problem of WPP for this type of wind farm is how to obtain an accurate forecasting wind speed.

In the research, there are three independent NWPs provided by different weather forecast companies. The problem we need to solve is to predict the wind power generation of the wind farm at intervals of 15 minutes from 00:00 to 24:00 of the next day according to the multi-sources NWPs. Thus there are 96 output prediction points which is used for dispatching power generations by power grid companies.

In this research, first, a relational model between wind speed and wind power was established based on a back propagation neural network (BPNN) which was trained by the historical data of actual measured wind speed and actual wind power. By using the data driven method, the inherent relationship of the wind power generation is described. Then, instead of using one NWP, a Bayesian method was designed to fuse the three independent NWPs to calculate the wind speed which is more accurate than that of any of the three NWP. Finally, the wind speed forecasted by the Bayesian method is imported into the relational model to predict the wind power.

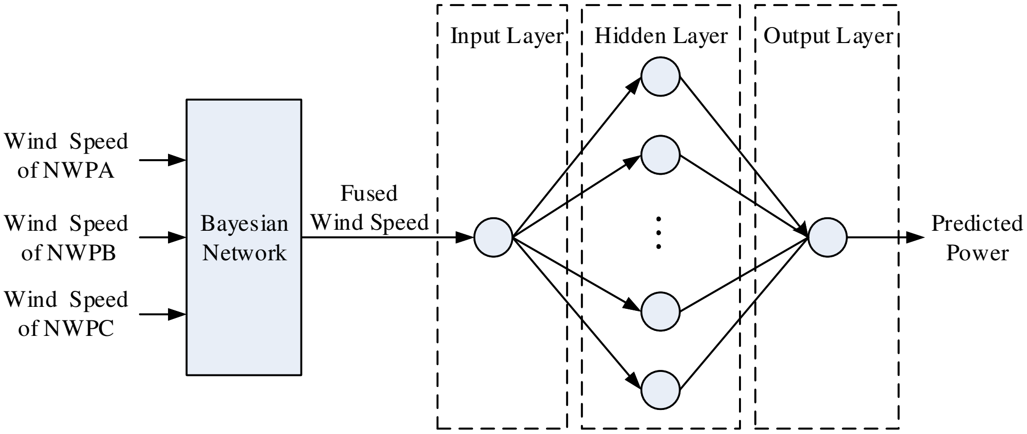


Figure 1. The scheme diagram of the proposed WPP method

In this paper, three layers BPNN structure is applied to build the relational model between wind speed and wind power, which contains one input node which is the wind speed and one output node which is the wind power. The incentive functions of the model are *logsig* and *purelin*, respectively. Since the basic idea of the BPNN is clear, this paper will not discuss further.

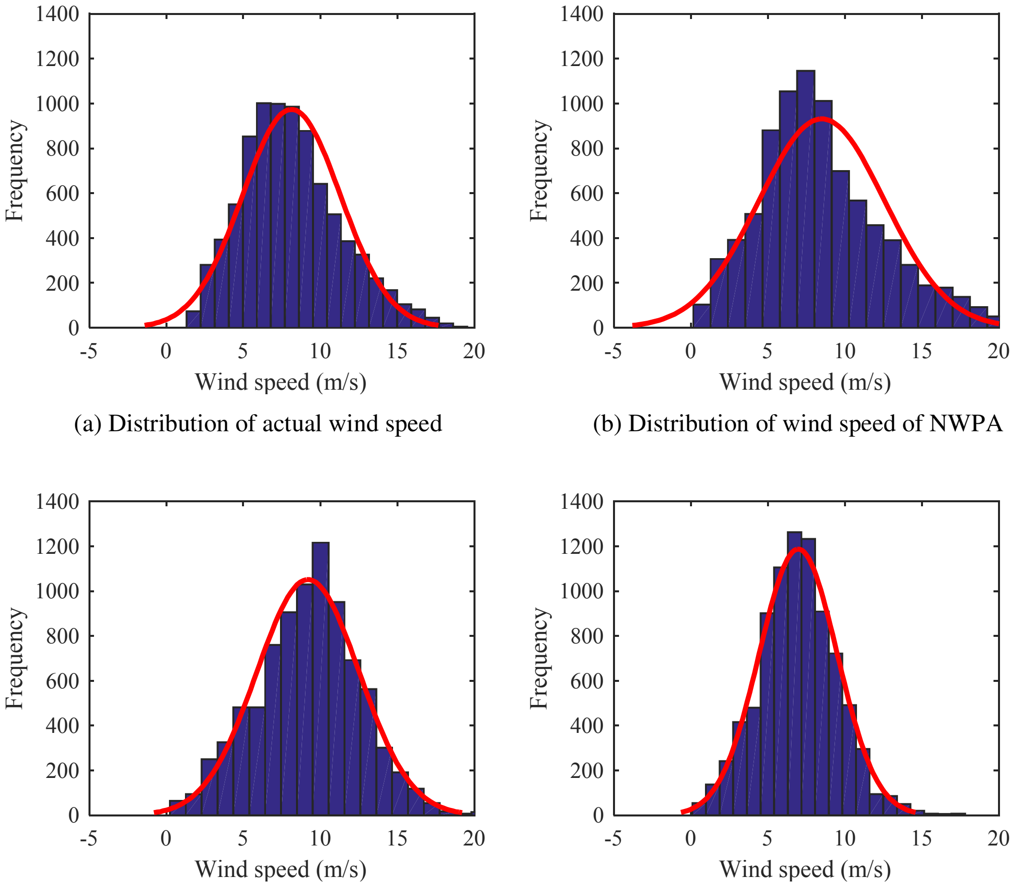
3. Analysis of Wind Speed Characteristics of Independent NWPs

Three NWPs, defined as NWPA, NWPB and NWPC, have different characteristics. In the view of the instantaneous characteristics, sometimes NWPA is more accurate; sometimes NWPB and/or NWPC are/is more accurate. In the view of statistical characteristics, the wind speed values of the three NWPs and the actual measured wind speed value basically follow the normal distribution. The statistical characteristics of the historical actual wind speed and the wind speed of historical three NWPs are shown in Figure 2, which illustrates that the means and variances of each wind speed value are different. The values of the means and variances of each wind speed are shown in the Table 1.

Figure 3 (a), (b) and (c) show the comparison of forecasting wind speed of each NWP and the actual measured speed value in a certain period of history. As is shown the figure, the errors between the forecasting wind speed and the actual speed are fluctuating. The three kinds of wind speed of the NWPs fluctuate around a certain value, which is shown in Figure 4. Figure 4 (a), (b) and (c) show the fluctuation of the wind speed of NWPA, NWPB and NWPC when the actual measured wind speed is between 6m/s and 6.5m/s. As mentioned above, sometimes NWPA is more accurate; sometimes NWPB and/or NWPC are/is more accurate. Therefore, a suitable method should be considered to fuse the three forecasting wind speed in order to calculate a more accurate wind speed.

Table.1 Mean value and variance of each wind speed

	Actual wind speed	Wind speed of NWPA	Wind speed of NWPB	Wind speed of NWPC
Mean value	7.78	7.84	9.01	6.73
Variance	10.09	13.96	11.93	6.49



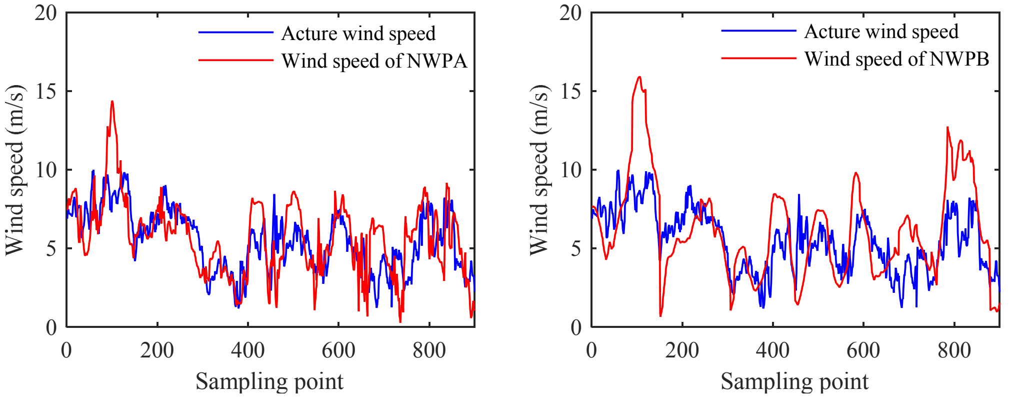
(a) Distribution of actual wind speed

(b) Distribution of wind speed of NWPA

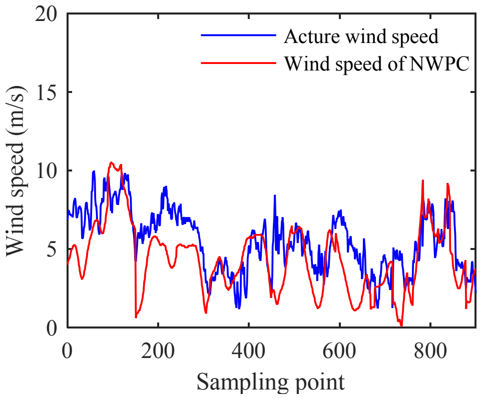
(c) Distribution of wind speed of NWPB

(d) Distribution of wind speed of NWPC

Figure 2. Distribution of each wind speed

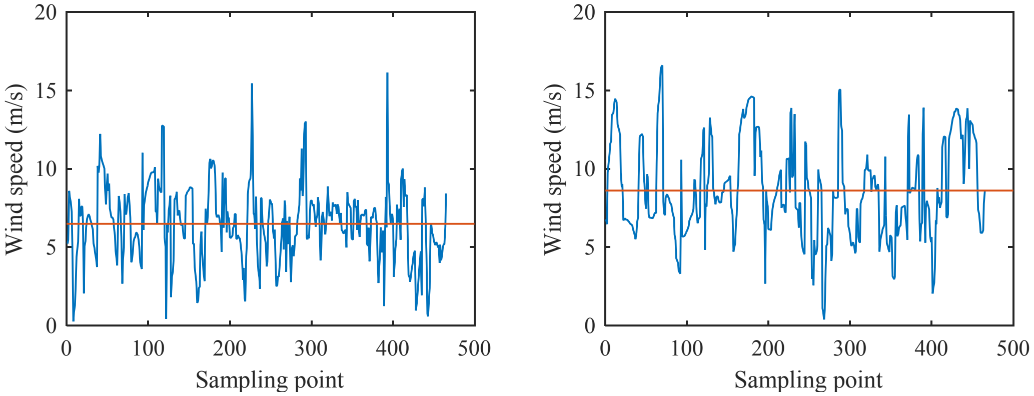


(a) Actual wind speed and wind speed of NWPA (b) Actual wind speed and wind speed of NWPB

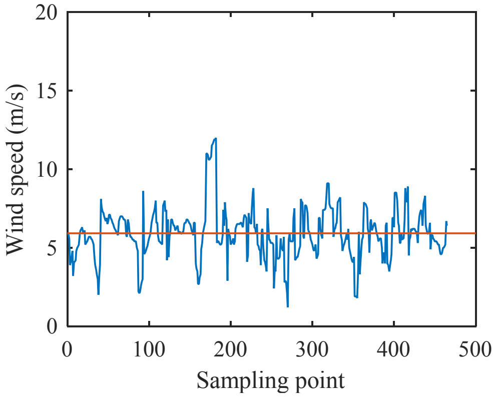


(c) Actual wind speed and wind speed of NWPC

Figure 3. The relationship between actual wind speed and wind speed of each NWP



(a) Fluctuation of wind speed of NWPA (b) Fluctuation of wind speed of NWPB



(c) Fluctuation of wind speed of NWPA

Figure 4. Fluctuation of the forecasting wind speed of each NWP when the actual wind speed is between 6m/s and 6.5m/s.

4. Bayesian Method for Fusing the Multi-source NWPs

In the view of the statistical characteristics, the wind speed values of three NWPs basically meet the normal distribution, but their feature are distinct account of different means and variance. It is possible to achieve a more accurate forecasting wind speed by fusing these three original forecasting wind speed values of the NWPs. The fused wind speed is more effective to predict the wind power.

In this paper, the three wind speed values of NWPs, defined as x_A , x_B and x_C , are fused by using Bayesian method. Since the means of the history data of x_A , x_B and x_C are not equal, we make a transformation to let the means of them be equal in order to facilitate the solution. The transformed wind speeds are defined as x'_A , x'_B and x'_C .The transformation method is shown in equation (1),

$$\begin{aligned} x'_A &= x_A + (u_r - u_A), \\ x'_B &= x_B + (u_r - u_B), \text{ and} \\ x'_C &= x_C + (u_r - u_C), \end{aligned} \tag{1}$$

where u_r denotes the mean value of the actual wind speed; u_A , u_B , and u_C denote the mean values of wind speed of NWPA, NWPB, and NWPC. After the transformation, the mean values of the three wind speed of NWPs are equal. Then adopting the Bayesian method, the transformed wind speed values are fused into one.

The fused wind speed is defined as *x*. The Bayesian fusion equation is given by

$$p(x \mid x'_A, x'_B, x'_C) = \frac{p(x'_A, x'_B, x'_C \mid x) p(x)}{p(x'_A, x'_B, x'_C)} = \frac{p(x'_A \mid x) p(x'_B \mid x) p(x'_C \mid x) p(x)}{p(x'_A, x'_B, x'_C)} \tag{2}$$

where $p(x)$ is the probability of the wind speed *x*; $p(x'_A, x'_B, x'_C \mid x)$ is the probability of forecast wind speed values of x'_A , x'_B , and x'_C when the wind speed is *x*; is the $p(x \mid x'_A, x'_B, x'_C)$ is the probability of the wind speed *x* when the forecast wind speed values of NWPA, NWPB, and NWPC are x'_A , x'_B and x'_C , respectively. $p(x)$ and $p(x'_A, x'_B, x'_C \mid x)$ are priori probabilities which can be calculated by the historical data. $p(x \mid x'_A, x'_B, x'_C)$ is the posteriori probability of the wind speed *x*. In this method, the purpose is to find the *x* whose posteriori probability $p(x \mid x'_A, x'_B, x'_C)$ is the highest.

The wind speed meet normal distribution, they are described as

$$\begin{aligned} x &\sim N(x_0, \sigma_0^2), \\ x'_A &\sim N(u_r, \sigma_A^2), \\ x'_B &\sim N(u_r, \sigma_B^2), \text{ and} \\ x'_C &\sim N(u_r, \sigma_C^2). \end{aligned} \tag{3}$$

Let $a = \dfrac{1}{p(x'_A, x'_B, x'_C)}$ which is a constant. The equation (2) can be transformed to

$$\begin{aligned} p(x \mid x'_A, x'_B, x'_C) &= a\frac{1}{\sqrt{2\pi\sigma_A}}\exp\{-\frac{1}{2}[\frac{x'_A - u_r}{\sigma_A}]^2\}\times\frac{1}{\sqrt{2\pi\sigma_B}}\exp\{-\frac{1}{2}[\frac{x'_B - u_r}{\sigma_B}]^2\} \\ &\times\frac{1}{\sqrt{2\pi\sigma_C}}\exp\{-\frac{1}{2}[\frac{x'_C - u_r}{\sigma_C}]^2\}\times\frac{1}{\sqrt{2\pi\sigma_0}}\exp\{-\frac{1}{2}[\frac{r - r_0}{\sigma_0}]^2\} \\ &=\text{aexp}\{-\frac{1}{2}\left[\frac{x'_A - u_r}{\sigma_A}\right]^2 -\frac{1}{2}\left[\frac{x'_B - u_r}{\sigma_B}\right]^2 -\frac{1}{2}\left[\frac{x'_C - u_r}{\sigma_C}\right]^2 -\frac{1}{2}\left[\frac{x - x_0}{\sigma_0}\right]^2\} \\ &\times\frac{1}{\sqrt{2\pi\sigma_A}}\times\frac{1}{\sqrt{2\pi\sigma_B}}\times\frac{1}{\sqrt{2\pi\sigma_C}}\times\frac{1}{\sqrt{2\pi\sigma_0}} \end{aligned} \tag{4}$$

where σ_A, σ_B, and σ_C denote the root mean square of x'_A, x'_B and x'_C, respectively.

The exponential part in the above equation is a quadratic function with respect to *x,* so $p(x \mid x'_A, x'_B, x'_C)$ is still normal distribution. It can be described as $N(x_N, \sigma_N^2)$. Then

$$p(x \mid x'_A, x'_B, x'_C) = \frac{1}{\sqrt{2\pi\sigma_N}}\exp\{-\frac{1}{2}[\frac{x - x_N}{\sigma_N}]^2\}. \tag{5}$$

x_N can be obtained by comparing the parameters between the equation (4) and (5),

$$x_N = [\frac{x'_A}{\sigma_A^2} + \frac{x'_B}{\sigma_B^2} + \frac{x'_C}{\sigma_C^2} + \frac{x_0}{\sigma_0^2}] / [\frac{1}{\sigma_A^2} + \frac{1}{\sigma_B^2} + \frac{1}{\sigma_C^2} + \frac{1}{\sigma_0^2}]. \tag{6}$$

So fused wind speed based on the Bayesian estimation is $\hat{x}$

$$\hat{x} = \int_\Omega x\frac{1}{\sqrt{2\pi\sigma_N}}\exp\{-\frac{1}{2}[\frac{x - x_N}{\sigma_N^2}]\}dx = x_N . \tag{7}$$

The wind speed calculated by Bayesian method is used to predict the wind power by the BPNN relational model between wind speed and wind power.

5. EXPERIMENTS AND RESULTS ANALYSIS

In this paper, the relational model between wind speed and wind power is built by a three layer BPNN. The model is trained and testified by historical data of the actual measured wind speed and wind power. Figure 5 shows the actual wind power and calculated wind power by actual wind speed. It illustrates that the accuracy of the model is above 95% which means that the model is able to describe the relationship between the wind speed and power. The data is obtained from one certain wind farm of China from March 2016 to May 2016.

Figure 6 shows the actual wind speed, forecasted wind speed of NWPA, NWPB, and NWPC and the fused wind speed based on Bayesian method. From Figure 6 we can find that the fused wind speed is more closed to the actual measured wind speed than the three wind speed of NWPA, NWPB, and NWPC. The mean squared errors (MSE) between actual wind speed and the speed forecasting accuracy of the fused wind, wind speed of NWPA, NWPB, and NWPC are 0.84, 1.58, 2.22, and 0.91, respectively.

The fused wind speed, wind speed of NWPA, NWPB, and NWPC were used to predict the wind power as the input of the BPNN, respectively. The results were shown in Figure 7 which shows that the MSE between actual wind speed and the speed forecasting accuracy of the fused wind,

wind speed of NWPA, NWPB, and NWPC are 3.02, 7.13, 4.32, and 4.23, respectively. The predicted power by fused wind speed is more accurate than the original forecasted wind speed of NWPA, NWPB, and NWPC.

The results demonstrate that the proposed Bayesian fusion method can achieve more accurate forecasting wind speed than that forecasted by the original NWPs. By using the fused wind speed, the accuracy of WPP is also much improved.

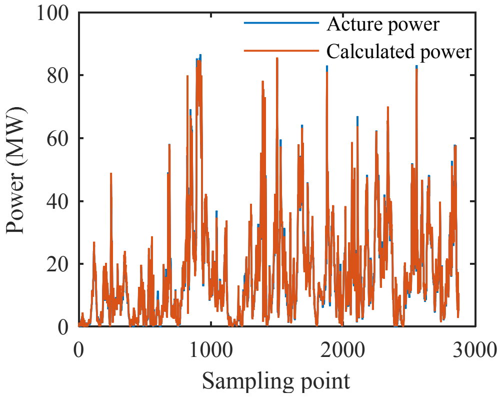


Figure 5. Actual wind power and calculated wind power by actual wind speed based on BPNN

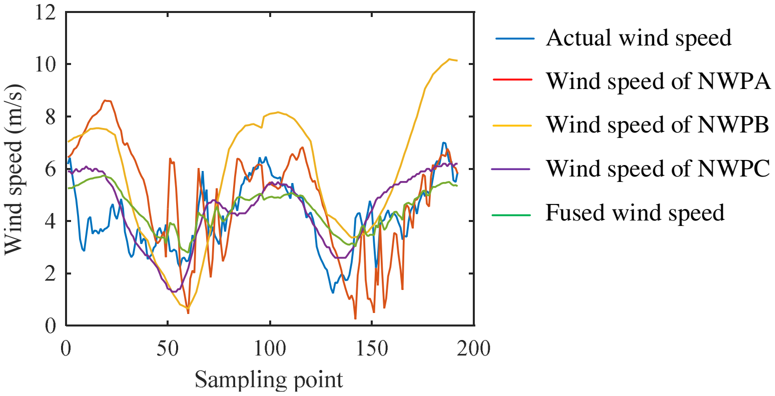


Figure 6. Actual wind speed, fused wind speed, and wind speed of NWPA, NWPB, and NWPC

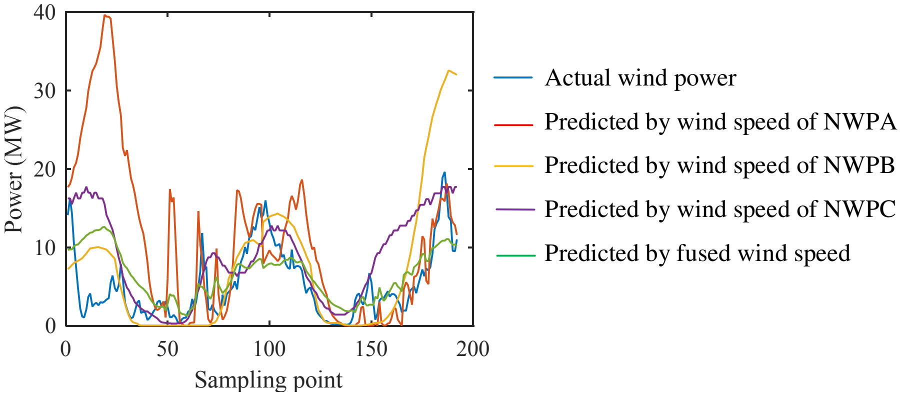


Figure 7. Actual wind power and wind power predicted by fused wind speed, the wind speed of NWPA, NWPB, and NWPC

6. CONCLUSION

This paper presents a new WPP method based on Bayesian method and multi-source NWPs. The proposed method fuses the three NWPs to get more accurate wind speed value; and the fused wind speed is employed to predict the wind power, the accuracy of which can satisfy the demand of the power grid companies. This method has the advantage of lower calculation, higher reliability and practicability.

ACKNOWLEDGEMENTS

The first author is an overseas researcher under Postdoctoral Fellowship of Japan Society for the Promotion of Science. This work was supported by the Hubei Provincial Natural Science Foundation of China under Grants No. 2016CFB480 and No. 2015CFA010, National Natural Science Foundation of China under Grants No. 61333002 and No. 61203017, and the Foundation Research Founds for China University of Geosciences under Grant No. 2015349120.

REFERENCES

[1] Rui Pinto. Botterud, Vladimiro Miranda, & Jean Sumaili, (2011) “Wind power forecasting uncertainty and unit commitment”, Applied Energy, vol.88, pp. 4014-4023.

[2] Jae-kun Lyu, J. Jae-Haeng Heo, Mun-Kyeom Kim, & Jong Keun Park, (2013) “Impacts of wind power integration on generation dispatch in power systems”, Journal of Electrical Engineering & Technology, vol.8, pp. 453-463.

[3] Ch. Ulam-Orgil, Hye-Won Lee, & Yong-Cheol Kang, (2012) “Evaluation of the wind power penetration limit and wind energy penetration in the mongolian central power system”, Journal of Electrical Engineering & Technology, vol.7, pp. 852-858.

[4] Saurabh S. Soman, Hamidreza Zareipour, Om Malik, & Paras Mandal, (2010) “A review of wind power and wind speed forecasting methods with different time horizons”, Proceedings of 2010 North American Power Symposium, pp. 1-8.

[5] G. Xydis, C. Koroneos, & M. Loizidou, (2009) “Exergy analysis in a wind speed prognostic model as a wind farm sitting selection tool: A case study in Southern Greece”, Applied Energy, vol.86, pp. 2411-2420.

[6] Colm Lowery & Mark O'Malley, (2012) “Impact of wind forecast error statistics upon unit commitment”, IEEE Transactions on Sustainable Energy, vol.3, pp. 760-768.

[7] Ramesh Babu. N & P. Arulmozhivarman. P, (2013) “Improving forecast accuracy of wind speed using wavelet transform and neural networks”, Journal of Electrical Engineering & Technology, vol. 8, pp. 559-563.

[8] Xiaojuan Han, Xilin Zhang, Fang Chen & Zhihui Song, (2012) “Short-term wind speed prediction method based on time series combined with LS-SVM”, Proceedings of the 31st Chinese Control Conference, pp. 7593-7597.

[9] Andrew Kusiak, Haiyang Zheng, & d Zhe Song, (2009) “Short-term prediction of wind farm power:a data mining approach”, IEEE Transactions on Energy Conversion, vol.24, pp.125-136.

[10] Wendan Zhang, Fang Liu, Xiaolei Zheng, & Yong Li, (2015) “A hybrid EMD-SVM based short-term wind power forecasting model”, Proceedings of 2015 IEEE PES Asia-Pacific Power and Energy Engineering Conference, pp. 1-5.

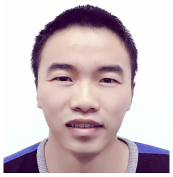
[11] GUO Pengfei, QI Zhiyuan, & HUANG Wei, (2016) "Short-term wind power prediction based on genetic algorithm to optimize RBF neural network", Proceedings of the 28th Chinese Control and Decision Conference, pp.1220-1223.

[12] Yanfeng Ge. Peng Liang. Liqun Gao, (2015) "A self-adaptive model for wind power prediction", Proceedings of the 27th Chinese Control and Decision Conference, pp. 1165-1169.

[13] Adnan Anwar & Abdun Naser Mahmood, (2014) "Enhanced estimation of autoregressive wind power prediction model using constriction factor particle swarm optimization", Industrial Electronics and Applications, pp. 1136-1140.

AUTHORS

Jianqi An is an associate professor of School of Automation, China University of Geosciences, and an overseas researcher (2016.11- 2018.11) in Department of Computer Science, School of Computing, Tokyo Institute of Technology under Postdoctoral Fellowship of Japan Society for the Promotion of Science. He received his B.S. in 2004, M.S. in 2007, and Ph.D. in 2011 from Central South University. His research interests include detection, Modelling, and advanced control methods for complex process and their applications etc. He is the member of the Chinese Association of Automation (CAA). He published more 20 academic papers, applied for 21 patents, published 2 textbooks.



Zhangbing Chen is a postgraduate student of China University of Geosciences. He was born in Hubei, China, 1992. He received his B.S. from the Hubei University of Technology in 2016. His research interests include wind power prediction, information fusion and data mining.

Min Wu received his B.S. and M.S. degrees in engineering from Central South University, Changsha, China, in 1983 and 1986, respectively, and his Ph.D. degree in engineering from the Tokyo Institute of Technology, Tokyo, Japan, in 1999. He was a faculty member of the School of Information Science and Engineering at Central South University from 1986 to 2014, attaining the position of full professor. In 2014, he moved to the China University of Geosciences, Wuhan, China, where he is a professor in the School of Automation. He was a visiting scholar with the Department of Electrical Engineering, Tohoku University, Sendai, Japan, from 1989 to 1990, and a visiting research scholar with the Department of Control and Systems Engineering, Tokyo Institute of Technology, from 1996 to 1999. He was a visiting professor at the School of Mechanical, Materials, Manufacturing Engineering and Management, University of Nottingham, Nottingham, UK, from 2001 to 2002. His current research interests include robust control and its applications, process control, and intelligent control.

Dr. Wu is a member of the Chinese Association of Automation, and a senior member of the IEEE. He received the IFAC Control Engineering Practice Prize Paper Award in 1999 (together with M. Nakano and J. She).

Takao Terano is a professor of Department of Computer Science, School of Computing, Tokyo Institute of Technology. His research interests include genetic algorithm-based machine learning, case-based reasoning, analogical reasoning, distributed artificial intelligence, cooperative agents, computa- tional organization theory, and knowledge dystem development methodology. He is a member of the editorial board of major AI-related academic societies in Japan and a member of IEEE, ACM, AAAI, and PAAA.

Min Ding is a Lecturer of School of Automation, China University of Geosciences. She received her B.S. from the Central South University in 2009. She had the Ph.D. in Engineering from the Waseda University. Her research interests include renewable energy, control and operation optimization of micro-grid and fault diagnosis of microgrid.

Hua Xie is a postgraduate student of China University of Geosciences. He was born in Shanxi, China, 1992. He received his B.S. from China University of Geosciences in 2015. His research interests include wind power prediction, ormation fusion and data mining.

INTENTIONAL BLANK

SEGMENTATION AND CLASSIFICATION OF BRAIN TUMOR CT IMAGES USING SVM WITH WEIGHTED KERNEL WIDTH

Kimia Rezaei[1] and Hamed Agahi[2]

[1]Corresponding author: Department of Electrical Engineering, Fars science and research branch, Islamic Azad University, Iran
[2]Associate professor, Department of Electrical Engineering, Shiraz branch, Islamic Azad University, Fars, Iran

ABSTRACT

In this article a method is proposed for segmentation and classification of benign and malignant tumor slices in brain Computed Tomography (CT) images. In this study image noises are removed using median and wiener filter and brain tumors are segmented using Support Vector Machine (SVM). Then a two-level discrete wavelet decomposition of tumor image is performed and the approximation at the second level is obtained to replace the original image to be used for texture analysis. Here, 17 features are extracted that 6 of them are selected using Student's t-test. Dominant gray level run length and gray level co-occurrence texture features are used for SVM training. Malignant and benign tumors are classified using SVM with kernel width and Weighted kernel width (WSVM) and k-Nearest Neighbors (k-NN) classifier. Classification accuracy of classifiers are evaluated using 10 fold cross validation method. The segmentation results are also compared with the experienced radiologist ground truth. The experimental results show that the proposed WSVM classifier is able to achieve high classification accuracy effectiveness as measured by sensitivity and specificity.

KEYWORDS

Brain tumor, Computed tomography, Segmentation, Classification, Support vector machine.

1. INTRODUCTION

A brain tumor or intracranial neoplasm occurs when some abnormal cells are shaped inside the brain. Two main types of tumors exist: malignant or cancerous tumors and benign tumors. Medical image processing has been developed rapidly in recent years for detecting abnormal changes in body tissues and organs. X-ray computed tomography (CT) technology uses computer-processed X-rays to produce tomographic images of a scanned object, which makes inside the object visible without cutting. CT images are most commonly used for detection of head injuries, tumors, and Skull fracture. Since various structures have similar radiodensity, there is some difficulty separating them by adjusting volume rendering parameters. The manual analysis of tumor based on visual interpretation by radiologist may lead to wrong diagnosis when the number of images increases. To avoid the human error, an automatic system is needed for

David C. Wyld et al. (Eds) : ITCS, SIP, CST, ARIA, NLP - 2017
pp. 39– 50, 2017. © CS & IT-CSCP 2017 DOI : 10.5121/csit.2017.70304

analysis and classification of medical images. Image segmentation is the process of partitioning a digital image into a set of pixels based on their characteristics and in medical images, texture contents are considered as pixels characteristics. There are various methods for segmentation. Here Support Vector Machine (SVM) with kernel function is constructed to segment the tumor region by detecting tumor and non-tumor areas. The segmentation results are obtained for the purpose of classifying benign and malignant tumors. Classification is the problem of identifying to which of a set of categories a new observation belongs, on the basis of a training set of data whose category membership had been defined. There are various algorithms for classification using a feature vector containing image texture contents. SVM, which is considered as a supervised learning system for classification, is used here.

2. LITERATURE SURVEY

There are a lot of literatures that focus on brain tumor CT images segmentation, classification and feature extraction. Padma et al. [1] proposed co-occurrence, gray level and new edge features by means of SVM classifier for segmentation of tumor from brain CT images. The method is applied on real data of 80 tumor images and it is inferred that better accuracy had been achieved compared with the fuzzy c-means clustering method. Nandpuru et al. [2] proposes SVM classification technique to recognize normal and abnormal brain Magnetic Resonance Images (MRI). First, skull masking applied for the removal of non-brain tissue like fat, eyes and neck from images. Then gray scale, symmetrical and texture features were extracted for classification training process.Rajini et al. [3] proposes a new approach for automatic diagnosis of normal an abnormal MR images which is consist of two stages of feature extraction using discrete wavelet transformation and image classification by means of feed forward back propagation artificial neural network and k-nearest neighbors(k-NN) classifier. Sridhar et al. [4] proposed a method to classify 5 classes of Brain tumors MR images using Probabilistic Neural Network. Discrete Cosine Transform was applied for dimensionality reduction and extraction of 16 different features. Sundararaj et al. [5] used various intensity based and texture features such as skewness and coarseness. They constructed a linear SVM classifier for the diagnosis of normal and abnormal brain CT images. Padma et al. [6] used wavelet co-occurrence texture features by means of bidirectional associative memory type artificial neural network for the segmentation of tumor in brain CT images. They performed Genetic Algorithm for feature selection. The algorithm performance evaluation represents the outperformance of this method. Kaur, T et al.[7] proposed an automatic segmentation method on brain tumor MR images that performs multilevel image thresholding, using the spatial information encoded in the gray level co-occurrence matrix. Kaur, T et al.[8] proposed a technique which exploits intensity and edge magnitude information in brain MR image histogram and GLCM to compute the multiple thresholds. Verma, A. K. et al.[9] decomposed corrupted images using symlet wavelet then proposed a denoising algorithm utilizes the alexander fractional integral filter which works by the construction of fractional masks window computed using alexander polynomial.

The above literature survey illustrates that all the above methods are considered co-occurrence texture features only and some of the methods are proposed for the purpose of classification only and some for segmentation only.

3. MATERIALS AND METHODS

First, image noises are removed using median and wiener filter. Some features must be extracted from brain tumor images for the purpose of classifier training. Hence, a two-level discrete wavelet decomposition of tumor image is performed and the approximation at the second level is obtained to replace the original image to be used for texture analysis. Here, 17 features are extracted that 6 of them are selected using Student's t-test. Dominant gray level run length and gray level co-occurrence texture features are used for SVM training. Malignant and benign tumors are classified using SVM with kernel width and weighted kernel width (WSVM) and k-Nearest Neighbors (k-NN) classifier. The proposed methodology is applied to real brain CT images datasets collected from Shiraz Chamran hospital. All images are in DICOM format with a dimension of 512 × 512. The proposed algorithm is implemented in Matlab software.

3.1 Image Pre-processing and Enhancement

Medical images corrupt through imaging process due to different kinds of noise. In pre-processing stage, noise and high frequency artifact present in the images are removed. The median filter is a nonlinear digital filtering method, often used for noise reduction on an image or signal [10]. This technique is performed to improve the results of later processing. Median filter is mostly used to remove noise from medical images. Wiener filter produces an estimate of a target random process by means of linear time-invariant filter [11]. Wiener filter is also a helpful tool for the purpose of medical images noise reduction. Here images noise removing process is carried out by using median filter and wiener filter.

3.2 Tumor Segmentation using SVM Classifier

In this paper, SVM classifier is chosen for tumor identification[12]. SVM is a machine learning technique combining linear algorithms with linear or non-linear kernel functions that make it a powerful tool for medical image processing applications. To apply SVM into non-linear data distributions, the data should be transformed to a high dimensional feature space where a linear separation might become feasible. In this study, a linear function is used.

Training an SVM involves feeding studied data to the SVM along with previously studied decision values, thus constructing a finite training set. To form the SVM segmentation model, feature vectors of tumor and non-tumor area, distinguished with the help of radiologist, are extracted. 25 points covering tumor area and 25 points covering the non-tumor area are selected. These points not only cover all the tumor and no-tumor areas but also are enough as an input for training a SVM classifier due to its powerful learning even through using few numbers of training inputs. For each point (pixel), two properties of position and intensity are considered to form the feature vector or training vector. Totally 50 feature vectors are defined as input to the SVM classifier to segment the tumor shape. Accordingly, there is a 25 × 3 matrix of tumor area and a 25 × 3 matrix of non-tumor area. In segmentation phase, matrix t is given as input to the SVM for training and pixels are labeled so that their classes can be designated.

$$t_i = (x_i, y_i, I_i(x_i, y_i)) \quad i = 1, \ldots, 50 \tag{1}$$

i represent the number of training vectors. (x_i, y_i) and $I_i(x_i, y_i)$ represent the position and intensity of the selected points, respectively. Pixel selection using Matlab is displayed in Figure 1.

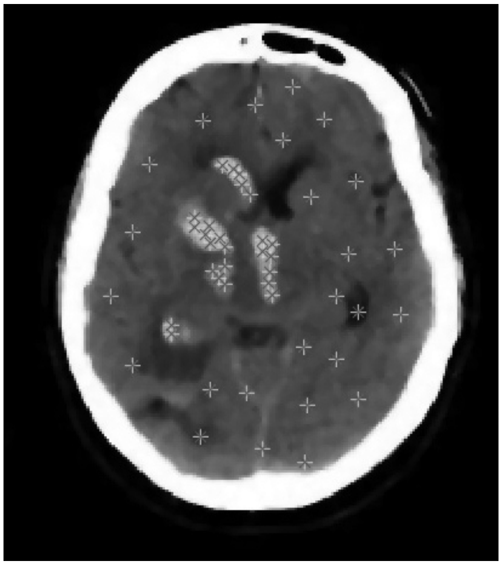


Figure 1. Pixel Selection using Matlab

3.3 Processing the Segmented Tumor Image on the Basis of 2D Discrete Wavelet Decomposition

Discrete Wavelet Decomposition is an effective mathematical tool for texture feature extraction from images. Wavelets are functions based on localization, which are scaled and shifted versions of some fixed primary wavelets. Providing localized frequency information about the function of a signal is the major advantage of wavelets.

Here a two-level discrete wavelet decomposition of tumor image is applied, which results in four sub-sets that show one approximation representing the low frequency contents image and three detailed images of horizontal, vertical and diagonal directions representing high frequency contents image [13]. 2D wavelet decomposition in second level is performed on the approximation image obtained from the first level. Second level approximation image is more homogeneous than original tumor image due to the removing of high-frequency detail information. This will consequence in a more significant texture features extraction process.

3.4 Feature Extraction

Texture is the term used to characterize the surface of an object or area. Texture analysis is a method that attempts to quantify and detect structural abnormalities in various types of tissues. Here dominant gray-level run length and gray-level co-occurrence matrix method is used for texture feature extraction.

The dominant gray-level run length matrix [14] is given as:

$$\varphi(d,\theta) = \left[p(i, j \mid d,\theta) \right] \qquad 0 < i \le N_g , \quad 0 < j \le R_{\max} \tag{2}$$

Where N_g is the maximum gray level and $R_{\max}$ is the maximum run length. The function $p\left(i,j\middle|\theta\right)$ calculates the estimated number of runs in an image containing a run length j for a gray level i in the direction of angle θ. Dominant gray-level run length matrices corresponding to θ = 0°, 45°, 90° and 135° are computed for approximation image derived from second level wavelet decomposition. Afterward, the average of all the features extracted from four dominant gray level run length matrices is taken.

A statistical method of analyzing texture considering the spatial relationship of pixels is the gray-level co-occurrence matrix (GLCM) [15]. The GLCM functions characterize the texture of the given image by computing how often pairs of pixel with certain values and in a specified spatial relationship occur in an image. The gray-level co-occurrence matrix is given as:

$$\varphi(d,\theta)=\left[p(i,j\mid d,\theta)\right] \quad 0<i\le N_g \quad , 0< j\le N_g \tag{3}$$

Where Ng is the maximum gray level. The element $p(i,j|d,\theta)$ is the probability matrix of two pixels, locating within an inter-sample distance d and direction θ that have a gray level i and gray level j. Four gray-level co-occurrence matrices, with θ = 0°, 45°, 90° and 135° for direction and 1and 2 for distance, are computed for approximation image obtained from second level wavelet decomposition. Then, 13 Haralick features [16] are extracted from each image's GLCM and the average of all the extracted features from four gray-level co-occurrence matrices is taken.

3.5 Feature Selection

Feature selection is a tool for transforming the existing input features into a new lower dimension feature space. In this procedure noises and redundant vectors are removed. Here, Two-sample Student's t-test is used for feature selection which considered each feature independently [17]. In this method, significant features are selected by computing the mean values for every feature in benign tumor class and malignant tumor class. Then, mean values of both classes are compared.

The T-test presumed that both classes of data are distributed normally and have identical variances. The test statistics can be calculated as follows:

$$t=x_b-x_m \,/\, \sqrt{\frac{\mathrm{var}_b}{n_b}+\frac{\mathrm{var}_m}{n_m}} \tag{4}$$

Where, x_b and x_m are mean values from benign and malignant classes. var_b and var_m represent variances of benign and malignant classes. n_b and n_m show the number of samples (images) in each class. This t value followed Student t-test with $\left(n_b+n_m-2\right)$ degrees of freedom for each class.

In statistics, the *p*-value is a function of the observed sample results, used to test a statistical hypothesis and figuring out that the hypothesis under consideration is true or false. Here, the p-value is calculated based on test statistics and degrees of freedom [18]. Then, the optimal features are selected on the basis of the condition $P < 0.001$.

3.6 Classification Using k-NN Classifier

In this paper, the main objective of classification is the identification of benign and malignant tumors in brain computed tomography images. The k-nearest neighbor classifier is a nonparametric supervised classifier that performs propitious for optimal values of k. k-NN algorithm consists of two stages of training and testing. In training stage, data points are given in n-dimensional space [19]. These training data are labeled so that their classes can be specified. In the testing stage, unlabeled data are given as input and the classifier generates the list of the k nearest data points (labeled) to the testing point. Then the class of the majority of that list is identified through the algorithm.

k-NN algorithm:

1. Define a suitable distance metric.

2. In training step, all the training data set P are put in pairs.

$$P = \{(y_i, C_i), i = 1, \ldots n\} \tag{5}$$

Where y_i is a training pattern in the training data set, C_i is its class and n is the number of training patterns.

3. In testing step, the distances between testing feature vector and training data are computed.

4. The k-nearest neighbors are chosen and the class of the testing example is specified.

The result of classification in testing stage is used to evaluate the precision of the algorithm. If it was not satisfactory, the k value can be changed till achieving the desirable result.

3.7 Classification using SVM classifier

Support vector machine algorithm depends on the structural risk minimization principle. Compared with artificial neural networks, SVM is less computationally complex and function well in high-dimensional spaces. SVM does not suffer from the small size of training dataset and obtains optimum outcome for practical problem since its decision surface is specified by the inner product of training data which enables the transformation of data to a high dimensional feature space. The feature space can be defined by kernel function K(x, y). The most popular kernel is the (Gaussian) radial basis function kernel, which is used here and is defined as follows:

$$k(x, y) = \exp(-\|x - y\|^2 / 2\sigma^2) \tag{6}$$

Where σ is the kernel width and chosen by the user. For the purpose of diminishing the coexisting over-fitting and under-fitting loss in support vector classification using Gaussian RBF kernel, the kernel width is needed to be adjusted, to some extent, the feature space distribution. The scaling rule is that in dense regions the width will be narrowed (through some weights less than 1) and in sparse regions the width will be expanded (through some weights more than 1) [20]. The Weighted Gaussian RBF kernel is as follows:

$$k(x,y)=\exp(-\lambda\_weight(x)\times\lambda\_weight(y)\times\lambda\times\|x-y\|^2) \tag{7}$$

Where λ_weight is a variable changing in a very small range around 1.

3.8 Sementation Performance Evaluation

The performance result of segmentation is evaluated by computation of segmentation accuracy. The segmentation accuracy is calculated as the direct ratio of the number of tumor pixels common for ground truth and the output of the segmented tumor to the total ground truth tumor pixels. The ground truth is indicated from the boundary drawings of the radiologist. segmentation accuracy and segmentation error are as fallows:

$$\begin{gathered}\text{Segmentation accuracy} = (\text{no. of pixels matche / total no. of tumor}\\ \text{pixels in ground truth})\times 100\end{gathered} \tag{8}$$

3.9 Classification Performance Evaluation

In order to compare the classification results, classifiers performances were evaluated using round robin (10-fold cross-validation) method [21]. In this method, the total number of data is divided into 10 subsets. In each step one subset is left out and the classifier is trained using the remainders. Next, the classifier is applied to the left out subset to validate the analysis. This process is iterated until each subset is left out once. For instance, in the n-sample images, the round robin method trains the classifier using n − 1 samples and then applies the one remaining sample as a test sample. Classification is iterated until all n samples have been applied once as a test sample. The classifier’s accuracy is evaluated on the basis of error rate. This error rate is defined by the terms true and false positive and true and false negative as follows:

$$sensetivity = TP\,/\,(TP+FN)\times 100 \tag{9}$$

$$specificity = TN\,/\,(TN+FP)\times 100 \tag{10}$$

$$accuracy = (TP+TN)\,/\,(TP+TN+FP+FN) \tag{11}$$

Where TN is the number of benign tumors truly identified as negative, TP is the number of malignant tumors truly identified as positive, FN, malignant tumors falsely identified as negative and FP, benign tumors falsely identified as positive. Sensitivity is the ability of the method to recognize malignant tumors. Specificity is the ability of the method to recognize benign tumors. Accuracy is the proportion of correctly identified tumors from the total number of tumors.

4. Results and Discussion

The proposed method is applied to real brain CT images. The data set consists of volume CT data of 20 patients (10-benign, 10- malignant). Number of slices varies across the patient's 4-6 benign slices and 4-6 malignant slices. In total there were 100 slices (50-benign, 50- malignant). All images were with a dimension of 512 × 512, gray scale and in DICOM format. The proposed algorithm is implemented in Matlab 2014b.

Result of an input real CT image and the segmented tumor image using SVM classifier is represented in Figure 2.

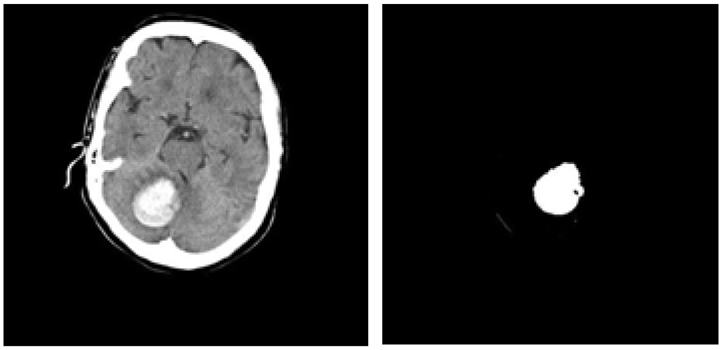


Figure 2. Input CT image and segmented tumor using SVM classifier

The quantitative results in terms of performance measures such as segmentation accuracy and segmentation error for real data of 50 benign slices of 10 patients (5 slices for each patient) and 50 malignant slices of 10 patients(5 slices for each patient), are calculated and tabulated in Table 1.

Table 1. Segmentaion Accuracy of 10 Patients with 100 Slices

Patients	Malignant slices (50)	Benign slices (50)
	Segmentation accuracy	
1	88.85	89.83
2	89.09	88.98
3	87.89	88.72
4	88.87	87.98
5	89.79	89.65
6	89.86	87.98
7	88.98	87.86

8	88.79	89.65
9	89.78	88.94
10	89.92	88.57
average accuracy	89.182	88.816

Tumor image wavelet approximation and its details in horizontal, vertical and diagonal directions at second level of wavelet decomposition can be observed in Figure 3.

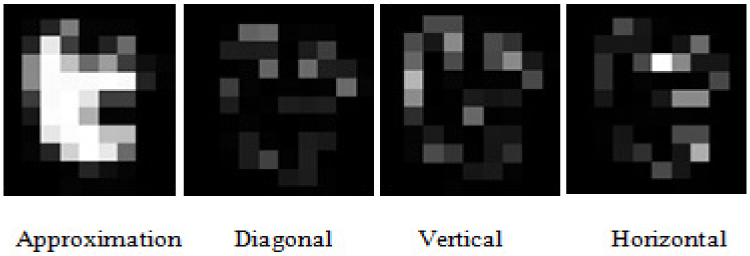


Figure 3. Wavelet Decomposition Images of Segmented Tumor

17 features are extracted from the wavelet approximation tumor image of each slice that 6 of them are selected by means of Student's t-test. The best textural features selected are long-run low-gray-level emphasis (LLGE), long-run high-gray-level emphasis (LHGE), energy, contrast, variance and inverse difference moment (IDM). The feature selection outcome is consistent with the knowledge of radiologist. For instance, feature variance computes the heterogeneity of a CT slice and LHGE captures the heterogeneous nature of the texture feature. According to radiologist, it can be inferred from the presence of heterogeneity that an abnormal slice is malignant. Feature IDM measures the homogeneity of a slice and feature LLGE demonstrate the homogeneous nature of the texture feature. Conforming to radiologist, the existence of homogeneity indicates that an abnormal slice is benign. These six features are given as inputs to the K-NN, SVM and WSM classifiers.

The performance of classifiers is evaluated using 10-fold cross-validation method and tabulated in Table 2.

Table 2. Classifier Performances Comparison

classifier	**K-NN**	**SVM**	**WSVM**
sensitivity	76%	77%	78%
specificity	72%	75%	76%
accuracy	74%	76%	77%

Classification accuracy effectiveness is measured by sensitivity and specificity. Compared to K-NN and SVM with the classification accuracy of 74% and 76% respectively, WSVM performed better with the accuracy of 77%. The stated results of the comparison of three classifier performances are represented using a bar graph as shown in Figure 4.

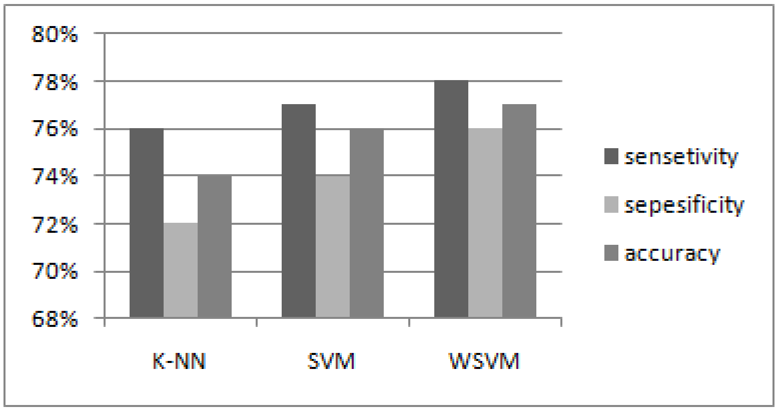


Figure 4. Classifiers Performance Evaluation

5. CONCLUSION

The work in this research involved using SVM with kernel function to classify Brain tumor CT images into benign and malignant. From the experimental results, it is inferred that the best classification performance is achieved using the WSVM. Furthermore, these results show that the proposed method is effective and efficient in predicting malignant and benign tumors from brain CT images. For future work, the proposed method can be applied to other types of imaging such as MRI and even can be used for segmentation and classification of tumors in other parts of body.

ACKNOWLEDGEMENT

The authors wish to acknowledge Shiraz Chamran hospital for providing the original brain tumor images in DICOM format.

REFERENCES

[1] Padma Nanthagopal, A., Sukanesh Rajamony, R., 2012, A region-based segmentation of tumour from brain CT images using Nonlinear Support Vector Machine classifier, J. Med. Eng. Technol., 36, (5), 271–277

[2] Nandpuru, H.B., Salankar, S.S., Bora, V.R., 2014, MRI Brain Cancer Classification Using Support Vector Machine, IEEE. Conf. Electrical Electronics and Computer Science, 1–6

[3] Hema Rajini N., Bhavani R., 2011, Classification of MRI Brain Images using k-Nearest Neighbor and Artificial Neural Network., IEEE-International Conference on Recent Trends in Information Technology, ICRTIT Chennai, Tamil Nadu., 563 – 568

[4] Sridharn. D., Murali Krishna.IV., 2013, Brain Tumor Classification U sing Discrete Cosine Transform and Probabilistic Neural Network, 2013 International Conference on Signal Processing Image Processing & Pattern Recognition (ICSIPR), Coimbatore, 92 – 96

[5] Sundararaj, G.K.; Balamurugan, V., 2014, Robust Classification of Primary Brain Tumor in Computer Tomography Images Using K-NN and Linear SVM, 2014 International Conference on Contemporary Computing and Informatics (IC3I), Mysore, 1315 – 1319

[6] Padma, A., Sukanesh, R., 2011, A wavelet based automatic segmentation of brain tumor in CT images using optimal statistical texture features, Int. J. Image Process., 5, (5), 552–563

[7] Kaur, T., Saini, B. S., & Gupta, S. (2016). Optimized Multi Threshold Brain Tumor Image Segmentation Using Two Dimensional Minimum Cross Entropy Based on Co-occurrence Matrix. In Medical Imaging in Clinical Applications (pp. 461-486). Springer International Publishing.

[8] Kaur, T., Saini, B. S., & Gupta, S. (2016) A joint intensity and edge magnitude-based multilevel thresholding algorithm for the automatic segmentation of pathological MR brain images. Neural Computing and Applications, 1-24.

[9] Verma, A. K., & Saini, B. S. ALEXANDER FRACTIONAL INTEGRAL FILTERING OF WAVELET COEFFICIENTS FOR IMAGE DENOISING. Signal & Image Processing : An International Journal (SIPIJ) Vol.6, No.3, June 2015

[10] Chun-yu ,N., 2009, Research on removing noise in medical image based on median filter method, IT in Medicine & Education, ITIME '09. IEEE International Symposium, Jinan, 384 – 388

[11] Benesty, J.; Jingdong Chen; Huang, Y.A., 2010, Study of the widely linear Wiener filter for noise reduction, Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference, Dallas, TX, 205- 208

[12] El-Naqa, I., Yang, Y., Wernick, M.N., Galatsanos, N.P., Nishikawa, R.M, 2002, A support vector machine approach for detection of microcalcifications, IEEE Trans. Med. Imag., 21, (12), 1552–1563

[13] Zhengliang Huan; Yingkun Hou, 2008, An Segmentation Algorithm of Texture Image Based on DWT, Natural Computation, 2008. ICNC '08. Fourth International Conference, Jinan, 5, 433- 436

[14] Tang, X., 1998, Texture information in run length matrices, IEEE Trans. Image Process., 7, (11), 234–243

[15] Khuzi, M., Besar, R., Zaki WMD, W., Ahmad, N.N., 2009, Identification of masses in digital mammogram using gray level co-occurrence matrices, Biomed. Imag. Interv. J., 5, (3), 109–119

[16] Haralick, R.M., Shanmugam, K., Dinstein, I., 1973, Texture features for Image classification, IEEE Trans. Syst. Man Cybern. 3, (6), 610–621

[17] Levner, I., Bulitko, V., Lin, G., 2006, Feature extraction for classification of proteomic mass spectra: a comparative study, Springer-Verlag Berlin Heidelberg, Stud Fuzz, 207, 607–624

[18] Soper D.S.: ‘P-value calculator for a student t-test (OnlineSoftware)’, 2011, http://www.danielsoper.com/statcalc3

[19] F. Latifoglu, K. Polat, S. Kara, S. Gunes, 2008, Medical diagnosis of atherosclerosis from carotid artery Doppler signals using principal component analysis (PCA), k-NN based weighting pre-processing and Artificial Immune Recognition System (AIRS), J. Biomed. Inform. 41, 15–23.

[20] Yuvaraj N., Vivekanandan P., 2013, An Efficient SVM based Tumor Classification with Symmetry Non-Negative Matrix Factorization Using Gene Expression Data , Information Communication and Embedded Systems (ICICES), 2013 International Conference, Chennai, 761– 768

[21] Liao Y.-Y., Tsui, P.-H., Yeh, C.-K., 2009, Classification of benign and malignant breast tumors by ultrasound B-scan and nakagami-based images, J. Med. Biol. Eng. 30, (5), 307–312

AUTHORS

Dr. Hamed Agahi, has obtained his doctoral degree from Tehran University, Iran. He has 4 years of teaching experience. He is currently working as an Assistant Professor and the head of researchers and elite club in Shiraz Azad University, Iran. He has published many papers in scientific journals and conference proceedings. His research interests include pattern recognition, image processing, signal processing and machine vision and applications.

Kimia Rezaei received her Bachelor degree from Fasa Azad University, Iran, and the Master degree in telecommunications engineering from Shiraz Azad University, Iran. She has published one paper in national conference in Iran. She is currently working as Telecommunicatons Engineer in Sahand Telecommunication company in Iran. Her research interest is focused on pattern recognition and Image processing related research programs targeted for practical applications.

INFORMATION SECURITY MATURITY MODEL FOR NIST CYBER SECURITY FRAMEWORK

Sultan Almuhammadi and Majeed Alsaleh

College of Computer Sciences and Engineering,
King Fahd University of Petroleum and Minerals,
Dhahran, Saudi Arabia

ABSTRACT

The National Institute of Standards and Technology (NIST) has issued a framework to provide guidance for organizations within critical infrastructure sectors to reduce the risk associated with cyber security. The framework is called NIST Cyber Security Framework for Critical Infrastructure (CSF). Many organizations are currently implementing or aligned to different information security frameworks. The implementation of NIST CSF needs to be aligned with and complement the existing frameworks. NIST states that the NIST CSF is not a maturity framework. Therefore, there is a need to adopt an existing maturity model or create one to have a common way to measure the CSF implementation progress. This paper explores the applicability of number of maturity models to be used as a measure to the security poster of organizations implementing the NIST CSF. This paper reviews the NIST CSF and compares it to other information security related frameworks such as COBIT, ISO/IEC 27001 and the ISF Standard of Good Practice (SoGP) for Information Security. We propose a new information security maturity model (ISMM) that fills the gap in the NIST CSF.

KEYWORDS

Information Security, Maturity Model, Cyber-Security.

1. INTRODUCTION

Many organizations could be aligned with one of the information security related best practice frameworks. This makes the alignment of the NIST CSF with such frameworks a must. NIST CSF is a set of industry standards and best practices [1]. The framework of NIST CSF clearly indicates that organizations planning to implement it can use their existing processes and place them on top of the NIST CSF to identify gaps with respect to the framework. This implies the comprehensiveness of the NIST CSF when compared with other frameworks such as COBIT, ISO/IEC 27001 and ISF Standard of Good Practices (SoGP). Thus, to ensure applicable alignment with any information security framework, we need to confirm the comprehensiveness or identify any possible gap in NIST CSF.

David C. Wyld et al. (Eds) : ITCS, SIP, CST, ARIA, NLP - 2017
pp. 51– 62, 2017. © CS & IT-CSCP 2017 DOI : 10.5121/csit.2017.70305

However, in this paper, we show that NIST CSF is not comprehensive to address all information security related processes that are addressed in some of those frameworks. The main objective of the framework is to manage cyber security risks within the organizations that implement it. In the NIST CSF, the \Framework Implementation Tiers" part, referred to as \Tiers", is detailed as one of three parts that the framework consists of [1]. However, the Tiers does not provide organizations with a mechanism to measure the progress of implementing NIST CSF or their maturity level and information security processes' capabilities. Tiers is just visionary tool that allows organizations to understand their cyber security risk management approach and what are the processes in place to manage the risk. NIST official web site [2] has stated that the Tiers are not intended to be measurement tool to maturity levels.

This paper is a comprehensive comparison between NIST CSF, COBIT, ISO/IEC 27001 and ISF frameworks. It identifies the gap of key information security processes that are addressed in some frameworks but not in NIST CSF. We fill this gap and propose a new capability maturity model (CMM) to measure NIST CSF implementation progress.

2. Overview of The NIST Cyber Security Framework

The NIST CSF consists of three main parts in which, cyber security is considered as a risk that is managed through the enterprise risk management process [1]. Thus, we identify the NIST CSF framework as risk-based framework. The three parts are: framework core, risk tiers, and framework profile.

Table 1: Frameworks Comparison

Framework	Control Categories	Control Objectives	Activities
NIST CSF [1]	Functions (5)	Categories (22)	Subcategories (98)
ISF [3]	Categories (4)	Areas (26)	Topics (118)
ISO27001 (2013) [4]	Clauses (14)	Control objective (35)	Controls (114)
COBIT5 (2013) [5]	Domains (5)	Processes (37)	Practices (210)

2.1 Framework Core

The framework core consists of a set of cyber security activities. These activities are grouped in \Subcategories" which are grouped too in \Categories". The categories are sorted in five different \Functions": Identify, Protect, Detect, Respond, and Recover. The NIST CSF five functions are concurrent and continuous. When the functions collectively implemented they form a high-level and strategic view of the cyber security risk management program. The Framework Core part has also the desired outcomes (controls objectives) and informative references. Informative references are list of cyber security activities in standards, guidelines, or practices such as as COBIT, ISO/IEC 27001 and the ISF SoGP. The comparison between the NIST CSF and other frameworks will be done on the level of the cyber security activities to ensure that all key information securities activities are addressed. Table 1 compares the structure of NIST CSF with the structure of selected sample of frameworks.

2.2 RISK TIERS

The Tiers part of the NIST CSF is a visionary tool that allows organizations to understand their cyber security risk management approach and what are the processes in place to manage the risk. Based on the identified processes in place, the organization may be classified in one of four tier levels. The tier levels range from \Partial" in Tier 1,\Risk Informed" in Tier 2, \Repeatable" in Tier 3, to \Adaptive" in Tier 4.

2.3 FRAMEWORK PROFILE

The framework profile, referred to as \profile", is a tool to document, implement, and track the organizations' opportunities for improving their cyber security posture. The profile has the current cyber security activities implemented by the organization, as well as the planned activities to be implemented in order to close the gap between the current and the \to-be" state. Organizations need to identify which cyber security activities are needed to improve the current state based on risk assessment to identify risks that may prevent achieving the business objectives.

3. RELATED WORK

We reviewed the \Baldrige Excellence Framework" and \Baldrige Excellence Builder" at NIST website [6]. We found that these two documents were not introduced to serve as Maturity Model. However, they are a continues effort linked to the Tiers, where the main aim is to help organizations to evaluate how effective is their cyber security risk management effort. The Baldrige Excellence Builder links the cyber security program with several areas such as leadership, customers, employees, and the outcome results. In [7], the authors proposed a method to select measures which evaluate the gap between the current and the target states based on the NIST CSF risk Tiers. In [8], on the other hand, the authors highlighted the need for Compliance Assessment in order to reduce the gap in the Processes pillar (one of three pillars including Human Resources and Technology). Therefore, they proposed a model that is generic to allow for overall compliance evaluation.

4. NIST CSF EVALUATION

NIST CSF, as a framework, has the following nature:

- Focus on information security high-level requirements.
- Applicable for the development of information security program and policy

Examples of other frameworks include, COBIT, ISO/IEC 27001 and the ISF SoGP for Information Security. However, the detailed cyber security activities are usually listed in standards, guidelines, and practices. They have the following nature:

- Focus on information security technical and functional controls (customizable).
- Applicable for developing checklists and conducting compliance/audit assessments.

Examples of standards and guidelines include NIST SP 800-53, ISO-27001 Annex, and ISF SoGP. The NIST CSF has mapped number of standards in the informative references. The mapped standards include NIST SP 800 series, COBIT 5, ISA 62443, ISO/IEC 27001:2013, and

CCS [1]. ISF SoGP was not mapped in the NIST CSF framework. Therefore, we will use the ISF SoGP mapping [9] to NIST CSF to conduct the comparison with NIST CSF.

NIST CSF clearly indicates that organizations planning to implement it can use their existing processes and place them on top of the NIST CSF to identify gaps with respect to the framework [1]. However, this assumes that NIST CSF will be comprehensive and adopted framework will be always equal or less than NIST. This is illustrated in Figure 1-a.

Of course the other scenario of NIST CSF being not comprehensive and has a gap when compared with other frameworks is possible. In order to verify this scenario (illustrated in Figure 1-b), we matched all mapped CSF informative references with the corresponding framework. Numeric statistics of this match are as follows:

Framework	CSF	Gap	Gap %
ISO 27001	93	21	18.4%
ISF	49	69	58.5%
COBIT5	165	45	21.4%

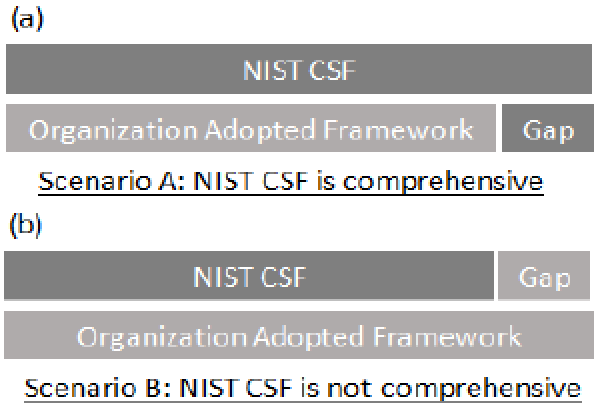

(a)
NIST CSF
Organization Adopted Framework
Gap
Scenario A: NIST CSF is comprehensive
(b)
NIST CSF
Gap
Organization Adopted Framework
Scenario B: NIST CSF is not comprehensive

Figure 1: Two gap scenarios for CSF being comprehensive

We found that the compliance process is one gap area, related to information security, that is identified and need to be addressed in future update to NIST CSF. For example, MEA03 (Monitor, Evaluate and Assess Compliance with External Requirements) is a COBIT process that is not mapped to NIST CSF. Also, SI2.3 (Monitoring Information Security Compliance) is ISF process that is not mapped to NIST CSF. In addition, ISO/IEC 27001 has one process (A.18: Compliance) that is partially mapped to NIST CSF. NIST CSF has mapped only the following five ISO/IEC processes: A.18.1 (Compliance with legal and contractual requirements), A.18.1.3 (Protection of records), A.18.1.4 (Privacy and protection of personally identifiable information), A.18.2.2 (Compliance with security policies and standards), and A.18.2.3 (Technical compliance review).

We traced the Compliance Assessment in NIST 800 series and found two main publications ([10] and [11]) that highlighted this topic. The Compliance Assessment was addressed under the Risk Monitoring process, roles and responsibilities associated with it. The two main objectives of the Risk Monitoring process are to verify the existence of the control (Compliance) and the efficiency of the control to mitigate the risk [11]. Compliance assessment is very essential to ensure that identified control to mitigate the risk is implemented correctly and operating as intended. For detailed responsibilities of each role in the compliance process refer to the following in [10]:

Role	Reference [10]
Info. system owner	Sec. D.9, Page D-5
Info. system security officer	Sec. D.10, Page D-6
info. security architect	sec. D.11, Page D-6
security control assessor	sec. D.13, Page D-7

We propose to add the compliance assessment process as a process in NIST CSF (be the category number 23). This category will contain the missed subcategories highlighted previously. The process should at least contain the following as subcategories:

- Legal and Regulatory Compliance
- Information Privacy
- Intellectual property
- Compliance with security policies and standards

5. Measuring Maturity of Organizations Implementing NIST CSF

The profile part of the NIST CSF is focused on tracking the organization progress in implementing the gaps to move from the current state to the defined target. NIST CSF has provided the Tiers as visionary tool that allows organizations to understand their cyber security risk characteristics. However, as we highlighted in Section 1, Tiers does not provide organizations with a mechanism to measure the progress of implementing NIST CSF or their maturity level and information security processes capabilities.

Therefore, a maturity model is needed to measure the information security processes capabilities. The main objective of such maturity model is to identify a baseline to start improving the security posture of an organization when implementing NIST CSF. The maturity model then is used in cycles to build consensus, set the priorities of investment in information security, and after all measure the implementation progress [12]. Some of the frameworks that we studied come with maturity model (like COBIT and ISF). For other frameworks that do not have maturity model like ISO 27001, other information security related maturity models like ONG C2M2 and SSE MM are used (Figures 2 and 3). We studied the different maturity models to verify if they map to each other in order to utilize any of them to measure the maturity of organizations implementing NIST CSF. The main focus of our study was to compare the scale used by each model and the domains evaluated by each model.

We reviewed the four maturity models SSE CMM [13], ONG C2M2 [14], ISF MM [15], and COBIT PAM MM [16]. Unlike the other three maturity models, ONG C2M2 is three scale model and assesses ten domains. Refer to Figure 2.

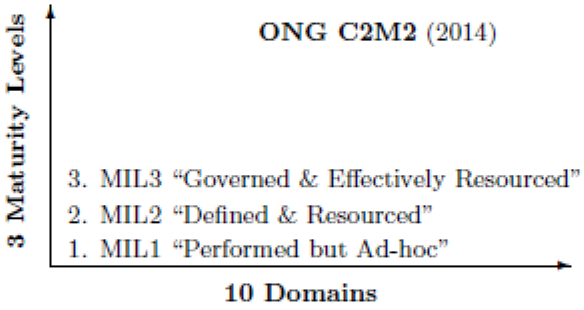

ONG C2M2 (2014)
3 Maturity Levels
3. MIL3 "Governed & Effectively Resourced"
2. MIL2 "Defined & Resourced"
1. MIL1 "Performed but Ad-hoc"
10 Domains

Figure 2: ONG C2M2 Maturity Model scales and domains

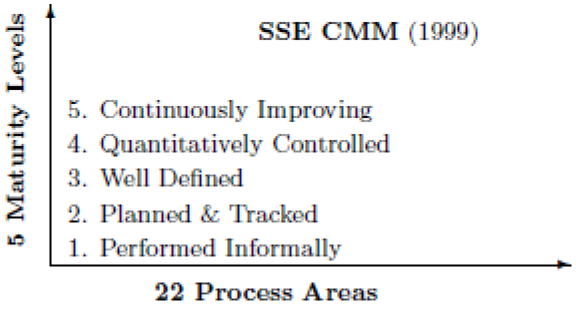


Figure 3: SSE Maturity Model scales and domains

While SSE CMM (Figure 3), ISF MM (Figure 4) and PAM MM (Figure 5) are the same scale maturity models, yet the problem of mapping exists. In Table 2, we identified that level 2 \Planned and Tracked" of SSE CMM is not mapped to any of the other maturity models. Figure 3 illustrates the levels and domains of SSE CMM. On the other hand, in ISF MM and PAM MM, level 2 and 3 is the opesite of each other. Figures 4 and 5 illustrate the levels and domains of ISF MM and PAM MM.

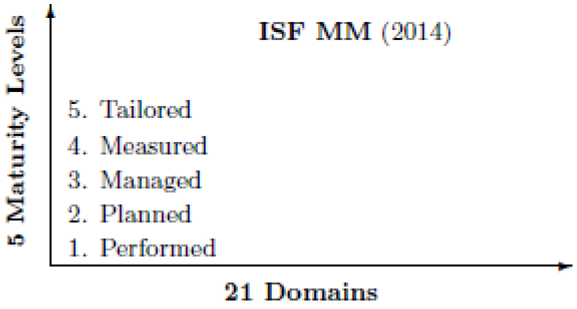


Figure 4: ISF Maturity Model scales and domains

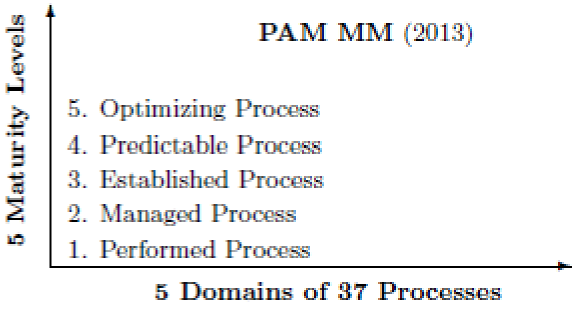


Figure 5: PAM Maturity Model scales and domains

We performed a comprehensive comparison of all the domains in the four maturity models. These domains are carefully examined in an attempt to verify the applicability of any maturity model regardless of the deployed framework. However, our study shows the lack of one-to-one mapping or any clear way to map these domains in an applicable way. There are items in certain models that are mapped to multiple items in other models. While other items have no mapping as show in Table 3. For example, \Monitor Posture" in SSE CMM is mapped to three items in ISF MM, three items in PAM, and two items in ONG C2M2. While \Monitor and Control Technical Effort" in SSE CMM and \Manage Operation" in PAM have no mapping in ISF MM and ONG C2M2. Moreover, the \Administer Security Controls" in SSE MM is mapped to seven items in PAM MM.

Table 2: Maturity Models Scale Comparison

SSE CMM [13]	ISF MM [15]	COBIT PAM [16]	ONG C2M2 [14]
L1 Performed Informally	L1 Performed	L1 Performed Process	L1 Performed but Ad-hoc
L2 Planned and Tracked	No Mapping	No Mapping	No Mapping
L3 Well Defined	L2 Planned	L3 Established Process	L2 Defined and Resourced
No Mapping	L3 Managed	L2 Managed Process	L3 Governed and Effectively Resourced
L4 Quantitatively Controlled	L4 Measured	L4 Predictable Process	No Mapping
L5 Tailored	L5 Continuously Improving	L5 Optimizing Process	No Mapping

Our study, as summarized in Tables 2 and 3, illustrates the mapping of ONG C2M2 with the other three maturity models in Table 2. It shows that ONG C2M2 has similarity and map to the first three levels of the ISF MM (Figure 4). However, there is a gap in the assessed areas due to the difference in the number of both maturity models (10 assessed areas in ONG C2M2 versus 21 in ISF MM). There are assessed areas in ISF MM which are not mapped to ONG C2M2 such as \Compliance", \Security Audit", \Security Architecture", and \Secure Application Development". Other areas of ONG C2M2 are mapped to more than one area in ISF MM. For example, \Cyber security Program Management" in ONG C2M2 was mapped to three areas in ISF MM, namely,\Security Strategy", \Security Governance", and \Security Policy".

Table 3: Maturity Models Domains Comparison

SSE CMM [13]	**ISF MM [15]**	**COBIT PAM [16]**	**ONG C2M2 [14]**
No Mapping	Security Strategy	Manage Strategy	Cybersecurity Program Management
Administer Security Controls*	Security Governance	Ensure Governance Framework Setting and Maintenance	
		Ensure Benefits Delivery	
		Ensure Risk Optimization	
		Ensure Resource Optimization	
		Ensure Stakeholder Transparency	
Specify Security Needs	Security Policy	Manage the IT Management Framework	
Assess Security Risks	Information Risk Management	Manage Risk	Risk Management
Manage Project Risk			
Assess Threats			
Assess Impact			
Verify and Validate Security	Compliance	Monitor, Evaluate and Assess Performance and Conformance	No Mapping
Build Assurance Augment		Monitor, Evaluate and Assess the System of Internal Control	
		Monitor, Evaluate and Assess Compliance . . .	
	Security Audit		No Mapping
Administer Security Controls*	Asset Management	Manage Assets	Asset, Change, and Configuration Management
	Change Management	Manage Configuration	
Manage Configuration		Manage Change	
		Manage Organizational Change Ennoblement	
		Manage Change Acceptance and Transitioning	
No Mapping	Identity and Access Management	Manage Security Services	Identity and Access Management
Assess Vulnerabilities	Vulnerability Management		Threat and Vulnerability Management
Monitor Posture	Threat Intelligence		
	Security Event Management	Manage Service Requests and Incidents	Event and Incident Response, Continuity of Operations
	Incident Management	Manage Problems	
No Mapping	Business Continuity	Manage Continuity	
No Mapping	Crisis Management		
Coordinate Security	No Mapping	Manage Security	No Mapping
Plan Technical Effort	No Mapping	Manage Programs and Projects	No Mapping
		Manage Portfolio	
		Manage Budget and Costs	
Monitor and Control Technical Effort	No Mapping	Manage Operations	No Mapping
Provide Security Input	Security Architecture	Manage Enterprise Architecture	No Mapping
Define Organization Systems Process		Manage Requirements Definition	
Improve Organization Systems Engineering Process		Manage Solutions Identification and Build	
Manage Systems Engineering Support Environment		Manage Availability and Capacity	
Manage Product Line Evolution		Manage Innovation	
Ensure Quality	Secure Application Development	Manage Quality	No Mapping
No Mapping	Digital Connections	Manage Assets*	Information Sharing and Communications

Provide On-Going Skills	Human Resources Security	Manage Knowledge Manage Human Resources	Workforce Management
Administer Security Controls*	Security Awareness and Behavior	Manage Security*	Situational Awareness
Coordinate with Suppliers	External Supplier Management	Manage Relationships Manage Service Agreements Manage Suppliers	Supply Chain and External Dependencies Management
No Mapping	No Mapping	Manage Business Process Controls	No Mapping

* This item is mapped to multiple items in other models

6. PROPOSED INFORMATION SECURITY MATURITY MODEL

In the previous sections, we discussed few critical issues about NIST CSF framework. In order to understand the importance of a new cabability maturity modle for the NIST CSF, we highlight the following factors:

- The need for business management to measure the maturity of the security program to assure the reliability of the IT services enabling and supporting their business [12].
- NIST CSF framework tiers are not intended to be measurement tool to maturity levels [2].
- The identified gap in NIST CSF.
- The lack of (one-to-one) mapping in both scale levels and the assessed areas of the different existing maturity modules.

Taking into considerations all the above factors, there is a need to define a new CMM for NIST CSF. Therefore, we propose a five-level scale with 23 assessed areas as shown in Figure 6. Our suggested assessed areas are shown in Table 4. These areas are the 22 in NIST CSF categories plus the compliance assessment (No. 6 in Table 4).

Three of the four maturity models we compared are five-level scales in addition of other five information security related maturity models reviewed by [12]. This supports our decision to make the proposed maturity model a five-level scale. However, detailed review of the required scale characteristics, such as the scale levels, scale level definitions, or scale measures (staged versus continuous), need to be addressed in future work.

The proposed ISMM will enable the organizations to measure their implementation progress over time. They will use the same measuring tool in a regular basis to ensure maintaining the desired security posture. Furthermore, using the same measuring tool by different organizations will allow the benchmarking between those organizations [12].

Table 4: The 23 assessed areas of the proposed maturity model

1	Asset Management
2	Business Environment
3	Governance
4	Risk Assessment
5	Risk Management Strategy
6	Compliance Assessment
7	Access Control
8	Awareness and Training
9	Data Security
10	Information Protection Processes and Procedures
11	Maintenance
12	Protective Technology
13	Anomalies and Events
14	Security Continuous Monitoring
15	Detection Processes
16	Response Planning
17	Response Communications
18	Response Analysis
19	Response Mitigation
20	Response Improvements
21	Recovery Planning
22	Recovery Improvements
23	Recovery Communications

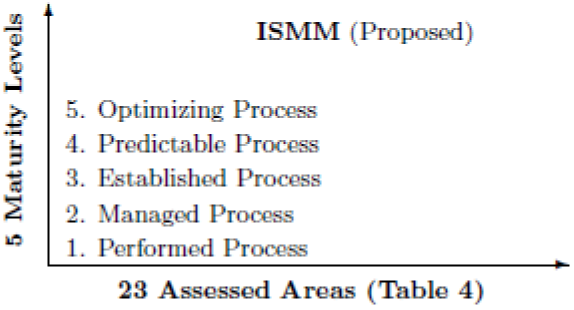


Figure 6: Proposed Information Security Maturity Model (ISMM)

7. CONCLUSION

NIST CSF has been introduced to organizations with critical infrastructure as an integrated framework to implement in order to improve their security postures. The NIST recommended to use the framework on top of and to complement any implemented framework within the organization. The ongoing enhancement nature of the information security programs drives the organizations to continuously measure their capabilities of achieving the desired outcome of the implemented framework. The organizations use the capability maturity models to evaluate their capabilities. This will give the management of the organization the bases of their decisions to define and prioritize their investment strategy in building the information security.

This paper considered the evaluation of the NIST CSF comprehensiveness to ensure that it will cover any existing framework. Moreover, the paper reviewed number of maturity models to assess the applicability to use with NIST CSF and the existence of mapping between the NIST CSF control objectives and the assessed areas. The paper used three information security related frameworks (ISO 27001, ISF, and COBIT5) and four maturity models (ISF, PAM, SSE CMM, and ONG C2M2).

The review considered the mapping made by NIST CSF to other frameworks and confirmed that the NIST CSF did not adequately address the compliance assessment process. The evaluation of the maturity models considered the scale levels definitions and the assessed areas. In both dimensions, there was no one-to-one mapping between the different maturity models. Therefore, we concluded that none of the evaluated maturity models can be used with NIST CSF to have a wide coverage and mapping to implemented framework. The paper proposed a new maturity model of five-level scale and include the twenty two NISCT CSF categories with addition of the compliance assessment process.

As for the future work, first, this paper shows the comparison between the assessed areas in different maturity models, but it did not compare them with the NIST CSF. This comparison is important to identify which maturity model can be used as a bases to define the scale levels of the proposed NIST CSF maturity model. The scope of the comparison also needs to be expanded to cover more cyber security and information security related maturity models such as the Community Cyber Security Maturity Model [17] and the Information Security Governance model [12]. Second, the current best practice of the information security business structure needs to be considered to resort the 23 assessed areas according to that structure grouping areas performed in one entity together. For example, business processes like the asset management, change management, threat monitoring, or risk management might be used to group related NIST CSF categories.

REFERENCES

[1] NIST, \Framework for improving critical infrastructure cybersecurity," http://www.nist.gov/cyberframework/upload/cybersecurity-framework-021214.pdf, 2014.

[2] N. Keller, \Cybersecurity framework faqs framework components," https://www.nist.gov/cyberframework/cybersecurity-framework-faqs-frameworkcomponents, 2015, accessed: December 11, 2016.

[3] ISF, \The standard of good practices for information security," in Information Security Forum ISF, 2014.

[4] S. Schweizerische, \Information technology-security techniques-information security management systems-requirements," ISO/IEC International Standards Organization, 2013.

[5] ISACA, \Cobit 5: A business framework for the governance and management of enterprise it," 2012.

[6] L. Scott, \Baldrige cybersecurity initiative," https://www.nist.gov/baldrige/productsservices/baldrige-cybersecurity-initiative, 2016, accessed: November 10, 2016.

[7] S. Fukushima and R. Sasaki, \Application and evaluation of method for establishing consensus on measures based on cybersecurity framework," in The Third International Conference on Digital Security and Forensics (DigitalSec2016), 2016, p. 27.

[8] N. Teodoro, L. Goncalves, and C. Serr~ao, \Nist cybersecurity framework compliance: A generic model for dynamic assessment and predictive requirements," in Trustcom/BigDataSE/ISPA, 2015 IEEE, vol. 1. IEEE, 2015, pp. 418{425.

[9] ISF, \Isf standard and nist framework poster," in Information Security Forum ISF, 2014.

[10] J. T. FORCE and T. INITIATIVE, \Guide for applying the risk management framework to federal information systems," NIST special publication, vol. 800, p. 37, 2010.

[11] P. D. Gallagher and G. Locke, \Managing information security risk organization, mission, and information system view," National Institute of Standards and Technology, 2011.

[12] M. Lessing, \Best practices show the way to information security maturity," http://hdl.handle.net/10204/3156, 2008, accessed: January 10, 2017.

[13] Carnegie-Mellon-University, \Systems security engineering capability maturity model (sse-cmm) model description document version 3.0," 1999.

[14] D. of Energy, \Oil and natural gas subsector cybersecurity capability maturity model (ong-c2m2 v1.1)," Department of Energy, Washington, DC: US, 2014.

[15] ISF, \Time to grow using maturity models to create and protect value," in Information Security Forum ISF, 2014.

[16] ISACA, COBIT Process Assessment Model (PAM): Using COBIT 5. ISACA, 2013.

[17] G. B. White, \The community cyber security maturity model," in Technologies for Homeland Security (HST), 2011 IEEE International Conference on. IEEE, 2011, pp. 173{178.

A SURVEY ON RECENT APPROACHES COMBINING CRYPTOGRAPHY AND STEGANOGRAPHY

Sultan Almuhammadi and Ahmed Al-Shaaby

College of Computer Sciences and Engineering,
King Fahd University of Petroleum and Minerals,
Dhahran, Saudi Arabia

ABSTRACT

Digital communication witnesses a noticeable and continuous development in many applications in the Internet. Hence, a secure communication sessions must be provided. The security of data transmitted across a global network has turned into a key factor on the network performance measures. Cryptography and steganography are two important techniques that are used to provide network security. In this paper, we conduct a comparative study of steganography and cryptography. We survey a number of methods combining cryptography and steganography techniques in one system. Moreover, we present a classification of these methods, and compare them in terms of the algorithm used for encryption, the steganography technique and the file type used for covering the information.

KEYWORDS

Cryptography, encryption, decryption, steganography, stego-image.

1. INTRODUCTION

Information security has grown as a significant issue in our digital life. The development of new transmission technologies forces a specific strategy of security mechanisms especially in state of the data communication [1]. The significance of network security is increased day by day as the size of data being transferred across the Internet [2]. Cryptography and steganography provide significant techniques for information security [3].

The most important motive for the attacker to benefit from intrusion is the value of the confidential data he or she can obtain by attacking the system [2]. Hackers may expose the data, alter it, distort it, or employ it for more difficult attacks [4]. A solution for this issue is using the advantage of cryptography and steganography combined in one system [5, 3].

This paper presents a historical background of the art of cryptography and steganography in Section 2, and shows the differences between these techniques. Section 3 gives a literature survey about methods which combine steganography techniques and cryptography techniques. Section 4

David C. Wyld et al. (Eds) : ITCS, SIP, CST, ARIA, NLP - 2017
pp. 63– 74, 2017. © CS & IT-CSCP 2017 DOI : 10.5121/csit.2017.70306

presents a comparative analysis of the surveyed methods. The conclusion is in Section 5 with some useful remarks.

2. BACKGROUND

Cryptography and steganography are two approaches used to secure information, either by encoding the information with a key or by hiding it [1, 6, 7, 8]. Combining these two approaches in one system gives more security [5, 9]. It is useful to explain these approaches and discuss the benefits of combining them.

2.1 CRYPTOGRAPHY

Cryptography is one of the traditional methods used to guarantee the privacy of communication between parties. This method is the art of secret writing, which is used to encrypt the plaintext with a key into ciphertext to be transferred between parties on an insecure channel. Using a valid key, the ciphertext can be decrypted to the original plaintext. Without the knowledge of the key, nobody can retrieve the plaintext. Cryptography plays an essential role in many services, like: confidentiality, key exchange, authentication and non-repudiation. Cryptography provides these services for secure communication across insecure channels, Figure 1 shows the cryptography system [10].

There are three types of cryptographic schemes for securing the data: public-key cryptography, secret key cryptography, and hash functions. These schemes are used to achieve different objectives. The length and type of the keys used depend on the type of encryption algorithm [10].

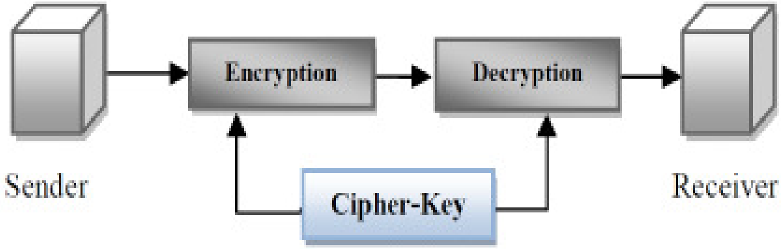


Figure 1: Cryptography System [11]

2.1.1. Symmetric-Key Cryptography

The technique of symmetirc-key encryption is also known as the symmetric-key, shared key, and single-key encryption. In this technique, the same secret-key is used for both encryption and decryption sides. The original information or plaintext is encrypted with a key by the sender. Then the same key is used by the receiver to decrypt the message and obtain the plaintext. The key is known only by those two parties who are authorized to do the encryption and decryption [12]. The technique provide good security for transmission. However, there is a difficulty in the key distribution. If the key is stolen the whole data security is compromised. Moreover, a secure mechanism is needed for safe key-exchange process. Examples of symmetric-key schemes include DES and AES algorithms [12].

2.1.2. Asymmetric-Key Cryptography

This technique is also known as public-key cryptography. It uses two keys, knows as public and private keys, which are mathematically associated, and separately used for encrypting and decrypting respectively. For each user, A, both keys are needed for the scheme to run. The key used for encryption is publicly available, hence it is called user A's public-key, K_{pub_A}. Therefore, all other users can access the public-key K_{pub_A} and encrypt messages to be sent to the user A. On the other hand, the private-key K_{pri_A} is only known by the user A who uses it for decryption. As a main requirement in this scheme, it is computationally infeasible to obtain private-key K_{pri_A} from the public-key K_{pub_A}. An example of asymmetric-key cryptosystem is RSA [10].

2.1.3. Hash Functions

A hash function is a one-way collision-free function with a fixed-length output. Hash functions are also called message digests. A hash function is an algorithm that does not use any key. However, a fixed-length hash value is calculate based on the input data such that it computationally infeasible to obtain the input data from the hash value, or even any input string that matches the given hash value. Hash functions are usually used to produce digital fingerprints of files and to guarantee the integrity of the files [10].

2.2 STEGANOGRAPHY

Steganography can be defined as the art of hiding data and communicating hidden data through apparently reliable carriers in attempt to hide the existence of the data itself. So, there is no knowledge of the existence of the message in the first place. Steganography techniques often use a cover, like an image or another file, to hide the secret information. If a person views the cover which the information is hidden within, there shall be no clue that there is any hidden data under the cover. In this way, the individual won't endeavour to decode the data. Figure 2 shows an overview of steganography system [10].

The secret information can be inserted into the cover media by the stego system encoder with using certain algorithm. A secret message can be plaintext, an image, ciphertext, or anything which can be represented in the form of a bit string. After the secret date is embedded in the cover object, the cover object is called a stego object. The stego object is sent to a receiver by selecting the suitable channel, where a decoder system is used with the same stego method to extract the secret information [10].

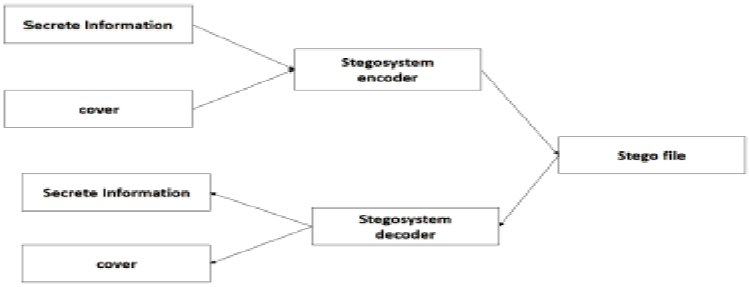


Figure 2: Steganography System

There are various types of steganography. Here are some of the common types:

1. **Text Files:** The technique of embedding secret data inside a text is identified as text stego. Text steganography needs a low memory because this type of file can only store text files. It affords fast transfer or communication of files from a sender to receiver [1].

2. **Image Files:** It is the procedure in which we embed the information inside the pixels of image. So that, the attackers cannot observe any change in the cover image. The least significant bit (LSB) approach is a common image steganography algorithm [1].

3. **Audio Files:** It is the process in which we hide the information inside an audio media. There are many approaches to hide secret information in an audio files, like: Phase Coding and LSB [1].

4. **Video Files:** It is the process of hiding some secret data inside the frames of a video [1].

2.3. Benefits of combine the Steganography and Cryptography

It is noted that steganography or cryptography alone is insufficient for the security of information in all scenarios. However, if we combine these systems, we can generate more reliable and strong systems [9].

The combination of these two strategies will improve the security of the secret information. This combination will fulfill some desired features, like: memory usage, security, and strength for sensitive information transmission across an open channel. Also, it will be a powerful mechanism which enables people to communicate without dragging the attention of eavesdroppers who does not even know of the existence of the secret information being transmitted [5].

3. LITERATURE REVIEW

The significance of network security is increasing day by day as the size and sensitivity of data being transferred across the Internet increase. This issue pushes the researchers to do many studies to provide the needed security. A solution for this issue is using the advantage of cryptography and steganography combined in one system. Many studies propose methods to combine cryptography with steganography systems in one system. These methods were deceased in previous surveys available on this topic, such as [1] published in 2014, which aims to give an overview of the methods proposed to combine cryptography with steganography systems. The authors introduced 12 methods which are combined steganography and cryptography and made a comparative analysis. This comparative has been implemented on the basis of the requirements of security, namely: authentication, confidentiality, and robustness. Another survey [12] was published in 2014. This survey presented many steganographic techniques combined with cryptography, AES Algorithm, Alteration Component, Random Key Generation, Distortion Process, Key Based Security Algorithm.

There has been a continuous rise in the number of data security threats in the recent decays. It has become a matter of concern for security experts. Cryptography and steganography are the best techniques to face these threats. Today, researchers are proposing a blended approach of both techniques to achieve a higher level of security when both techniques are used together.

In [13], the authors proposed an encrypting technique by combining cryptography and steganography to hide the data. In the cryptography process, they proposed an effective technique for data encryption using one's complement method, that they called as SC-MACS. It used a symmetric key method where both sender and receiver share the same key for encryption and decryption. In steganography part, they used the LSB method.

In [14], the authors proposed a highly-secure steganography technique by combining DNA sequence with Hyper-elliptic Curve Cryptography. This approach achieved the benefits of both techniques to obtain a high level of secure communication, besides other benefits of applying DNA cryptography and steganography. The algorithm hides a secret image in another cover image by converting them into DNA sequence using the nucleotide to the binary transformation table. On the sender side, the embedding method includes three steps. First, it converts the values of a pixel of both the cover image and secret image to their respective DNA triplet value utilizing characters to the DNA triplet conversion. Second, it converts the triplet values to binary values format. In the final stage, it applies the XOR logic between binary values of both secret image and cover image to generate a new image which called stego-image.

In [15], the authors presented a new technique called multi-level secret data hiding which integrates two different methods of encryption, namely: visual cryptography and steganography. The first step of this method is to use a method called halftoning, which is used to reduce the pixels and simplify the processing. After that visual cryptography is performed, it produces the shares which form the first level of security, and then steganography is applied using the LSB method to hide the shares in different media like image, audio, and video.

The work in [16] presents a method based on combining both strong encryption algorithm and steganographic technique to make the communication of confidential information safe, secure and extremely hard to decode. An encryption algorithm is employed first to encrypt the secret message before encoding it into a QR code. They used AES-128 to encrypt a message, in UTF-8 format, and converted it into base 64 format to make it compatible for further processing. The encoded image is scrambled to achieve another security level. The scrambled QR code is finally embedded in a suitable cover image, which is then transferred securely to deliver the secret information. They utilized a LSB method to accomplish the digital image steganography. At the receivers side, the secret data is retrieved through the decoding process. Thus, a four-level security has been rendered for a secret message to be transferred.

In [17], the authors presented an image steganography method. At first, they used DES algorithm to encrypt the text message. They used a 16 round DES with 64-bit block size. After that the K-Means Clustering of the pixels which clusters the image into numerous segments and embedded data in every segment. There are many clustering algorithms used for image segmentation. Segmentation includes a huge set of information in the form of pixels, where every pixel additional has three components namely red, green and blue (RGB). After the formation of the clusters, the encrypted text is separated into K segments. These segments are to be hidden in each cluster. They used the LSB method for this purpose.

In [18], the authors concluded that cryptography or steganography alone cannot be used for transmission of data because each has its own weaknesses. So, they proposed a system in which both techniques are used to create a secure system. They claimed that it is nearly impossible for a third party to breach the system and gain confidential data. The system used the TwoFish algorithm for encryption, while a new approach for performing the steganography is used which called Adaptive B45 steganography technique.

In [19], the authors presented a method to extend the embedding capacity and to enhance the quality of stego-images. The Adaptive Pixel Value Differencing, which is an improved form of Pixel Value Differencing, was utilized as the Steganographic system. AES was utilized as the Cryptographic system. In their method, they used an image as a cover to hide the secret data. These covering images should be in grayscale of size 256 x 256 bits. If the size is higher, they brought it to this range. If the cover image is a color image, they changed it into the grayscale range. They used APVD algorithm to embed the data into the cover image. The resultant stego-image is then encrypted using AES algorithm. It is important to notice here that the encypted stego-image is left uncovered.

In [20], the authors conducted a performance analysis survey on various algorithms like DES, AES, RSA combined with LSB substitution technique which serves well to draw conclusions on the three encryption techniques based on their performances in any application. It has been concluded from their work that AES encryption is better than the other techniques as it accounts for less encryption and decryption times, and uses less memory as buffering space.

In [21], the authors performed a modern method in which Huffman encoding is used to hide data. They took a gray level image of size *m* x *n* as a covering image and a *p* x *q* secret image. Then, they executed the Huffman encoding over the secret image and every bit of Huffman code of a secret image is hidden into a cover image using LSB method.

In [22], the authors suggested a new steganographic technique based on gray-level modification for true color images using a secret key, cryptography and image transposition. Both the secret key and the secret information are first encrypted using multiple encryption algorithms (Bit-Xor operation, stego key-based encryption, and bits shuffling). These are, later, hidden in the cover image pixels. In addition, the input image is changed before data hiding. Image transposition, Bit-Xoring, stego key-based encryption, bits shuffling, and gray-level modification introduces five various security levels to the suggested technique, making the recovery of data is very difficult for attackers.

In [23], the authors proposed an approach which uses Blowfish to encrypt the secret information before embedding it in the image using LSB method.

In [24], the authors encrypted the secret data using AES algorithm and hashed the key using SHA-1 to prevent attacks. After that, they used the LSB technique to embed the encrypted information in an image, video or audio. The receiver must recover the key which is hashed at the sender side. The secret data can be hidden in any type of media which affords more security.

In [25], hiding information using steganography and cryptography is discussed. A new approach is explained to secure data without decreasing the quality of the image as a cover medium. The steganographic method is used by finding the similarity bit of the message with a bit of the most significant bit (MSB) of the covering image. They used divide and conquer approach for finding the similarity. The outcomes are the bit index position, which is later encrypted using DES algorithm.

In [26], the authors proposed a new method. First, the secret message is changed into cipher text using RSA algorithm and next they hide the cipher text in an audio media using LSB audio steganography technique. At the receiver side, the ciphertext is extracted from audio media then decrypted it into a message by using RSA decryption. So, this technique combines the characteristic of both public-key cryptography and steganography to provide a higher level of security.

In [27], the authors used Blowfish algorithm to encrypt a secret image. They claimed that Blowfish is faster, stronger, and provides better performance than RC6, RC4, DES, 3DES, and AES. They selected a secret image in BMP format and encrypted it by Blowfish. Then, they used LSB method to embed the encrypted image into video frames. This method provides authenticity, integrity, confidentiality and non-repudiation.

The paper [28], is similar to the method mentioned in [27] but the only difference is that the text is selected to be a secret message instead of an image, and it is encrypted using Blowfish algorithm. Next, an image is used to be a cover object with the LSB method to embed the encrypted text into this cover.

In [29], the authors proposed a new strategy employs RSA algorithm with a key of size 128 for encrypting the secret information before embedding it into a cover image, and use F5 steganographic algorithm to embed the encrypted message in the cover image gradually. They selected Discrete Courier Transform (DCT) with random coefficients to embed the secret message by the F5 algorithm. They applied matrix embedding to reduce the changes to be made to the length of the message. This strategy gives a fast system, with a high steganographic capacity, and can prevent observing and analytical attacks.

In [30], the authors have proposed a novel visual cryptographic technique. This technique is suitable for both Grayscale and Bitmap color images. In this approach, the theory of Residual Number System (RNS) was utilized based on the Chinese Remainder Theorem (CRT) for shares creation and shares stacking of a given image. First, they embedded a secret image in a cover image to make a stego-image. An 8-bit pixel of the stego-image is selected and added with an 8-bit key to produce a cipher pixel. They use addition modulo 256 and a pseudo-random number generator with a mixed key generation technique to generated the key. After they encrypted the stego-image, they mapped the cipher pixel into RNS of n elements. Finally, they collected and send the n elements. This approach is extremely fast, secure, reliable, efficient and easy to implement.

In [31], the combination of cryptography and image steganography is achieved by utilizing both AES and LSB algorithms. The authors uses the LSB method to embed the secret information into an image file and they used AES algorithm for encrypting the stego-image. The authors concluded that this technique is effective for secret communication.

4. COMPARATIVE ANALYSIS OF SURVEYED METHODS

In this section, we briefly summarize the differences between cryptography and steganography. Then present a comparative analysis of the methods surveyed in Section 3. Table 1 shows the differences between the steganography and cryptography. The comparison is in terms of:

definition, objective, carrier, number of input files, importance of the key, visibility, security services offered, type of attack, attacks, resultant output, and applications.

Table 1: Cryptography vs Steganography

Criteria/Method	Steganography	Cryptography
Definition	Cover writing [7, 1]	Secret writing [7, 1]
Objective	Maintaining existence of a message secret, Secret communication [7, 1,5]	Maintaining contents of a message secret, Data protection [7, 1, 5]
Carrier	Any digital media [7, 1, 6, 10, 8]	Usually text based [7, 1, 6, 10, 8]
Input file	At least two [6]	One [6]
Key	Optional [6, 7, 8, 1]	Necessary [6, 7, 8, 1]
Visibility	Never [6, 1, 7]	Always [6, 1, 7]
Security services offered	Authentication, Confidentiality,Identification [10]	Confidentiality, Identification, Data Integrity and authentication Non-repudiation [6, 7, 1, 10]
Type of Attack	Steganalysis: Analysis of a file with an aim of finding whether it is stego file or not [6, 1, 10, 8]	Cryptanalysis [6, 1, 10, 8]
Attacks	Broken when attacker reveals that steganography has been used. known as Steganalysis. [6, 5, 7, 1]	Broken when attacker can under-stand the secret message. known as Cryptanalysis [6, 5, 7, 1].
Resultant Output	Stego file [6, 1, 8]	Ciphertext [6, 1, 8]
Applications	Used for securing information against potential eavesdroppers [10]	Used for securing information against potential eavesdroppers [10]

According to the methods surveyed in Section 3, we observed that most of these approaches apply the encryption (cryptography) before the covering (steganography). We classify these methods as Class-A. On the other hand, we classify the methods in which steganography is performed before cryptography as Class-B. This is a useful classification since the methods in the same class usually have similar features. The proposed classes, A and B, of the surveyed methods are illustrated in Figures 3 and 4 respectively.

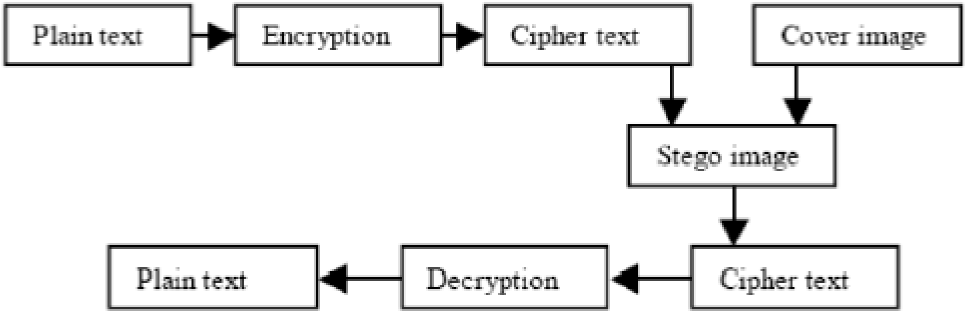


Figure 3: Class A [32]

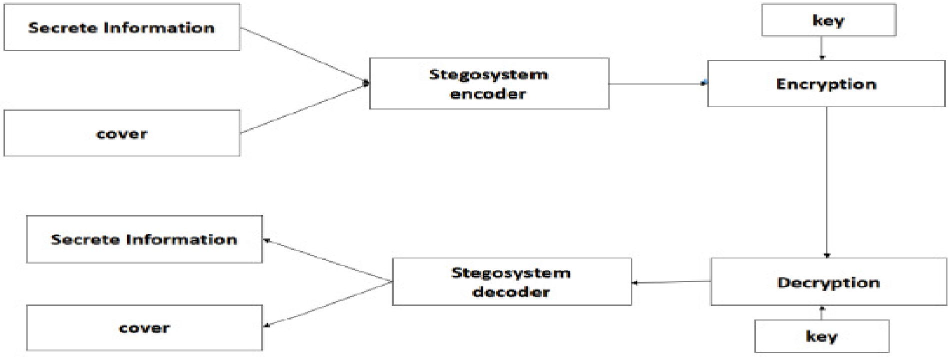


Figure 4: Class B

We included in this study the encryption algorithm used in surveyed methods. The algorithms used in these methods are: AES, DES, Twofish, Blowfish, RSA, etc. Another aspect of this study is the steganography technique and the file type used for covering.

Our study shows that Class-A methods are more popular in the research than the ones of Class-B. Class-A methods have higher security levels and less risk exposing than Class-B since ciphertexts in Class-A is hidden by the steganography technique. While in Class-B, the encrypted stego-image is exposed.

According to the authors of the methods in Class-B, which are mentioned in the literature review, the method of Class-B usually provides larger space for hiding information inside the cover object, because the encryption process is applied to all data inside the cover object. The drawback of this class is that the output file of the encryption process will be vulnerable to suspect of the existence of a secret data inside it. Table 2 summarizes these results.

Table 2: A Comparative Analysis of Surveyed Methods

System	Year	Class	Encryptosystem	Stegosystem	File Type
[13]	2015	A	SCMACS	LSB	Any In Image
[14]	2016	B	HECC	DNA&XOR	Image In Image
[15]	2016	A	VCS	LSB	Image In any
[16]	2016	A	AES-128	QR Code&LSB	Text In Image
[17]	2016	A	DES	LSB	Text In Image
[18]	2016	A	TwoFish	B45	Text In Image
[19]	2015	B	AES	APVD	Any In Image
[20]	2013	A	AES	LSB	Text In Image
[22]	2015	A	Multiple Encryption	XOR	Text In Image
[23]	2015	A	Blowfish	LSB	Text In Image
[24]	2012	A	AES KEY SHA-1	LSB	Any in Any
[25, 32]	2015	B	DES	Same Bit & MSB	Text In Image
[26]	2015	A	RSA	LSB	Text In Audio
[27]	2013	A	Blowfish	LSB	Image In Video
[28]	2013	A	Blowfish	LSB	Text In Image
[29]	2014	A	RSA Key 1024bit	F5	Text In Image
[30]	2013	B	VCS	LSB	Image In Image
[31]	2013	B	AES	LSB	Text In Image

5. CONCLUSION

In this paper, the concepts of cryptography, steganography and their applications in the security of digital data communication across network is studied. A comprehensive technical survey of recent methods which combined steganography and cryptography is presented. Combining these two techniques is found to be more secure than applying each one of them separately.

A detailed comparison of methods combining cryptography and steganography techniques is presented. A useful classification of these methods is proposed. Our study shows that Class-A methods are more common than Class-B and provide better security with less exposing of the encrypted data. The only advantage of Class-B as claimed by the authors of the surveyed methods in this class is providing more capacity for the secret information.

REFERENCES

[1] M. K. I. Rahmani and N. P. Kamiya Arora, \A crypto-steganography: A survey," International Journal of Advanced Computer Science and Application, vol. 5, pp.149{154, 2014.

[2] J. V. Karthik and B. V. Reddy, \Authentication of secret information in image stenography," International Journal of Computer Science and Network Security (IJCSNS), vol. 14, no. 6, p. 58, 2014.

[3] M. H. Rajyaguru, \Crystography-combination of cryptography and steganography with rapidly changing keys," International Journal of Emerging Technology and Advanced Engineering, ISSN, pp. 2250{2459, 2012.

[4] D. Seth, L. Ramanathan, and A. Pandey, \Security enhancement: Combining cryptography and steganography," International Journal of Computer Applications (0975{8887) Volume, 2010.

[5] H. Abdulzahra, R. AHMAD, and N. M. NOOR, \Combining cryptography and steganography for data hiding in images," ACACOS, Applied Computational Science, pp. 978{960, 2014.

[6] P. R. Ekatpure and R. N. Benkar, \A comparative study of steganography & cryptography," 2013.

[7] N. Khan and K. S. Gorde, \Data security by video steganography and cryptography techniques," 2015.

[8] M. K. I. Rahmani and M. A. K. G. M. Mudgal, \Study of cryptography and steganography system," 2015.

[9] C. P. Shukla, R. S. Chadha, and A. Kumar, \Enhance security in steganography with cryptography," 2014.

[10] P. Kumar and V. K. Sharma, \Information security based on steganography & cryptography techniques: A review," International Journal, vol. 4, no. 10, 2014.

[11] J. K. Saini and H. K. Verma, \A hybrid approach for image security by combining encryption and steganography," in Image Information Processing (ICIIP), 2013 IEEE Second International Conference on. IEEE, 2013, pp. 607{611.

[12] H. Sharma, K. K. Sharma, and S. Chauhan, \Steganography techniques using cryptography-a review paper," 2014.

[13] A. Dhamija and V. Dhaka, \A novel cryptographic and steganographic approach for secure cloud data migration," in Green Computing and Internet of Things (ICG-CIoT), 2015 International Conference on. IEEE, 2015, pp. 346{351.

[14] P. Vijayakumar, V. Vijayalakshmi, and G. Zayaraz, \An improved level of security for dna steganography using hyperelliptic curve cryptography," Wireless Personal Communications, pp. 1{22, 2016.

[15] S. S. Patil and S. Goud, \Enhanced multi level secret data hiding," 2016.

[16] B. Karthikeyan, A. C. Kosaraju, and S. Gupta, \Enhanced security in steganography using encryption and quick response code," in Wireless Communications, Signal Processing and Networking (WiSPNET), International Conference on. IEEE, 2016, pp. 2308{2312.

[17] B. Pillai, M. Mounika, P. J. Rao, and P. Sriram, \Image steganography method using k-means clustering and encryption techniques," in Advances in Computing, Communications and Informatics (ICACCI), 2016 International Conference on. IEEE, 2016, pp. 1206{1211.

[18] A. Hingmire, S. Ojha, C. Jain, and K. Thombare, \Image steganography using adaptive b45 algorithm combined with pre-processing by twofish encryption," International Educational Scientific Research Journal, vol. 2, no. 4, 2016.

[19] F. Joseph and A. P. S. Sivakumar, \Advanced security enhancement of data before distribution," 2015.

[20] B. Padmavathi and S. R. Kumari, \A survey on performance analysis of des, aes and rsa algorithm along with lsb substitution," IJSR, India, 2013.

[21] R. Das and T. Tuithung, \A novel steganography method for image based on huffman encoding," in Emerging Trends and Applications in Computer Science (NCETACS), 2012 3rd National Conference on. IEEE, 2012, pp. 14{18.

[22] K. Muhammad, J. Ahmad, M. Sajjad, and M. Zubair, \Secure image steganography using cryptography and image transposition," arXiv preprint arXiv:1510.04413, 2015.

[23] T. S. Barhoom and S. M. A. Mousa, \A steganography lsb technique for hiding image within image using blowfish encryption algorithm," 2015.

[24] S. E. Thomas, S. T. Philip, S. Nazar, A. Mathew, and N. Joseph, \Advanced cryptographic steganography using multimedia files," in International Conference on Electrical Engineering and Computer Science (ICEECS-2012), 2012.

[25] M. A. Muslim, B. Prasetiyo et al., \Data hiding security using bit matching-based steganography and cryptography without change the stego image quality," Journal of Theoretical and Applied Information Technology, vol. 82, no. 1, p. 106, 2015.

[26] A. Gambhir and A. R. Mishra, \A new data hiding technique with multilayer security system." 2015.

[27] M. H. Sharma, M. MithleshArya, and M. D. Goyal, \Secure image hiding algorithm using cryptography and steganography," IOSR Journal of Computer Engineering (IOSR-JCE) e-ISSN, pp. 2278{0661, 2013.

[28] A. Singh and S. Malik, \Securing data by using cryptography with steganography," International Journal of Advanced Research in Computer Science and Software Engineering, vol. 3, no. 5, 2013.

[29] M. Mishra, G. Tiwari, and A. K. Yadav, \Secret communication using public key steganography," in Recent Advances and Innovations in Engineering (ICRAIE), 2014. IEEE, 2014, pp. 1{5.

[30] R. H. Kumar, P. H. Kumar, K. Sudeepa, and G. Aithal, \Enhanced security system using symmetric encryption and visual cryptography," International Journal of Advances in Engineering & Technology, vol. 6, no. 3, p. 1211, 2013.

[31] D. R. Sridevi, P. Vijaya, and K. S. Rao, \Image steganography combined with cryptography," Council for Innovative Research Peer Review Research Publishing System Journal: IJCT, vol. 9, no. 1, 2013.

[32] P. Budi, R. Gernowo, M. Si, B. Noranita, S. Si, and M. Kom, \The combination of bit matching-based steganography and des cryptography for data security," 2013.

ATTRIBUTE REDUCTION-BASED ENSEMBLE RULE CLASSIFIERS METHOD FOR DATASET CLASSIFICATION

Mohammad Aizat bin Basir[1] and Faudziah binti Ahmad [2]

[1] Universiti Malaysia Terengganu (UMT)Terengganu, Malaysia
[2] Universiti Utara Malaysia(UUM) Kedah, Malaysia

ABSTRACT

Attribute reduction and classification task are an essential process in dealing with large data sets that comprise numerous number of input attributes. There are many search methods and classifiers that have been used to find the optimal number of attributes. The aim of this paper is to find the optimal set of attributes and improve the classification accuracy by adopting ensemble rule classifiers method. Research process involves 2 phases; finding the optimal set of attributes and ensemble classifiers method for classification task. Results are in terms of percentage of accuracy and number of selected attributes and rules generated. 6 datasets were used for the experiment. The final output is an optimal set of attributes with ensemble rule classifiers method. The experimental results conducted on public real dataset demonstrate that the ensemble rule classifiers methods consistently show improve classification accuracy on the selected dataset. Significant improvement in accuracy and optimal set of attribute selected is achieved by adopting ensemble rule classifiers method.

KEYWORDS

Attribute Selection, Ensemble, and Classification.

1. INTRODUCTION

Real world dataset usually consist a large number of attributes. It is very common some of those input attributes could be irrelevant and consequently give an impact to the design of a classification model. In situations where a rule has too many conditions, it becomes less interpretable. Based on this understanding, it becomes important to reduce the dimensionality (number of input attributes in the rule) of the rules in the rule set. In practical situations, it is recommended to remove the irrelevant and redundant dimensions for less processing time and labor cost. The amount of data is directly correlated with the number of samples collected and the number of attributes. A dataset with a large number of attributes is known as a dataset with high dimensionality [1]. The high dimensionality of datasets leads to the phenomenon known as the curse of dimensionality where computation time is an exponential function of the number of the dimensions. It is often the case that the model contains redundant rules and/or variables. When faced with difficulties resulting from the high dimension of a space, the ideal approach is to decrease this dimension, without losing the relevant information in the data. If there are a large number of rules and/or attributes in each rule, it becomes more and more vague for the user to understand and difficult to exercise and utilize. Rule redundancy and/or attribute complexity could overcome by reducing the number of attributes in a dataset and removing irrelevant or less

David C. Wyld et al. (Eds) : ITCS, SIP, CST, ARIA, NLP - 2017
pp. 75– 82, 2017. © CS & IT-CSCP 2017 DOI : 10.5121/csit.2017.70307

significant roles. This can reduce the computation time, and storage space. Models with simpler and small number of rules are often easier to interpret.

The main drawback of rule/attributes complexity reduction is the possibility of information loss. It is important to point out that two critical aspects of the attribute reduction problem are the degree of attribute optimality (in terms of subset size and corresponding dependency degree) and time required to achieve this attribute optimality. For example, existing methods such as Quick Reduct and Entropy-Based Reduction (EBR) methods find reduced in less time, but could not guarantee a minimal subset [1] –[3] whereas other hybrid methods which combine rough sets and swarm algorithm such as GenRSAR, AntRSAR, PSO-RSAR and BeeRSAR methods improve the performance but consume more time [1], [2].

In feature selection, also known as variable selection, attribute selection or variable subset selection is the process of selecting a subset of relevant features (attributes) for use in model construction. It is the process of choosing a subset of original features so that the feature space is optimally reduced to evaluation criterion. Feature selection can reduce both the data and the computational complexity. The raw data collected is usually large, so it is important to select a subset of data by creating feature vectors. Feature subset selection is the process of identifying and removing much of the redundant and irrelevant information possible.

However, the use of a subset of a feature set may disregard important information contained in other subsets. Consequently, classification performance is reduced. Therefore, this paper aims to find the optimal set of attributes and improve the classification accuracy by adopting the ensemble classifier method. Firstly, an optimal set of attribute subsets are extracted by applying various search method and a reduction algorithm to the original dataset. Then an optimal set of attributes further classified by adopting a classification ensemble approach. In the experiment, 6 various datasets were used. The experiment results showed that the performance of the ensemble classifier was improved the classification accuracy of the dataset. This paper is organized as follows: in Section II, related works are discussed. The proposed methodology is presented in Section III. In Section IV, the results and discussion are given. Finally, the conclusions presented in Section V.

2. RELATED WORKS

There many research in feature selection methods for constructing an ensemble of classifiers. The ensemble feature selection method is where a set of the classifiers, each of which solve the same original task, are joined in order to obtain a better combination global classifier, with more accurate and reliable estimates or decisions than can be obtained from using a single classifier. The aim of designing and using the ensemble method is to achieve a more accurate classification by combining many weak learners.

Previous studies show that methods like bagging improve generalization by decreasing variance. In contrast, methods similar to boosting achieve this by decreasing the bias [4]. [5] demonstrated a technique for building ensembles from simple Bayes classifiers in random feature subsets.
[6] explored tree based ensembles for feature selection. It uses the approximately optimal feature selection method and classifiers constructed with all variables from the TIED dataset.

[7] presented the genetic ensemble feature selection strategy, which uses a genetic search for an ensemble feature selection method. It starts with creating an initial population of classifiers where each classifier is generated by randomly selecting a different subset of features. The final ensemble is composed of the most fitted classifiers.

[8] suggested a nested ensemble technique for real time arrhythmia classification. A classifier model was built for each 33 training sets with enhanced majority voting technique. The nested ensembles can relieve the problem of the unlikelihood of a classifier being generated when learning the classifier by an old dataset and limited input features. One of the reasons that make the ensemble method popular is that ensemble methods tend to solve dataset problems.

3. METHODOLOGY

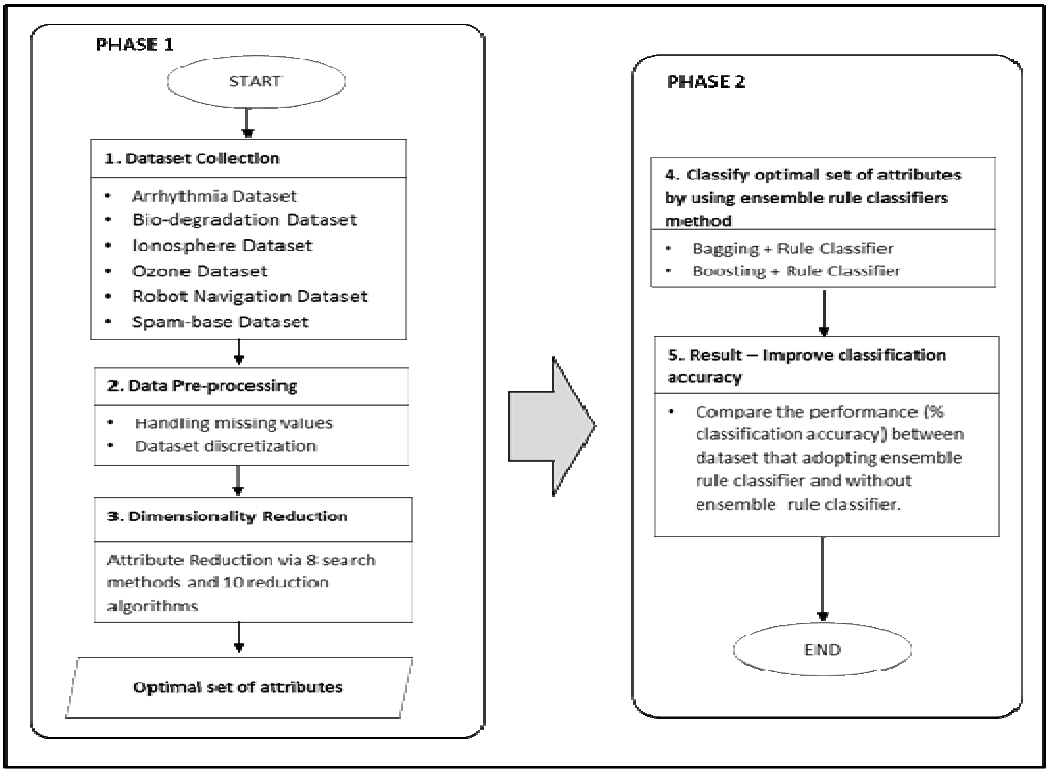

PHASE 1
START
1. Dataset Collection
• Arrhythmia Dataset
• Bio-degradation Dataset
• Ionosphere Dataset
• Ozone Dataset
• Robot Navigation Dataset
• Spam-base Dataset
2. Data Pre-processing
• Handling missing values
• Dataset discretization
3. Dimensionality Reduction
Attribute Reduction via 8 search methods and 10 reduction algorithms
Optimal set of attributes
PHASE 2
4. Classify optimal set of attributes by using ensemble rule classifiers method
• Bagging + Rule Classifier
• Boosting + Rule Classifier
5. Result – Improve classification accuracy
• Compare the performance (% classification accuracy) between dataset that adopting ensemble rule classifier and without ensemble rule classifier.
END

Figure 1. Methodology

The methodology is shown in Fig. 1. It consists of five (5) steps: (1) data collection; (2) data pre-processing; (3) dimensionality reduction; (4) classify an optimal set of attributes by using the ensemble rule classifier method; (5) Result-improved classification accuracy: ensemble rule classifier methods have been compared with datasets that do not use the ensemble rule classifier method. The output of phase 1 (step 1 – 3) is the optimal set of attributes. For phase 2 (step 4 – 5), the output is the improved classification accuracy by adopting an ensemble rule classifier method for the classification task. The details of the steps involved are described below:-

Step 1 (Data Collection): Six (6) different datasets were selected from UCI Machine Learning Repository. Arrhythmia dataset is one of the dataset selected due to its many features that make it challenging to explore [9]. Other five (5) datasets also were taken from different domain in order to confirm the suitability of the ensemble classifiers.

Step 2 (Data Pre-processing): Dataset that has missing values has been pre-processed in order to make sure that the dataset is ready to be experimented. All datasets were discretized since it has numeric data but needs to use classifier that handles only nominal values.

Step 3 (Dimensionality Reduction): 8 search methods and 10 reduction algorithms have been used in order to get the optimal set of attributes. The output of this step is the optimal set of attributes.

Step 4 (Classify optimal set of attributes by using the ensemble rule classifier method): In this step, the optimal sets of attributes obtained from previous step were classified by adopting the ensemble classifier method.

Step 5 (Model with good accuracy): In this step, the performance (% classification accuracy) of the dataset that used ensemble rule classifier methods has been compared with datasets that do not use the ensemble rule classifier method. The output of this step is the improved classification accuracy with optimal number of attributes.

Standard six datasets namely Arrhythmia, Bio-degradation, Ionosphere, Ozone, Robot Navigation and Spam-base from the UCI [10] were used in the experiments. These datasets include discrete and continuous attributes and represent various fields of data. The reason for choosing this dataset is to confirm the ensemble classifier is suited to all fields of data. The information on the datasets is shown in Table I.

Table 1. Dataset Characteristics.

Dataset	# of Attributes	# of Instances	# of Classes
Arrhythmia	279	452	16
Bio-degradation	41	1055	2
Ionosphere	34	351	2
Ozone	72	2536	2
Robot Navigation	24	5456	4
Spam-base	57	4601	2

All six (6) datasets were tested using 8 search methods and 10 reduction algorithms.

3. RESULTS AND DISCUSSION

The outputs for phase 1 and phase 2 are presented in this section. The performance results are presented in the percentage of classification accuracy with the optimal set of attributes.

3.1. Phase 1 (Step 1 – 3)

Table 2. List of an optimal set of attributes selected.

Dataset	Search Method	Reduction Algorithm	# of Attr	#of Sel Attr
Arrhythmia	Best First Search	WrapperSubsetEval	279	19
Bio-degradation	Best First Search	WrapperSubsetEval	41	10
Ionosphere	Greedy Stepwise	WrapperSubsetEval	34	8
Ozone	Race Search	ClassifierSubsetEval	72	5
Robot Navigation	SubsetSizeForwardSelec	CFSSubsetEval	24	6
Spam-base	Genetic Search	WrapperSubsetEval	57	18

Table 2 shows the results of an optimal set of attributes selected by using various search method and reduction algorithm. In phase one (1), eight (8) search methods, namely Best First Search, Genetic Search, Exhaustive Search, Greedy Stepwise Search, Linear Forward Selection Search, Scatter Search, Subset Size Forward Selection Search and Ranker Search were applied. In addition, ten (10) reduction algorithms that are CfsSubsetEval, ClassifierSubsetEval, ConsistencySubsetEval, FilteredSubsetEval, ChisquaredAttributeEval, FilteredAttributeEval, GainRatioAttributeEval, InfoGainAttributeEval, PrincipalComponent and WrapperSubsetEval were adopted. It can be seen that Arrhythmia and Ozone dataset produced a massive attribute reduction, which is more than 90% reduction. Best first search (BSF) was used with WrapperSubsetEval for Arrhythmia dataset since BFS is a robust searching [11] and better for dataset studied [12]. The rest of the dataset achieved more than 60% attribute reduction. Wrapper method (WrapperSubsetEval) performed better for 4 out of 6 datasets selected with combination of various search method. These experiments confirmed that significance attribute reduction can be accomplished by combining the right search method and reduction algorithm.

3.2. Phase 2 (Step 4 – 5)

In phase 2, each selected set of attributes for the six (6) various dataset namely Arrhythmia, Bio-degradation, Ionosphere, Ozone, Robot Navigation and Spam-base were classified using ensemble rule classifier methods of boosting, bagging and voting. In this phase, rule classifiers like Repeated Incremental Pruning to Produce Error Reduction (RIPPER), PART, Prism, Nearest Neighbor With Generalization (NNge) and OneR were evaluated with ensemble method. 70% of the dataset being used as training and the remaining 30% was used for testing data. The results are shown in Table 3 through Table 6.

Table 3. Classification Result of using RIPPER and RIPPER with Ensemble Rule Classifier Method.

Dataset	Without Ensemble Rule Classifier	With Ensemble Rule Classifier	
	RIPPER	Boosting + RIPPER	Bagging + RIPPER
	Acc (%)	Acc (%)	Acc (%)
Arrhythmia	73.67	73.41	73.80
Bio-degradation	83.50	83.94	83.72
Ionosphere	92.87	93.63	93.54
Ozone	93.62	93.67	93.62
Robot Navigation	96.28	97.73	97.16
Spam-base	92.65	93.24	92.92

Table III shows the classification result of using RIPPER and RIPPER with the ensemble method. RIPPER [13] with boosting and bagging method improves the classification accuracy of 4 datasets namely Bio-degradation, Ionosphere, Robot Navigation and Spam-base. These results are in line with the strength of the RIPPER that it tries to increase the accuracy of rules by replacing or revising individual rules [14]. It uses a reduced error pruning, which isolates some training data in order to decide when to stop adding more conditions to a rule. It also used a heuristic based on the minimum description length principle as stopping criterion.

Table 4. Classification Result of using PART and PART with Ensemble Rule Classifier Method

Dataset	Without Ensemble Rule Classifier	With Ensemble Rule Classifier	
	PART	Boosting + PART	Bagging + PART
	Acc (%)	Acc (%)	Acc (%)
Arrhythmia	74.13	74.98	76.93
Bio-degradation	83.94	83.86	84.69
Ionosphere	90.78	92.62	91.95

Ozone	93.76	93.86	93.81
Robot Navigation	96.88	99.06	97.74
Spam-base	93.53	93.81	93.98

Table 4 shows the classification result of using PART and PART with ensemble method. PART rule classifier with bagging method increased the classification accuracy of all the datasets. The PART algorithm [15] is a simple algorithm that does not perform global optimization to produce accurate rules. It adopts the separate-and-conquer strategy by building a rule, removes the instances; it covers, and continues creating rules recursively for the remaining instances until there are no more instances left. In addition, many studies have shown that aggregating the prediction of multiple classifiers can improve the performance achieved by a single classifier [16]. In this case, Bagging is known as a "bootstrap" ensemble method that creates individuals for its ensemble by training each classifier on a random redistribution of the training set. In contrast, Boosting method with PART rule classifier performed better accuracy for Robot Navigation dataset with more than 3% accuracy. In this case, these results are consistent with data obtained in [17] which proved that PART algorithm is the effective algorithm to be used for classification rule hiding.

Table 5. Classification Result of using PRISM and PRISM with Ensemble Rule Classifier Method

Dataset	**Without Ensemble Rule Classifier**	**With Ensemble Rule Classifier**	
	Prism	**Boosting + Prism**	**Bagging + Prism**
	Acc (%)	**Acc (%)**	**Acc (%)**
Arrhythmia	62.36	61.71	66.07
Bio-degradation	53.42	58.71	54.61
Ionosphere	89.77	91.87	91.03
Ozone	93.79	93.75	93.88
Robot Navigation	95.21	95.35	97.38
Spam-base	80.46	81.08	81.22

Table 5 shows the Classification result of using Prism and Prism with the ensemble rule classifier method. Prism is an algorithm used different strategy to induce rules which are modules that can avoid many of the problems associated with decision trees [18]. Prism rule classifier with bagging method performed well to enhance all the dataset. In addition, boosting method with Prism produced better accuracy result for Ionosphere and Spam-base Dataset.

Table 6. Classification Result of using OneR and OneR with Ensemble Rule Classifier Method

Dataset	**Without Ensemble Rule Classifier**	**With Ensemble Rule Classifier**	
	OneR	**Boosting + OneR**	**Bagging + OneR**
	Acc (%)	**Acc (%)**	**Acc (%)**
Arrhythmia	59.76	59.69	59.37
Bio-degradation	77.03	81.69	77.03
Ionosphere	87.26	91.53	87.17
Ozone	93.88	93.84	93.82
Robot Navigation	76.01	85.09	75.99
Spam-base	79.19	90.80	79.79

Table 6 shows the classification result of using OneR and OneR with the ensemble rule classifier method. Boosting Method with OneR rule classifier performed a lot better accuracy for Bio-degradation, Ionosphere, Robot Navigation and Spam-base dataset. Huge accuracy improvement using OneR rule classifier with Boosting method for Spam-base dataset which is more than 10% accuracy increased. In this case, OneR demonstrated the efficacy as an attribute subset selection algorithm in similar cases in [20].

In summary, results have shown significant improvement in term of classification accuracy when using the ensemble rule classifier method.

4. CONCLUSIONS

In this paper, eight (8) search methods with ten (10) reduction algorithms were tested with 6 datasets. Experimental results benchmark dataset demonstrates that the ensemble method, namely bagging and boosting with rule classifiers which are (RIPPER), PART, Prism, (NNge) and OneR significantly perform better than other approaches of not using the ensemble method. Beside these, it is found that right combination between search methods and reduction algorithms shown good performance on extracting an optimal number of attributes. For future research, methods of finding the suitable match between search method, reduction algorithm and ensemble classifiers can be developed to get a better view of the datasets.

ACKNOWLEDGEMENTS

The authors wish to thank Universiti Malaysia Terengganu (UMT), Universiti Utara Malaysia (UUM) and Kementerian Pendidikan Malaysia (KPM). This work was supported by UMT, UUM and KPM, Malaysia.

REFERENCES

[1] R. Jensen and Q. Shen, “Finding rough set reducts with ant colony optimization,” Proc. 2003 UK Work., vol. 1, no. 2, pp. 15–22, 2003.

[2] N. Suguna and K. Thanushkodi, “A Novel Rough Set Reduct Algorithm for Medical Domain Based on Bee Colony,” vol. 2, no. 6, pp. 49–54, 2010.

[3] B. Yue, W. Yao, A. Abraham, and H. Liu, “A New Rough Set Reduct Algorithm Based on Particle Swarm Optimization,” pp. 397–406, 2007.

[4] R. E. Schapire, Y. Freund, P. Bartlett, and W. S. Lee, “Boosting the margin: A new explanation for the effectiveness of voting methods,” Ann. Stat., vol. 26, no. 5, pp. 1651–1686, 1998.

[5] A. Tsymbal, S. Puuronen, and D. W. Patterson, “Ensemble feature selection with the simple Bayesian classification,” Inf. Fusion, vol. 4, no. 2, pp. 87–100, 2003.

[6] E. Tuv, “Feature Selection with Ensembles , Artificial Variables , and Redundancy Elimination,” J. Mach. Learn. Res., vol. 10, pp. 1341–1366, 2009.

[7] D. W. Opitz, “Feature selection for ensembles,” Proc. Natl. Conf. Artif. Intell., pp. 379–384, 1999.

[8] M. E. A. Bashir, M. Akasha, D. G. Lee, G. Yi, K. H. Ryu, E. J. Cha, J.-W. Bae, M.-C. Cho, and C. W. Yoo, “Nested Ensemble Technique for Excellence Real Time Cardiac Health Monitoring.,” in International Conference on Bioinformatics & Computational Biology, BIOCOMP 2010, July 12-15, 2010, Las Vegas Nevada, USA, 2 Volumes, 2010, pp. 519–525.

[9] E. Namsrai, T. Munkhdalai, M. Li, J.-H. Shin, O.-E. Namsrai, and K. H. Ryu, “A Feature Selection-based Ensemble Method for Arrhythmia Classification,” J Inf Process Syst, vol. 9, no. 1, pp. 31–40, 2013.

[10] D. Aha, P. Murphy, C. Merz, E. Keogh, C. Blake, S. Hettich, and D. Newman, "UCI machine learning repository," University of Massachusetts Amherst, 1987. .

[11] M. GINSBERG, Essentials of Artificial Intelligence. Elsevier, 1993.

[12] R. Kohavi and G. H. John, "Wrappers for feature subset selection," Artif. Intell., vol. 97, no. 1–2, pp. 273–324, 1997.

[13] W. W. Cohen, "Fast effective rule induction," in Proceedings of the Twelfth International Conference on Machine Learning, 1995, pp. 115–123.

[14] F. Loen, M. H. Zaharia, and D. Galea, "Performance Analysis of Categorization Algorithms," in Proceeding of th 8th International Symposium on Automatic Control and Computer Science, lasi, 2004.

[15] E. Frank and I. H. Witten, "Generating accurate rule sets without global optimization," in Work, 1998, pp. 144–151.

[16] C. D. Sutton, "Classification and Regression Trees, Bagging, and Boosting," Handbook of Statistics, vol. 24. pp. 303–329, 2004.

[17] S. Vijayarani and M. Divya, "An Efficient Algorithm for Classification Rule Hiding," Int. J. Comput. Appl., vol. 33, no. 3, pp. 975–8887, 2011.

[18] J. Cendrowska, "PRISM: An algorithm for inducing modular rules," lnL J. Man-Machine Stud., vol. 27, pp. 349–370, 1987.

[19] A. Tiwari and A. Prakash, "Improving classification of J48 algorithm using bagging,boosting and blending ensemble methods on SONAR dataset using WEKA," Int. J. Eng. Tech. Res., vol. 2, no. 9, pp. 207–209, 2014.

[20] C. Nevill-Manning, G. Holmes, and I. H. Witten, "The Development of Holte's 1R Classifier."

AUTHORS

Mohammad Aizat bin Basir is currently a lecturer in Universiti Malaysia Terengganu (UMT), Malaysia.

Faudziah binti Ahmad is currently Assoc. Prof. in computer science in Universiti Utara Malaysia (UUM), Malaysia.

MULTILINGUAL CONVERSATION ASCII TO UNICODE IN INDIC SCRIPT

Dr. Rajwinder Singh[1] and Charanjiv Singh Saroa[2]

[1]Department of Punjabi, Punjabi University, Patiala, India
[2]Department of Computer Engraining, Punjabi University Patiala, India

ABSTRACT

In this paper we discuss the various ASCII based scripts that are made for Indian languages and the problems associated with these types of scripts. Then we will discuss the solution we suggest to overcome these problems in the form of "Multilingual ASCII to Unicode Converter". We also explain the need of regional languages for the development of a person. This paper also contains information of UNICODE and various other issues related to regional languages.

KEYWORDS

Keywords: NLP, Punjabi, Mother Tongue, Gurmukhi, Font Conversion, UNICODE, ASCII, Keyboard Layout.

1. INTRODUCTION

According UNESCO reports About half of the 6,000 or so languages spoken in the world are under threat. Over the past three centuries, languages have died out and disappeared at a dramatic and steadily increasing pace, especially in the Americas and Australia. Today at least 3,000 tongues are endangered, seriously endangered or dying in many parts of the world. [1] A language disappears when its speakers disappear or when they shift to speaking another language. [2] It is also proved from various researches that the primary education of the child should be in the mother tongue of the child instead of in any other language.

In this new world of technology, most of the information is available on internet in e-form. But in regional languages, due to various technical issues like ASCII based fonts, keyboard layouts, lack of awareness of UNICODE, non availability of spell checkers, it is not easy. In regional languages, most of the available fonts are ASCII based instead of UNICODE. We need an intelligent code converter that can change ASCII to UNICODE based scripts.

2. REGIONAL LANGUAGE

A regional language is a language spoken in an area of a state or country, whether it is a small area, a state, a county, or some wider area. Regional languages, as defined by the European Charter for Regional or Minority Languages are traditionally used by part of the population in a state, but which are not official state language dialects, migrant languages or artificially created languages. [3]

David C. Wyld et al. (Eds) : ITCS, SIP, CST, ARIA, NLP - 2017
pp. 83– 94, 2017. © CS & IT-CSCP 2017 DOI : 10.5121/csit.2017.70308

Regional language is mainly spoken in small parts. It changes with the change in culture, religion and economy of the region. In a country, there may be hundreds of regional languages and each language may have further variations. A language is not always limited within the boundaries of a country. One language may be part of more than one country. The eighth schedule of the constitution of India lists 22 scheduled languages. [4] The 22 is for scheduled languages as per the Indian Constitution. It is hard to use computer with all the languages. We need to train computer in each particular language. Computational linguistics is the study of computer system for understanding and generating natural language. [5] Linguistics is the scientific study of language. [6] V.Rajaraman writes in 1998 the government took proactive steps to promote Information Technology by giving incentives such as tax breaks and reduced import duties. [7] Communication infrastructure also improved. The cost of computers came down. All these resulted in a rapid growth of the software services industry with annual growth rate exceeding 30%. We will identify the significant events during each of the above referred periods and explain their impact on the development of IT in India.

2.1 IMPORTANCE OF REGIONAL LANGUAGES

We learn culture, religion and respect from our mother tongue. Regional languages contain lots of sources of understanding community and culture. Regional language/mother tongue gives us:

a) The connections to our roots.
b) Knowledge of our culture.
c) Sense of belonging.
d) Better linguistic skills.
f) Sharper children.
g) A better society.

2.2 NEED OF EDUCATION IN REGIONAL LANGUAGES

Primary education of the child should be in the mother tongue of the child instead of in any other language. Some of the statements are listed below also point toward this.

The following statement from the book titled “The Use of Vernaculars in Education” published by the United Nation‘s Educational Scientific and Cultural Organization (UNESCO) in 1953 is an eye opener. The book presents the essence of international research and wisdom on the issue:

It is axiomatic that the best medium for teaching a child is his mother tongue. Psychologically, it is the system of meaningful signs that in his mind works automatically for expression and understanding. Sociologically, it is means of identification among the members of community to which he belongs. [8]

Children learn best when they are taught in their mother tongue, particularly in the earliest years. Experience in many countries shows that bilingual education, which combines instruction in the mother tongue with teaching in the dominant national language, can open educational and other opportunities. In the Philippines students proficient in the two languages of the bilingual education policy (Tagalog and English) outperformed students who did not speak Tagalog at home. [9]

It is axiomatic that the best medium for teaching a child is his mother tongue. Psychologically, it is the system of meaningful signs that in his mind works automatically for expression and understanding sociologically. it is means of identification among the members of the community to which he belongs. [10] Educationally, he learns more quickly through it than through an unfamiliar linguistic medium.

2.3 CHALLENGES TO USE REGIONAL SCRIPTS IN E-FORM OF INFORMATION

It is very challenging to provide information in e-form using regional languages. Some of the challenges are:

2.3.1 Fonts & Keyboard Layouts:

There are 100s of fonts for every language. In most of the regional languages mainly each font has its own keyboard layout. That result in changing the content of the matter with the change of font and most of the information become useless. Like if Correct sentence is “ I am Going ” in font Arial , with change in font it becomes something like “r kj pjras”. This will never happens in English because all the fonts are created with same keyboard layout and same coding system. But this type of problem is very common in regional languages. A problem with ASCII based fonts for Regional scripts is that there is no standardization of mapping of script characters with keyboard keys. We presently work on 5 Indic Scripts and some of the ASCII based tables of various fonts of these scripts are:

Table 1. For Gurmukhi script of various ASCII Based Fonts

	Remington Style							Phonetic Style							
Decimal Code	Joy	Asees	BJanmeja5A	Prime Ja	GurmukhiLys 010	TERAFONT-Maharaja	Gul-P5Bold	AnmolLipi	LMP_Amrik	Akhar	Amritboli	DRChatrikWeb	GurmukhiIIGS	Chatrik	Sukhmani
65	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A
66	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B
67	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C
68	D	D	D	D	D	D	D	D	D	D	D	D	D	D	D
69	E	E	E	E	E	E	E	E	E	E	E	E	E	E	E
70	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F
71	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G
72	H	H	H	H	H	H	H	H	H	H	H	H	H	H	H
73	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I
74	J	J	J	J	J	J	J	J	J	J	J	J	J	J	J
75	K	K	K	K	K	K	K	K	K	K	K	K	K	K	K
76	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L
77	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M
78	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N
79	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O
80	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P
81	Q	Q	Q	Q	Q	Q	Q	Q	Q	Q	Q	Q	Q	Q	Q
82	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R
83	S	S	S	S	S	S	S	S	S	S	S	S	S	S	S
84	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T
85	U	U	U	U	U	U	U	U	U	U	U	U	U	U	U
86	V	V	V	V	V	V	V	V	V	V	V	V	V	V	V
87	W	W	W	W	W	W	W	W	W	W	W	W	W	W	W
88	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
89	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
90	Z	Z	Z	Z	Z	Z	Z	Z	Z	Z	Z	Z	Z	Z	Z
91	[	[	[	[	[	[	[	[	[	[	[	[	[	[	[
92	\	\	\	\	\	\	\	\	\	\	\	\	\	\	\

93	]	]	]	]	]	]	]	]	]	]	]	]	]	]	]
94	^	^	^	^	^	^	^	^	^	^	^	^	^	^	^
95	_	_	_	_	_	_	_	_	_	_	_	_	_	_	_
96	`	`	a	`	`	`	`	`	`	`	`	`	`	`	`
97		a	a	a	a	a	a	a	a	a	B	a	a	a	a
98	B	b	b	b	b	b	b	b	b	b	b	b	b	b	b
99	c	c	c	c	c	c	c	c	c	c	c	c	c	c	c
100	d	d	d	d	d	d	d	d	d	d	d	d	d	d	d
101	e	e	e	e	e	e	e	e	e	e	e	e	e	e	e
102	f	f	f	f	f	f	f	f	f	f	f	f	f	f	f
103	g	g	g	g	g	g	g	g	g	g	g	g	g	g	g
104	h	h	h	h	h	h	h	h	h	h	h	h	h	h	h
105	i	i	i	i	i	i	i	i	i	i	i	i	i	i	i
106	j	j	j	j	j	j	j	j	j	j	j	j	j	j	j
107	k	k	k	k	k	k	k	k	k	k	k	k	k	k	k
108	l	l	l	l	l	l	l	l	l	l	l	l	l	l	l
109	m	m	m	m	m	m	m	m	m	m	m	m	m	m	m
110	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n
111	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o
112	p	p	p	p	p	p	p	p	p	p	p	p	p	p	p
113	q	q	q	q	q	q	q	q	q	q	q	q	q	q	q
114	r	r	r	r	r	r	r	r	r	r	r	r	r	r	r
115	s	s	s	s	s	s	s	s	s	s	s	s	s	s	s
116	t	t	t	t	t	t	t	t	t	t	t	t	t	t	t
117	u	u	u	u	u	u	u	u	u	u	u	u	u	u	u
118	v	v	v	v	v	v	v	v	v	v	v	v	v	v	v
119	w	w	w	w	w	w	w	w	w	w	w	w	w	w	w
120	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
121	y	y	y	y	y	y	y	y	y	y	y	y	y	y	y
122	z	z	z	z	z	z	z	z	z	z	z	z	z	z	z

Table 2. For Hindi(Devnagri), Gujrati, Malayalam, Tamil script of various ASCII Based Fonts

Hindi (Devnagri)					Gujrati				Malayalam				Tamil			
Decimal code	Aakriti	Chanakya	xdvng	Kundli	Gujrati Saral-1	LMG-Laxmi	Saumil_guj2	Chitra	Manorama	Aruna	Deepa	Gayathri	ELCOT-Kanchi	TM-TTValluvar	Baamini	divya
65	A	A	A	A	A	A	A	A	A	A	A	A	A	அ	A	A
66	B	B	B	B	B	B	B	B	B	B	B	B	B	ஆ	B	B
67	C	C	X	C	C	C	C	C	C	C	C	C	C	இ	C	C
68	D	D	Δ	D	D	D	D	D	D	D	D	D	D	ஈ	D	D
69	E	E	E	E	E	E	E	E	E	E	E	E	E	உ	E	E
70	F	F	Φ	F	F	F	F	F	F	F	F	F	F	ஊ	F	F
71	G	G	Γ	G	G	G	G	G	G	G	G	G	G	எ	G	G
72	H	H	H	H	H	H	H	H	H	H	H	H	H	ஏ	H	H
73	I	I	I	I	I	I	I	I	I	I	I	I	I	ஐ	I	I
74	J	J	ϑ	J	J	J	J	J	J	J	J	J	J	ஒ	J	J

75	K	K	K	K	K	K	K	K	K	K	K	K	K	ஓ	K	K
76	L	L	Λ	L	L	L	L	L	L	L	L	L	L	க	L	L
77	M	M	M	M	M	M	M	M	M	M	M	M	M	ங	M	M
78	N	N	N	N	N	N	N	N	N	N	N	N	N	ச	N	N
79	O	O	O	O	O	O	O	O	O	O	O	O	O	ஞ	O	O
80	P	P	Π	P	P	P	P	P	P	P	P	P	P	ட	P	P
81	Q	Q	Θ	Q	Q	Q	Q	Q	Q	Q	Q	Q	Q	ண	Q	Q
82	R	R	P	R	R	R	R	R	R	R	R	R	R	த	R	R
83	S	S	Σ	S	S	S	S	S	S	S	S	S	S	ந	S	S
84	T	T	T	T	T	T	T	T	T	T	T	T	T	ப	T	T
85	U	U	Y	U	U	U	U	U	U	U	U	U	U	ம	U	U
86	V	V	ς	V	V	□	V	V	V	V	V	V	V	ய	V	V
87	W	W	Ω	W	W	□	W	W	W	W	W	W	W	ர	W	W
88	X	X	Ξ	X	X	X	X	X	X	X	X	X	X	ல	X	X
89	Y	Y	Ψ	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	வ	Y	Y
90	Z	Z	Z	Z	Z	Z	Z	Z	Z	Z	Z	Z	Z	ழ	Z	Z
91	[	[	[	[	[	[	[	[	[	[	[	[	[	ள	[	[
92	\	\	∴	\	\	\	\	\	\	\	\	\	\	ற	\	\
93	]	]	]	]	]	]	]	]	]	]	]	]	]	ன	]	]
94	^	^	⊥	^	^	^	^	^	^	^	^	^	^	ஸ	^	^
95	_	_	_	_	_	_	_	_	_	_	_	_	_	ஜ	_	_
96	`	`	α	`	`	`	`	`	`	`	`	`	`	ஷ	`	`
97	a	a	α	a	a	a	a	a	a	a	a	a	a	ஹ	a	a
98	b	b	β	b	b	b	b	b	b	b	b	b	b	க்ஷ	b	b
99	c	c	χ	c	c	c	c	c	c	c	c	c	c	ஸ்ரீ	c	c
100	d	d	δ	d	d	d	d	d	d	d	d	d	d	க்	d	d
101	e	e	ε	e	e	e	e	e	e	e	e	e	e	ங்	e	e
102	f	f	φ	f	f	f	f	f	f	f	f	f	f	ச்	f	f
103	g	g	γ	g	g	g	g	g	g	g	g	g	g	ஞ்	g	g
104	h	h	η	h	h	h	h	h	h	h	h	h	h	ட்	h	h
105	i	i	ι	i	i	i	i	i	i	i	i	i	i	ண்	i	i
106	j	j	φ	j	j	j	j	j	j	j	j	j	j	த்	j	j
107	k	k	κ	k	k	k	k	k	k	k	k	k	k	ந்	k	k
108	l	l	λ	l	l	l	l	l	l	l	l	l	l	ப்	l	l
109	m	m	μ	m	m	m	m	m	m	m	m	m	m	ம்	m	m
110	n	n	ν	n	n	n	n	n	n	n	n	n	n	ய்	n	n
111	o	o	o	o	o	o	o	o	o	o	o	o	o	ர்	o	o
112	p	p	π	p	p	p	p	p	p	p	p	p	p	ல்	p	p
113	q	q	θ	q	q	q	q	q	q	q	q	q	q	வ்	q	q

114	r	r	ρ	r	r	r	r	r	r	r	r	r	r	ழ்	r	r
115	s	s	σ	s	s	s	s	s	s	s	s	s	s	ள்	s	s
116	t	t	τ	t	t	t	t	t	t	t	t	t	t	ற்	t	t
117	u	u	υ	u	u	u	u	u	u	u	u	u	u	ன்	u	u
118	v	v	ϖ	v	v	v	v	v	v	v	v	v	v	ஸ்	v	v
119	w	w	ω	w	w	w	w	w	w	w	w	w	w	ஜ்	w	w
120	x	x	ξ	x	x	x	x	x	x	x	x	x	x	ஷ்	x	x
121	y	y	ψ	y	y	y	y	y	y	y	y	y	y	ஹ்	y	y
122	z	z	ζ	z	z	z	z	z	z	z	z	z	z	க்ஷ்	z	z

Above tables explain the variation in various fonts with same code. Like in ASCII code 65 in Joy, Asees, Bjanmeja5A, GurmukhiLys 010 represents Bindi ("ਂ"), in Prime Ja it represents adak("ੱ") in AnmolLipi, AmritLipi, Amritboli 65 is used to represents aira("ਅ")in Akhar font it is for ura+hora("ਓ") and in Chatrik font it is for ura("ੳ"),. And when we change the font of the text information of the text is also changed and become useless as shown in following table.

Tabel 3: Various Punjabi Fonts with same key impression

Font Name	Text
AnmolLipi	pMjwbI XUnIvristI, pitAwlw
Akhar	pMjwbI XUnIvristI, pitAwlw
Chatrik	pMjwbI XUnIvristI, pitAwlw
Joy	pMjwbI XUnIvristI, pitAwlw
Asees	pMjwbI XUnIvristI, pitAwlw
Bjanmeja5A	pMjwbI XUnIvristI, pitAwlw
GurmukhiLys 010	pMjwbI XUnIvristI, pitAwlw

Computer understand data in the form of 0 and 1. According to that ASCII codes are created for Roman script. But these codes are not compatible with other Indic scripts. Some of the variations are;

Table 4: Script and code Compare

Dec Code	Joy	Asees	BJanmeja5A	Prime Ja	GurmukhiLys 010	AnmolLipi	AmritLipi	Akhar	Amritboli	DRChatrikWeb
65	A	A	A	A	A	A	A	A	A	A
69	E	E	E	E	E	E	E	E	E	E
70	F	F	F	F	F	F	F	F	F	F
87	W	W	W	W	W	W	W	W	W	W

Table 5: Script and code Compare

Hindi (Devnagri)				Gujrati			Malayalam		Tamil	
	Aakriti	Chanakya	xdvng	Gujrati Saral-1	LMG-Laxmi	Saumil_guj2	Manorama	Aruna	ELCOT-Kanchi	TM-TTValluvar
65	A	A	A	A	A	A	A	A	A	அ
69	E	E	E	E	E	E	E	E	E	உ
77	M	M	M	M	M	M	M	M	M	ஙு
87	W	W	Ω	W	□	W	W	W	W	ரு

2.3.2 Non Unicode Fonts:

Mostly the data available in regional languages are in ASCII based Indic scripts. we want to display that data on website we have to upload the required font for that data and user firstly have to download and install that font, Only then the user can view that information. There are thousands of fonts used for create data in Indian Scripts. Only in Punjabi (Gurmukhi) alone has more than 225 popular fonts which are still in use to publishing books, magazine, news paper etc. Publishers are still working with ASCII coding based fonts. They have not used Unicode based fonts to following reasons:

- People resist to change, due to Unicode typing issues. [11]
- Lack of awareness of Unicode standard.
- Little support of Unicode system in publishing software that they are using.
- Less availability of Unicode fonts to represent text in different style and designs

It is always better to display information in Unicode based fonts while displaying the information on website. The information presently available to us is mainly in ASCII based fonts. So we convert that ASCII based information into Unicode based fonts so that it can be available on internet. The information displayed in Unicode can be seen on any computer without installing font. Other advantage of Unicode based fonts is that it is searchable on search engines like Google, Yahoo, Ask, Bing etc.

2.3.3 Some special Symbols:

Some text can contain some unique type of symbols that are not available in ASCII codes and even not in UNICODE system.

2.3.4 Typing problem:

By default Mainly each computer contain English (roman) keyboard. And most of the user are not aware of UNICODE based system and fonts. Without the knowledge of UNICODE based fonts user cannot type in Unicode.

2.3.5 Spell check:

All the available spellcheckers mainly work with English language.

3. ASCII/UNICODE

ASCII abbreviated from American Standard Code for Information Interchange, is a character encoding standard (the Internet Assigned Numbers Authority (IANA) prefers the name US-ASCII). ASCII codes represent text in computers, telecommunications equipment, and other devices. Most modern character-encoding schemes are based on ASCII, although they support many additional characters ASCII coding system can code only 128 characters [0-127] in ASCII 7bit and 256 characters (0-255) in ASCII 8bit. [12] These are allocated to characters of roman script, special symbols and to alphanumeric characters. No place for other scripts in ASCII .On the other hand the Unicode coding system provide much more range of codes that help to give unique code to various scripts. The Unicode Standard, the latest version of Unicode contains more than 110,000 characters covering 100 scripts. [13]

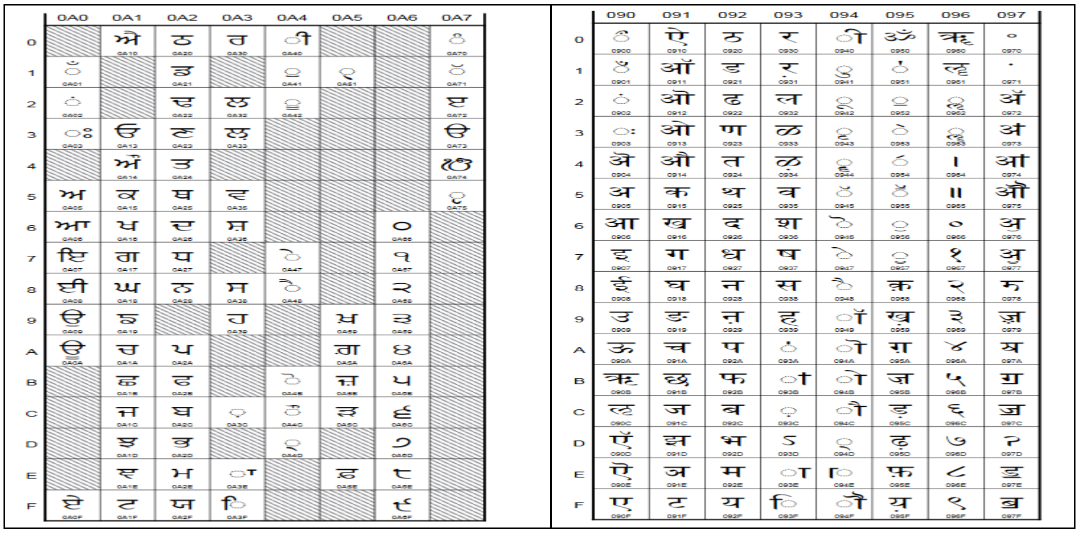
First version of Unicode (1.0.0) is released on October 1991 that contain total 7,161 of 24 scripts some of the scripts are Arabic, Armenian, Bengali, Bopomofo, Cyrillic, Devanagari, Gujarati, Gurmukhi, Hangul, Hebrew, Hiragana, Kannada, Katakana, Lao, Latin, Malayalam, Oriya, Tamil, Telugu, Thai, Tibetan etc. Version 7.0 is released in June 2014 that contain 113,021 characters of 123 scripts new scripts that are included are Bassa Vah, Caucasian Albanian, Duployan, Elbasan, Grantha, Khojki, Khudawadi, Linear A, Mahajani, Manichaean, Mende Kikakui, Modi, Mro, Nabataean, Old North Arabian, Old Permic, Pahawh Hmong, Palmyrene, Pau Cin Hau, Psalter Pahlavi, Siddham, Tirhuta, Warang Citi, and Dingbats. [14] The latest version was released in June 2016 that contain 128,237 characters and 135 scripts.

3.1 HOW UNICODE WORKS

As in ASCII code each roman character get its unique code so on every computer it will display as the user type it. when user write anything he/she did not worry about choosing which font to write in. User knows that other users will be able to read this article without any problems. This is not happened with regional languages ASCII based fonts. User need to provides font with the information so that other user can read it. But in Unicode each character gets its own individual code. But when user use Unicode font, users would not have to decide which font to use. ASCII uses the limited set of codes to store the character information whereas Unicode gives a uniqe code to every character which it recognises. Thats why ASCII may change its characters when the font is changed.

3.2 ADVANTAGES OF UNICODE

1) Allows for multilingual text in single document without bothering about fonts.
2) Support of Unicode is available on all modern technologies which extend life and scope of application.
3) Full internet support for Unicode system so information written using Unicode based font is easily viewed on internet.
4) Text in any language can be exchanged worldwide.



	0A0	0A1	0A2	0A3	0A4	0A5	0A6	0A7
0		ਐ 0A10	ਠ 0A20	ਰ 0A30	ੀ 0A40			ੰ 0A70
1	ਁ 0A01		ਡ 0A21		ੁ 0A41	ੑ 0A51		ੱ 0A71
2	ਂ 0A02		ਢ 0A22	ਲ 0A32	ੂ 0A42			ੲ 0A72
3	ਃ 0A03	ਓ 0A13	ਣ 0A23	ਲ਼ 0A33				ੳ 0A73
4		ਔ 0A14	ਤ 0A24					ੴ 0A74
5	ਅ 0A05	ਕ 0A15	ਥ 0A25	ਵ 0A35				ੵ 0A75
6	ਆ 0A06	ਖ 0A16	ਦ 0A26	ਸ਼ 0A36			੦ 0A66	
7	ਇ 0A07	ਗ 0A17	ਧ 0A27		ੇ 0A47		੧ 0A67	
8	ਈ 0A08	ਘ 0A18	ਨ 0A28	ਸ 0A38	ੈ 0A48		੨ 0A68	
9	ਉ 0A09	ਙ 0A19		ਹ 0A39		ਖ਼ 0A59	੩ 0A69	
A	ਊ 0A0A	ਚ 0A1A	ਪ 0A2A			ਗ਼ 0A5A	੪ 0A6A	
B		ਛ 0A1B	ਫ 0A2B		ੋ 0A4B	ਜ਼ 0A5B	੫ 0A6B	
C		ਜ 0A1C	ਬ 0A2C	਼ 0A3C	ੌ 0A4C	ੜ 0A5C	੬ 0A6C	
D		ਝ 0A1D	ਭ 0A2D		੍ 0A4D		੭ 0A6D	
E		ਞ 0A1E	ਮ 0A2E	ਾ 0A3E		ਫ਼ 0A5E	੮ 0A6E	
F	ਏ 0A0F	ਟ 0A1F	ਯ 0A2F	ਿ 0A3F			੯ 0A6F	

	090	091	092	093	094	095	096	097
0	ऀ 0900	ऐ 0910	ठ 0920	र 0930	ी 0940	ॐ 0950	ॠ 0960	॰ 0970
1	ँ 0901	ऑ 0911	ड 0921	ऱ 0931	ु 0941	॑ 0951	ॡ 0961	ॱ 0971
2	ं 0902	ऒ 0912	ढ 0922	ल 0932	ू 0942	॒ 0952	ॢ 0962	ॲ 0972
3	ः 0903	ओ 0913	ण 0923	ळ 0933	ृ 0943	॓ 0953	ॣ 0963	ॳ 0973
4	ऄ 0904	औ 0914	त 0924	ऴ 0934	ॄ 0944	॔ 0954	। 0964	ॴ 0974
5	अ 0905	क 0915	थ 0925	व 0935	ॅ 0945	ॕ 0955	॥ 0965	ॵ 0975
6	आ 0906	ख 0916	द 0926	श 0936	ॆ 0946	ॖ 0956	० 0966	ॶ 0976
7	इ 0907	ग 0917	ध 0927	ष 0937	े 0947	ॗ 0957	१ 0967	ॷ 0977
8	ई 0908	घ 0918	न 0928	स 0938	ै 0948	क़ 0958	२ 0968	ॸ 0978
9	उ 0909	ङ 0919	ऩ 0929	ह 0939	ॉ 0949	ख़ 0959	३ 0969	ॹ 0979
A	ऊ 090A	च 091A	प 092A	ऺ 093A	ॊ 094A	ग़ 095A	४ 096A	ॺ 097A
B	ऋ 090B	छ 091B	फ 092B	ऻ 093B	ो 094B	ज़ 095B	५ 096B	ॻ 097B
C	ऌ 090C	ज 091C	ब 092C	़ 093C	ौ 094C	ड़ 095C	६ 096C	ॼ 097C
D	ऍ 090D	झ 091D	भ 092D	ऽ 093D	् 094D	ढ़ 095D	७ 096D	ॽ 097D
E	ऎ 090E	ञ 091E	म 092E	ा 093E	ॎ 094E	फ़ 095E	८ 096E	ॾ 097E
F	ए 090F	ट 091F	य 092F	ि 093F	ॏ 094F	य़ 095F	९ 096F	ॿ 097F

Figure 1: Unicode Character Code to Rendered Glyphs for Gurmukhi and Devnagri Script

3.3 MULTILINGUAL ASCII TO UNICODE SCRIPT CONVERTER:

(www.gurmukhifontconverter.com)

To overcome the problem of ASCII based fonts we create font converter software that can convert many fonts of Indic scripts into other scripts without changing its original meaning and content. Some of the software that are available before it are not very accurate and it convert text without formatting, if a document contains some text in other scripts it converts the whole document to target font. But the font converter that we have created converts the only required code to target code without effecting the text of any other scripts. This software is then converted into an online website (www.gurmukhifontconverter.com) So that everyone can use it. Now more than 30,000 users are using it to convert their text from one font to another. To create this converter we had done manual mapping of around 168 fonts that results around 8,000 page document. This document has helped us to create this font converter. This converter also converts ASCII based fonts to Unicode based fonts.

i. Algorithm:

ii. Input text in the converter.

iii. Identify the script.

iv. By matching words from corpus

v. By matching tries from corpus

vi. Select maximum frequency of word and tri in each language script.

vii. Source script is automatically identified from step 2.

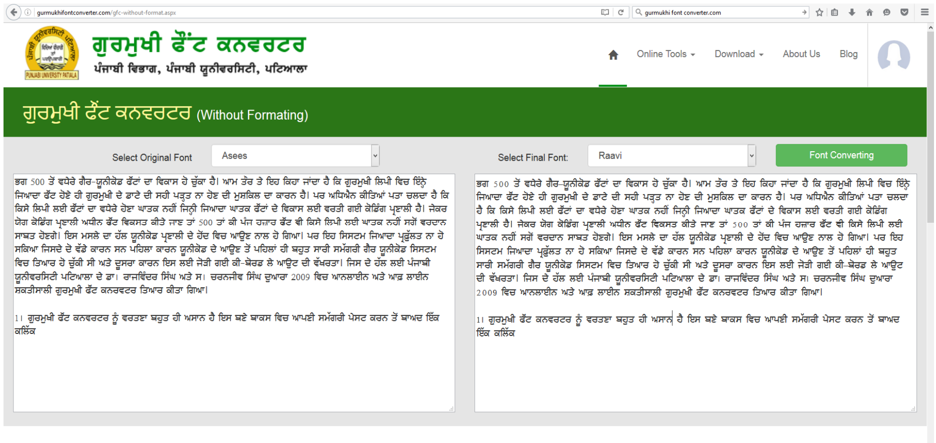
viii. Converter automatically identifies the target Unicode system.

ix. Convert the ASCII characters one by one into Unicode system.

x. Repeat stem 2 to 5 for each word and sentence.

Our developed system that will convert multilingual (Devnagri, Gurmukhi , Malyalam, gujrati, Telgu, Roman) data which automatically identify the script and then system automatically convert it into its UNICODE based script. To identify the script we created a database of around

200 Thousands of words of each script. Further tries are created comprising 400 Thousands tries of each script. “Multilingual Conversation ASCII to Unicode in Indic Script” fulfils following requirements:


gurmukhifontconverter.com/gfc-without-format.aspx
gurmukhi font converter.com
ਗੁਰਮੁਖੀ ਫੌਂਟ ਕਨਵਰਟਰ
ਪੰਜਾਬੀ ਵਿਭਾਗ, ਪੰਜਾਬੀ ਯੂਨੀਵਰਸਿਟੀ, ਪਟਿਆਲਾ
PUNJABI UNIVERSITY PATIALA
Online Tools
Download
About Us
Blog
ਗੁਰਮੁਖੀ ਫੌਂਟ ਕਨਵਰਟਰ (Without Formating)
Font Converting

Select Original Font: Asees

ਭਗ 500 ਤੋਂ ਵਧੇਰੇ ਗੈਰ-ਯੂਨੀਕੋਡ ਫੌਂਟਾਂ ਦਾ ਵਿਕਾਸ ਹੋ ਚੁੱਕਾ ਹੈ। ਆਮ ਤੌਰ ਤੇ ਇਹ ਕਿਹਾ ਜਾਂਦਾ ਹੈ ਕਿ ਗੁਰਮੁਖੀ ਲਿਪੀ ਵਿਚ ਇੰਨ੍ਹੇ ਜਿਆਦਾ ਫੌਂਟ ਹੋਣੇ ਹੀ ਗੁਰਮੁਖੀ ਦੇ ਡਾਟੇ ਦੀ ਸਹੀ ਪੜ੍ਹਤ ਨਾ ਹੋਣ ਦੀ ਮੁਸ਼ਕਿਲ ਦਾ ਕਾਰਨ ਹੈ। ਪਰ ਅਧਿਐਨ ਕੀਤਿਆਂ ਪਤਾ ਚਲਦਾ ਹੈ ਕਿ ਕਿਸੇ ਲਿਪੀ ਲਈ ਫੌਂਟਾਂ ਦਾ ਵਧੇਰੇ ਹੋਣਾ ਘਾਤਕ ਨਹੀਂ ਜਿਨ੍ਹੀ ਜਿਆਦਾ ਘਾਤਕ ਫੌਂਟਾਂ ਦੇ ਵਿਕਾਸ ਲਈ ਵਰਤੀ ਗਈ ਕੋਡਿੰਗ ਪ੍ਰਣਾਲੀ ਹੈ। ਜੇਕਰ ਯੋਗ ਕੋਡਿੰਗ ਪ੍ਰਣਾਲੀ ਅਧੀਨ ਫੌਂਟ ਵਿਕਸਤ ਕੀਤੇ ਜਾਣ ਤਾਂ 500 ਤਾਂ ਕੀ ਪੰਜ ਹਜ਼ਾਰ ਫੌਂਟ ਵੀ ਕਿਸੇ ਲਿਪੀ ਲਈ ਘਾਤਕ ਨਹੀਂ ਸਗੋਂ ਵਰਦਾਨ ਸਾਬਤ ਹੋਣਗੇ। ਇਸ ਮਸਲੇ ਦਾ ਹੱਲ ਯੂਨੀਕੋਡ ਪ੍ਰਣਾਲੀ ਦੇ ਹੋਂਦ ਵਿਚ ਆਉਣ ਨਾਲ ਹੋ ਗਿਆ। ਪਰ ਇਹ ਸਿਸਟਮ ਜਿਆਦਾ ਪ੍ਰਫੁੱਲਤ ਨਾ ਹੋ ਸਕਿਆ ਜਿਸਦੇ ਦੋ ਵੱਡੇ ਕਾਰਨ ਸਨ ਪਹਿਲਾ ਕਾਰਨ ਯੂਨੀਕੋਡ ਦੇ ਆਉਣ ਤੋਂ ਪਹਿਲਾਂ ਹੀ ਬਹੁਤ ਸਾਰੀ ਸਮੱਗਰੀ ਗੈਰ ਯੂਨੀਕੋਡ ਸਿਸਟਮ ਵਿਚ ਤਿਆਰ ਹੋ ਚੁੱਕੀ ਸੀ ਅਤੇ ਦੂਸਰਾ ਕਾਰਨ ਇਸ ਲਈ ਜੋੜੀ ਗਈ ਕੀ-ਬੋਰਡ ਲੇ ਆਉਟ ਦੀ ਵੱਖਰਤਾ। ਜਿਸ ਦੇ ਹੱਲ ਲਈ ਪੰਜਾਬੀ ਯੂਨੀਵਰਸਿਟੀ ਪਟਿਆਲਾ ਦੇ ਡਾ। ਰਾਜਵਿੰਦਰ ਸਿੰਘ ਅਤੇ ਸ। ਚਰਨਜੀਵ ਸਿੰਘ ਦੁਆਰਾ 2009 ਵਿਚ ਆਨਲਾਈਨ ਅਤੇ ਆਫ਼ ਲਾਈਨ ਸ਼ਕਤੀਸਾਲੀ ਗੁਰਮੁਖੀ ਫੌਂਟ ਕਨਰਵਟਰ ਤਿਆਰ ਕੀਤਾ ਗਿਆ।

1। ਗੁਰਮੁਖੀ ਫੌਂਟ ਕਨਵਰਟਰ ਨੂੰ ਵਰਤਣਾ ਬਹੁਤ ਹੀ ਅਸਾਨ ਹੈ ਇਸ ਬਣੇ ਬਾਕਸ ਵਿਚ ਆਪਣੀ ਸਮੱਗਰੀ ਪੇਸਟ ਕਰਨ ਤੋਂ ਬਾਅਦ ਇੱਕ ਕਲਿੱਕ

Select Final Font: Raavi

ਭਗ 500 ਤੋਂ ਵਧੇਰੇ ਗੈਰ-ਯੂਨੀਕੋਡ ਫੌਂਟਾਂ ਦਾ ਵਿਕਾਸ ਹੋ ਚੁੱਕਾ ਹੈ। ਆਮ ਤੌਰ ਤੇ ਇਹ ਕਿਹਾ ਜਾਂਦਾ ਹੈ ਕਿ ਗੁਰਮੁਖੀ ਲਿਪੀ ਵਿਚ ਇੰਨ੍ਹੇ ਜਿਆਦਾ ਫੌਂਟ ਹੋਣੇ ਹੀ ਗੁਰਮੁਖੀ ਦੇ ਡਾਟੇ ਦੀ ਸਹੀ ਪੜ੍ਹਤ ਨਾ ਹੋਣ ਦੀ ਮੁਸ਼ਕਿਲ ਦਾ ਕਾਰਨ ਹੈ। ਪਰ ਅਧਿਐਨ ਕੀਤਿਆਂ ਪਤਾ ਚਲਦਾ ਹੈ ਕਿ ਕਿਸੇ ਲਿਪੀ ਲਈ ਫੌਂਟਾਂ ਦਾ ਵਧੇਰੇ ਹੋਣਾ ਘਾਤਕ ਨਹੀਂ ਜਿਨ੍ਹੀ ਜਿਆਦਾ ਘਾਤਕ ਫੌਂਟਾਂ ਦੇ ਵਿਕਾਸ ਲਈ ਵਰਤੀ ਗਈ ਕੋਡਿੰਗ ਪ੍ਰਣਾਲੀ ਹੈ। ਜੇਕਰ ਯੋਗ ਕੋਡਿੰਗ ਪ੍ਰਣਾਲੀ ਅਧੀਨ ਫੌਂਟ ਵਿਕਸਤ ਕੀਤੇ ਜਾਣ ਤਾਂ 500 ਤਾਂ ਕੀ ਪੰਜ ਹਜ਼ਾਰ ਫੌਂਟ ਵੀ ਕਿਸੇ ਲਿਪੀ ਲਈ ਘਾਤਕ ਨਹੀਂ ਸਗੋਂ ਵਰਦਾਨ ਸਾਬਤ ਹੋਣਗੇ। ਇਸ ਮਸਲੇ ਦਾ ਹੱਲ ਯੂਨੀਕੋਡ ਪ੍ਰਣਾਲੀ ਦੇ ਹੋਂਦ ਵਿਚ ਆਉਣ ਨਾਲ ਹੋ ਗਿਆ। ਪਰ ਇਹ ਸਿਸਟਮ ਜਿਆਦਾ ਪ੍ਰਫੁੱਲਤ ਨਾ ਹੋ ਸਕਿਆ ਜਿਸਦੇ ਦੋ ਵੱਡੇ ਕਾਰਨ ਸਨ ਪਹਿਲਾ ਕਾਰਨ ਯੂਨੀਕੋਡ ਦੇ ਆਉਣ ਤੋਂ ਪਹਿਲਾਂ ਹੀ ਬਹੁਤ ਸਾਰੀ ਸਮੱਗਰੀ ਗੈਰ ਯੂਨੀਕੋਡ ਸਿਸਟਮ ਵਿਚ ਤਿਆਰ ਹੋ ਚੁੱਕੀ ਸੀ ਅਤੇ ਦੂਸਰਾ ਕਾਰਨ ਇਸ ਲਈ ਜੋੜੀ ਗਈ ਕੀ-ਬੋਰਡ ਲੇ ਆਉਟ ਦੀ ਵੱਖਰਤਾ। ਜਿਸ ਦੇ ਹੱਲ ਲਈ ਪੰਜਾਬੀ ਯੂਨੀਵਰਸਿਟੀ ਪਟਿਆਲਾ ਦੇ ਡਾ। ਰਾਜਵਿੰਦਰ ਸਿੰਘ ਅਤੇ ਸ। ਚਰਨਜੀਵ ਸਿੰਘ ਦੁਆਰਾ 2009 ਵਿਚ ਆਨਲਾਈਨ ਅਤੇ ਆਫ਼ ਲਾਈਨ ਸ਼ਕਤੀਸਾਲੀ ਗੁਰਮੁਖੀ ਫੌਂਟ ਕਨਰਵਟਰ ਤਿਆਰ ਕੀਤਾ ਗਿਆ।

1। ਗੁਰਮੁਖੀ ਫੌਂਟ ਕਨਵਰਟਰ ਨੂੰ ਵਰਤਣਾ ਬਹੁਤ ਹੀ ਅਸਾਨ ਹੈ ਇਸ ਬਣੇ ਬਾਕਸ ਵਿਚ ਆਪਣੀ ਸਮੱਗਰੀ ਪੇਸਟ ਕਰਨ ਤੋਂ ਬਾਅਦ ਇੱਕ ਕਲਿੱਕ

Figure 2. http:/gurmukhifontconverter.com/gfc-without-format.aspx

IDENTIFICATION

Identification is done on the basis of words. We need to identify the language from minimum words which are typed/paste by the user.

Convert To Unicode

The system efficiently converts all the ASCII based scripts in to Unicode system.

Retain Formatting

“Multilingual Conversation ASCII to Unicode in Indic Script” retains the original formatting of the text. It converts the text without changing its formatting.

4. RESULTS

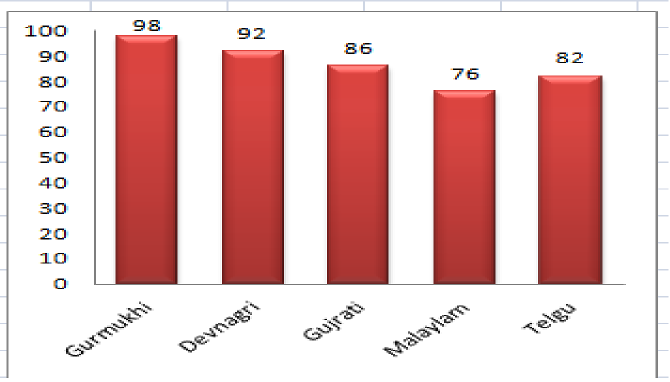


Figure 3. Accuracy of various scripts

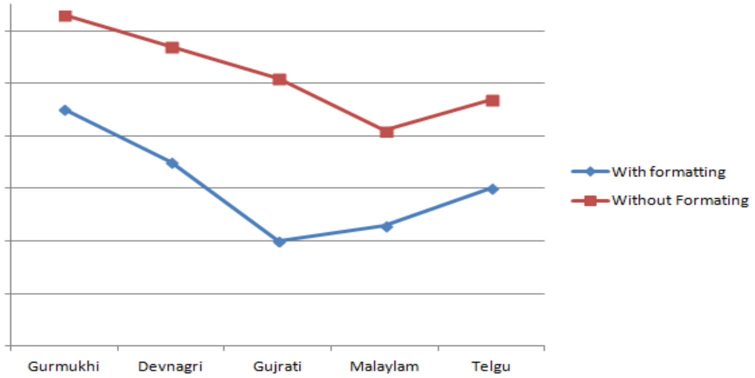


Figure 4. Speed with formatting and without formatting

REFERENCES

[1] UNESCO Bangkok, (2008) Improving the Quality of Mother Tongue-based Literacy and Learning: Case Studies from Asia, Africa and South America, pp7.

[2] Marcia Langton and Zane Marhea, (2003) Traditional Lifestyles and Biodiversity use Regional Report: Australia, Asia and The middle east, PP21

[3] Strasbourg, (1992) EUROPEAN CHARTER FOR REGIONAL OR MINORITY LANGUAGES, ETS 148–Charter for Minority Languages, 5.XI, pp14-15.

[4] Government of India, (2007), The Constitution of India, Govt. of India Ministry of law and justice (As modified up to the 1st December 2007), pp358-360.

[5] Ralph Grishman, (1986) Computational linguistics an introduction, Cambridge University Press, pp24.

[6] John Lyons, (1981) Language and Linguistics an Introduction, Cambridge University Press, pp33

[7] V. Rajaraman, (2012) History of computing in India, 1995-2010, Supercomputer education and research centre Indian institute of science, Bangalore 560012, pp14.

[8] Unesco Education Position Paper, (2003) Education in a multilingual World, Published by the United Nations Education, Scientific and Cultural Organization, pp13-14

[9] Human Development Report, (2004) Cultural Liberty in Today‟s Diverse World, Carfax Publishing, Taylor and Francis Ltd. Customer Services Department, pp77

[10] UNESCO. 1953. The Use of Vernacular Languages in Education. Monographs on Fundamental Education, No. 8. Paris, pp54

[11] Gurpreet Singh LEHAL1 Tejinder Singh SAINI, (2012) An Omni-font Gurmukhi to Shahmukhi Transliteration System, Proceedings of COLING, Mumbai PP314

[12] https://en.wikipedia.org/wiki/ASCII

[13] Ms.M.Kavitha1 , Ms.S.Kawsalya, (2013) Secured Crypto-Stegano Communication through Unicode and Triple DES, International Journal of Innovative Research in Computer and Communication Engineering Vol. 1, Issue 2, PP396

[14] The Unicode Consortium, (2015), The Unicode Standard Version 8.0 – Core Specification, Published in Mountain View, CA, pp913.

AUTHORS

Dr. Rajwinder Singh is Assistant Professor in Department of Punjabi, Editor and Coordinator Punjabipedia (www. punjabipedia.org) world famous project in Punjabi University Patiala, where he has been since 2009. He also currently working various projects related to technical development for Punjabi language, literature and culture. He has completed his Ph.D. (Linguistics) at Punjabi University, Patiala (2008) and his undergraduate also complete this University. His research interests Computational Linguistics, NLP, Grammar, Punjabi Linguistics and area of programming languages.

Er. Charanjiv Singh Saroa is Assistant Professor in Computer department of Engraining, Punjabi University Patiala. He is also working co-coordinator Punjabipedia world famous project in Punjabi University Patiala. His area of interest is NLP.

HOW TO MATCH BILINGUAL TWEETS?

Karima Abidi[1] Kamel Smaili[2]

[1] Ecole Supérieure d'Informatique (ESI), Algiers,Algeria
[2] Campus Scientifique LORIA , Nancy, France

ABSTRACT

In this paper, we propose a method that aligns comparable bilingual tweets which, not only takes into account the specificity of a Tweet, but treats also proper names, dates and numbers in two different languages. This permits to retrieve more relevant target tweets. The process of matching proper names between Arabic and English is a difficult task, because these two languages use different scripts. For that, we used an approach which projects the sounds of an English proper name into Arabic and aligns it with the most appropriate proper name. We evaluated the method with a classical measure and compared it to the one we developed. The experiments have been achieved on two parallel corpora and shows that our measure outperforms the baseline by 5.6% at R@1 recall.

KEYWORDS

Comparability measure ; Arabic stemming, Proper names; Soundex, Twitter

1. INTRODUCTION

The parallel corpora are extremely valuable resources for many applications in natural language processing, in particular for machine translation which needs massive corpus to train statistical models. However, this type of data are not always available and especially for certain pairs of languages. An attractive option is to collect data from the web and the social networks for which, nowadays data are abundant. However, this task is not easy to handle because the collected documents have to be aligned in accordance to their topic. When the corpora are aligned, they are considered as comparable. These corpora consist of a set of documents expressed in several languages which are not parallel in the strict sense, but deal with analogous subjects. These last decades this issue has grown considerably [10] [6] [13][16]. The community of Cross-Lingual Information Retrieval contributed significantly to propose different solutions to this issue [12], [7][14].

These last couple of years, a particular attention has been given to harvest data from social networks and particularly from Twitter. This is due to the fact that, so many people through the world adopted this social network to express their opinions. Consequently, it would be useful to investigate this media, in order to collect cross-lingual posts and align them in terms of topics. This will lead to get comparable corpora of Tweets. These documents could be then, used to extract parallel fragments or phrases. In order to have relevant parallel fragments, we need relevant comparable Tweets, to achieve that, we should propose an efficient measure to estimate the comparability. In this work, we will show that the classical dictionary-based measure [8] cannot be used directly for this task and especially for Arab Tweets. Specific knowledge related

David C. Wyld et al. (Eds) : ITCS, SIP, CST, ARIA, NLP - 2017
pp. 95– 106, 2017. © CS & IT-CSCP 2017 DOI : 10.5121/csit.2017.70309

to Arabic will be taken into account in order to tackle, not only, the issue of the specificity of Arabic, but also this free way of writing the posts.

This paper is organized as follows: In section 2, we present related works concerning the extraction of comparable corpora. Then, we describe the corpus, we collect, and we give details about the preprocessing steps in sections 3 and 4. In section 5, we present the process we use for matching bilingual Tweets. Several experiments and results are described in section 6. Finally, we conclude and present some future works.

2. RELATED WORKS

The construction of comparable corpora is performed using similarity measures. These measures can be based on three different approaches: vocabulary overlapping, vector space and Machine translation. Among existing work to align the comparable corpus, we can mention the following: Li and Gaussier in [8] defined the degree of comparability between two corpora as the expectation of finding, for each word of the source corpus, its translation in the target one. They use this definition to propose a measure which estimates the comparability of two parallel corpora to which noises have been added. They showed that the comparability degree decreased proportionally with the added noises. A similar approach proposed by Etchegoyhen et al. in [5] termed STACC, is based on expanded lexical sets and Jaccard similarity coefficient. The idea is to get rid of a manual bilingual dictionary. The bilingual dictionary is built on a large parallel corpus by using Giza ++[11]. Since, it is independent from languages, the approach has been evaluated on a large dataset of ten languages. Zhu et al. in [16] utilized a bilingual LDA model to match similar documents. They proved that this approach can obtain similar documents with consistent topics. Huang et al. in [6] describe a method based on techniques inspired from Cross Lingual Information Retrieval. With the translation of the keywords of the source documents, they retrieve the target documents which contain these translated words. Then, the mapping between source and target documents is achieved in accordance to a similarity value. A method based on word embedding has been proposed by Vulic et al in [14]. The model has the possibility to learn bilingual word embeddings from already comparable corpora. The crucial idea in this work is the fact that the method allows to share the cross-lingual embedding space.

Works on comparable corpus containing Arabic are not as popular as those used for English or French, we can mention those proposed in [10],[1], [12]. In this last work, different comparability measures based on bilingual dictionaries or on numerical methods such as Latent Semantic Indexing (LSI) have been proposed.

3. EXPERIMENTAL MATERIAL

As presented before, we propose to identify comparable corpora by extracting them from Twitter. In the following, we will be interested by two languages: Arabic and English. In order to identify comparable Tweets, we decided to set the topic which will be used to crawl the Tweets: *Syria's war*. Our objective is to crawl all the Tweets related to this subject and then to align Arabic and English at the Tweet level. For this purpose, we selected the 7th-top English Hashtags concerning the war in Syria (Table 1). The same process has been done on Arabic (Table 2).

Due to the particularity of the free way to write Arabic in Twitter, we used different Hashtags such as: #سوريا #سوريه , both of them correspond to the word *Syria*. One ending with *Ta* (ة) and the other with *alif* (ا).

Table 1. Number of English tweets collected for each Hashtag.

English Hashtag	N_{Twts}
#SyrianRefugees	10895
#refugeescrisis	2856
#Syrianarmy	3211
#freesyrianarmy	3119
#SyriaCrisis	6260
#syria	17000
#syrian	17000
Total	**57771**

Table 2. Number of Arabic tweets collected for each Hashtag.

English Hashtag		N_{Twts}
#اطفال-سورية	Children of Syria	1599
#الثورة-السورية	Syrian revolution	10092
#اللاجئين-السوريين	Syrian refugees	4000
#سوريا	Syria	17000
#سورية	Syria	17000
#الجيش-العربي-السوري	Syrian Arab army	916
#الجيش-السوري-الحر	Free Syrian army	4318
Total		**59452**

Table 1 and Table 2 show that the global number of crawled Tweets is approximatively equivalent. But the number of Tweets concerning the *Syrian Refugees* is not the same. It is twice much more in English than in Arabic. This could be explained by the fact that this topic has been very popular in the West, since the corresponding countries were directly concerned by the problem. While, the number of Tweets concerning *Syria* is much more important in Arabic than in English, since the Arab world is very involved in the Syrian issue.

4. PREPROCESSING TWEETS

In natural language and especially for processing Tweets, one needs to rewrite some words, to clean some of them, to homogeneous the way of writing, to transform digits, proper names, cities and so on. This step is referred as language preprocessing. In the following, we pre-process both Arabic and English by taking into account the linguistic specificity of each of them. This step is very crucial since our objective is to identify comparable Tweets. More the treatments of homogenization of Tweets in both languages are precise, and more the process of identifying comparable Tweets is relevant. Figure 1 illustrates an example of the differences between an English and its equivalent Tweet in Arabic. The fragments on red and green correspond to respectively the way to write the date and the number in Arabic.

English tweet	Syria issues decree no 63 the general parliamentary elections will be on Wednesday april th 13 2016
Arabic tweet	اصدر بشار الاسد المرسوم رقم ٦٣ القاضي بتحديد يوم الاربعاء ١٣ نيسان ٢٠١٦ موعد لانتخاب الاعضاء

Figure 1. Example of comparable Tweets.

4.1. Preprocessing English Tweets

In the following, we present the main treatments we achieve on the English Tweet corpus. Since Twitter is used by hundred million of users and because a Tweet is limited to 140 characters, people take some freedom in writing their posts, for instance by shortening the words. To handle this issue, we use a SMS dictionary [1] which contains abbreviations, acronyms and their literal corresponding text. We used this dictionary to replace abbreviations by their literal forms. For example, *ppl* will be replaced by *people*.

Sometimes, in an English Tweet, we can find some references to foreign languages. This could be a serious problem for the further linguistic treatments. Based on a list of stop words of few languages, we discard all the fragments which contain at least one of these stops words.

Contrary to Arabic Tweets corpus, the English one contains many Hashtags embedded in the Tweet itself. Sometimes they are in the middle of a post, consequently we cannot just remove them, since they are used as any word. In some Tweets, a Hashtag might be composed of several Hashtags *#SyrianArabArmy*, *#SyrianCivilians*, ...etc which should be split. In the case where words are separated by capital letters or special characters, it is easy to determine the border of words composing the compound Hashtag. But, when the Hashtag is completely written in lowercase, we need to perform differently. To do so, we decided to use a dictionary sorted alphabetically and by the word's length. We seek a word in the compound Hashtag by looking for it in this dictionary.

Because there are several ways to write a date in English, all of them will be homogenized such as: *DD/MM/Year*. In Table 3, we give some examples before and after the rewriting process.

In social network, a letter of some words could be duplicated several times to express an emotion or a sentiment, phenomenon known as an elongation such as in: *woahhh*. These kind of words are transformed by removing the duplicated letters: *woah*

Table 3. Examples of rewritten dates.

Before	After
April 13,2016	13/4/2016
february 1,2016	1/2/2016
22 feb 2016	22/2/2016
2.2.2016	2/2/2016

4.2. Preprocessing Arabic Tweets

Arabic is very different from Indo-European languages, that is why specific treatments have to be achieved, in order to make Tweets ready for the further processings.

For reasons of freedom writing in Twitter, users replace the letter ة by ه only at the end of words. This could be very surprising since these two letters have different linguistic roles in Arabic, but graphically they are similar, except that the first one has two dots above. For convenience, some people use indifferent one or the other. This is why we homogenized the script in all the tweets. For almost the same reasons, we replaced all the forms of symbol *Alif* with *Hamza* such as in: آ, أ, إ with a simple *Alif* ا . Concerning the diacritics, we removed them since people can read without short vowels.

In Arabic, two sets of digits are used for writing numbers, the first set is the one used around the world, known by Arabic digits and the second is the Indo-Arab digits which are much more used

[1] www.illumasolutions.com/omg-plz-lol-idk-idc-btw-brb-jk.htm

in Middle-East: ٠١٢.. ٩. For this purpose, we decided to keep only one numeral notation for numbers by using the English coma for decimal number, for example ٥,١٣will be rewritten as 5.13.

Concerning the dates, in the Arab world, three types of calendars could be used: Assyrian, Hijri and Gregorian. The first and the second one are much more used in the East than in the West of the Arab world. For example: *January* could be written: كانون الثاني , جانفي or يناير depending on the Arab region. That is why each date, whatever its form, is rewritten in accordance to the following pattern *DD/MM/Year* (see examples in Table 4).

In social networks, sometimes users stretch words to accentuate their opinion or just write the words such as they pronounce them, then it is necessary to normalize the way of writing these words. The stretched or duplicated letters are removed such as in the following Arabic examples: عااااجل →عاجل and يااااارب → يارب .

Table 4. Examples of date before and after the rewriting treatment.

Before	After
الاول من شباط 2016	1/2/2016
22 فبراير 2016	22/2/2016
2.2.2016	2/2/2016
13,5	13.5

4.3. Stemming

Arabic language is morphologically rich due to the fundamental rules used for building the words. In fact, often a root is considered as a producer of words, since it is agglutinated to affixes and suffixes to form new words. For example: the root كتب (*to write*), with specific affixes produces different words with different meanings: يكتب (*he writes*), مكتبة (*library*), مكتب(*office*), etc.

To use statistical methods, the words have to be segmented in order to reduce the number of entries in the vocabulary and then to have relevant statistics. That is why, a stemming procedure is run in order to segment the Tweets. The idea is to replace different words which share the same root by the root itself. This will reduce the size of the list of distinct words and leads to a better coverage of the corpus.

In this work, we applied different techniques to retrieve the most representative form of Arabic words. To do so, we combined Buckwalter Arabic Morphological Analyzer with a method based on Light Stemming (LS) presented in previous work [2].

For English, we also used a morphological analyzer to reduce the flexional forms of English words, for that we used the TreeTagger tool [2].

4.4. Proper names in Arabic

Generally, detection of proper names is a critical task for natural language processing applications especially for Arabic. The problem becomes harder when the processing concerns the Tweets. For Latin languages, proper names start with a capital letter, unfortunately, for Arabic, this notion

[2] http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/

does not exist. In addition, personal names refer, in general to common words. For example, the first name كريمة means *generous.*

Another issue is that the proper names could be agglutinated to prefixes and suffixes which requires lemmatization step before performing identification of proper names. For example: بسوريا (*in syria*) must be lemmatized بـ سوريا (Syria).

The majority of research work concerning the extraction of proper names have been dedicated to Modern Standard Arabic (MSA) by Pos-tagging, while only little attention has been given to unstructured text like tweets [15].

Our goal is how to match the English and Arabic proper names, from two bilingual Tweets. Name matching can be defined such as the process of determining, whether two name strings are instances of the same name [4]. This task is not difficult, if the two languages use the same alphabet. Otherwise, a transliteration of a source proper name has to be performed. Transliteration is the action of representing the signs of an alphabet of the source language by the signs of the target language. Transliterating Arabic is more difficult, because for each proper name, several acceptable transliteration candidates could be proposed depending on how it is pronounced in the target language. For example, for the first name: سليمان, the following transliterations are possible: *Sulayman, Seleiman, Sliman and Selayman.*

5. IDENTIFICATION OF COMPARABLE TWEETS

In general building comparable corpora consists in collecting multilingual documents concerning or not a specific topic, then documents are aligned by estimating the degree of their closeness. When we would like to build a comparable corpus of Tweets, the task is a bit complicated because the shortness of the post which makes the matching process more difficult. In fact, the matching process is based on the number of common words between two bilingual documents, unfortunately the number of words, in a Tweet, is very weak which makes this task harder. In the following, we propose a method dedicated to the extraction of comparable Tweets. From two Twitter corpus S and T, in two different languages and for a post s_i^d published at date d, we look for the Tweet $t_j^{\acute{d}}$ published at date $\acute{d}$ respecting the constraint $d-1 \leq \acute{d} \leq d+1$.

We hope that, with this constraint, we retrieve Tweets which concern the same topic. In order to align Tweets, other processing steps are necessary, they are described in the following.

5.1. Number and dates identification

As presented in section 4.2 and 4.1, the processing of dates and numbers is a crucial step allowing to identify similar dates and numbers in Arabic and English Tweets. An homogenization of these items is done in accordance to Tables 3 and 4.

5.2. Proper names identification

To identify Arabic proper names several treatments have been applied. In Arabic, proper names could be simple such as علي (*Ali*) or compounded such as: ابن عبد الرحمان (*ibn Abdul Rahman*) or علاء الدين (*Alaa Aldine*). These compounded proper names are generally composed with a single proper name preceded or followed by particles given in Table 5. We decided to merge them to facilitate their transliteration. For instance, ابن عبد الرحمان (ibn Abdul Rahman) is rewritten into ابن_عبد_الرحمان *ibn_Abdul_Rahman.*

To facilitate the process of matching the proper names, for each Arabic particle, several transliterations are proposed (see Table 5).

Table 5. Particles used in the compound proper names.

Particles	Arabic (English Transliteration)
Prefix	(ابن, بن) (ibn, bin, ben), عبد (3bd, abd), ابو(abu, abo, abou), بنت (bint, bent), ام (oum)
Suffix	الدين (eldin, aldin, uldin,eldin)

For the compounded proper names, it is easy to identify them thanks to the previous particles. While, for the single proper names the task is more difficult. That is why we encode them by taking advantage from their phonetic form. The encoding is done in both English and Arabic Tweet. In the English Tweet, all the words which are not in the bilingual dictionary, are encoded. The hypothesis is that we might consider them as proper names. Then all the words of the Arabic Tweet are encoded. If two strings from respectively Arabic and English Tweet have the same code, we can conclude that, one is the transliteration of the other. For that, we used Soundex [3] which proposes to replace each letter by the index of a group of characters. Each group is constituted by the letters corresponding to almost the same sound class (see Table 6). The characters of Group 0, are ignored unless they appear in the first position of the supposed proper name. Encoding consists in keeping the first character without any change and the following are encoded in accordance to Table 6. Any supposed proper name will be represented by a letter followed by three digits. For example, encoding the proper name جميلة, will give three codes (Figure 2) corresponding to the possible transliterations: *Djamila*, *Jamila* or *Gamila*.

Table 6. Encoding Table of Soundex.

English character	Index	Arabic character
A E H I O U W Y	0	ي و ه ع ح ا
B P F	1	ب ف
C S K G J Q X Z	2	ك ق غ ص ش س ز ج ح
D T	3	ط ظ ض ذ د ث ت
L	4	ل
M N	5	م ن
R	6	ر

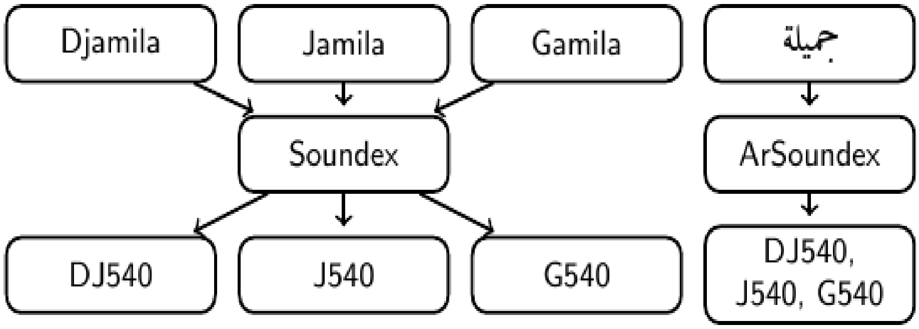


Figure 1. Encoding the proper name جميلة .

The weakness of Soundex is that, it encodes only the first four consonants. This constitutes a serious problem for the compound proper names. For example, *Abdel Aziz* will have the same code (A134) as *Abdel Rahman*. To overcome this constraint, we decided to encode each item of a compounded proper name. That is why, *Abdel Aziz* will be encoded by: *A134A220* and *Abdel Rahman* will be encoded by *A134R550*.

Comparing codes is not enough, in fact the encoding process does not produce a unique code for each proper name. So when an English word is encoded, it is then transformed into an Arabic word thanks to a transliteration table. Then the transliterated word is compared to those in the Arabic Tweet which have the same code as the English encoded word.

5.3. The comparability measure

In order to find similar Tweets, we used an adapted version of Li and Gaussier measure which is based on a bilingual dictionary [8]. The similitude between two Tweets is defined as follows: Let assume that T is an Arabic-English Tweet corpus consisting of an Arabic part T_a and an English part T_e. The comparability measure can be defined as the maximum score between an Arabic Tweet T_a and all the Tweets T_i for 1≤ i ≤ N_{Te} Where N_{Te} is the size of T_e.

For each Arabic Tweet, Score is calculated as follows:

$$Score(t_a, t_e) = \frac{\sum_{w\in T_e \cap D_e} \sigma(w,T_a) + \sum_{w\in T_a \cap D_a} \sigma(w,T_e)}{|T_e \cap D_e| + |T_a \cap D_a|} \qquad (1)$$

where D_e (respectively D_a) is the English side (respectively Arabic side) of a bilingual dictionary. σ is a function which indicates if one of the potential translations T(w) of the word w does exist in the vocabulary V_x.

$$\sigma(w, V_x) = \begin{cases} 1 \; if \; T(w) \cap V_x \neq \emptyset \\ \quad 0 \; else \end{cases} \qquad (2)$$

The adapted comparability measure of Li and Gaussier referred as LGT is calculated as follows:

$$LGT(t_a) = \max_{1\le i \le N_{Te}} Score(t_a, t_e^i) \qquad (3)$$

The size and quality of dictionary may heavily affect the result of comparability measure. In this work, we used the Open Multilingual WordNet (OMW) which contains 17 785 pairs of Arabic and English word. In the previous sections, we proposed several procedures in order to take into account the specificities of Arabic and particularities of the Tweets. In other words, when we identify a number, a date or a proper name, this has to be taken into account in the comparability measure. That is why, we modified the score mentioned in (1) as follows:

$$MScore(t_a, t_e) = \frac{\sum_{w\in T_e} \sigma(w,T_a) + \sum_{w\in T_a} \sigma(w,T_e)}{|T_e| + |T_a|} \qquad (4)$$

σ returns 1 if w has a translation in the target Tweet or if it has been identified such as a number, a date, a proper name, etc. (see section 5). And the new comparability measure is calculates as follows:

$$MLGT(t_a) = \max_{1\le i \le N_{Te}} MScore(t_a, t_e^i) \qquad (5)$$

6. EXPERIMENTAL RESULTS

The idea of identifying comparable corpora is an intermediate milestone, the final goal is to look for the best matched Tweets, in order to retrieve parallel phrases which could be used in machine translation. First of all, we need to measure the comparability of the collected bilingual corpus of Tweets by using the measures presented in Section 5.3. To evaluate the reliability of this measure we run an experiment on two parallel corpora: The first extracted from Twitter [9] which is available at [3] referred in the following as C_{Ling}. This corpus contains just 2006 parallel tweets. To our knowledge, it is the only available parallel corpus Arabic-English. Since this corpus is small, we decided to test on a parallel newspaper corpus which contains 11942 sentences extracted from ANN[4], referred in the following as C_{ANN} .

Concerning the Twitter corpus of Ling et al., unfortunately when we investigated its content, we discovered that the data are not really parallel. Consequently, we cannot used it as a reference corpus for our tests. To illustrate the problems we encountered, we give in Table 7, few examples of what has been considered by the authors as parallel Tweets.

Table 7. Example of parallel tweets extracted by [9] .

Nb	source and target tweet
1	**Source**: اعرف منين امال ماهر **Target**: a3raf menen amal maher
2	**Source**: ★∗○○○∗★∗★∗○○○★★-★ **Target**: ★∗○○○∗★∗★∗○○○★★
3	**Source**: الدراما هذي جمممال فيها المجدول الي ب 39 يوم $$---$$ **Target**: $$,$$
4	**Source**: قطع الماس الماس!!! **Target**: diamond cut diamond !!.. diamante taglio
5	**Source**: تم العثور علي شخص فاهم خطاب مرسي جاري التحقيق معه **Target**: c'est la vie
6	**Source**: ستار اكاديمي 9 في المسبح **Target**: فضائح ستار اكاديمي9

In this corpus some tweets are considered as parallel, though they contain only characters such as in the 2 *th* and 3 *th* examples. This is due to the absence of preprocessing before alignment.
To identify the language of tweets, the authors used a binary function which yields 1 if a word w contains characters of a specific language L, and 0 otherwise. This function is useful if we would like to differentiate Mandarin from English, but not English from French. From a subset of 1000 Tweets, 0.6% of tweet are different from the desired language such as in the 4 *th* and 5 *th* examples.

We found in this parallel corpus, a Tweet in a language which is aligned with itself in the same language which is unacceptable for our purpose (see the 6 *th*).

Due to all these problems, we decided to select 1000 multilingual tweets considered by the authors as parallel and we extracted the real parallel Tweets. Only 34% are strictly parallel. The following tests have been achieved on this subset of parallel Tweets calledC_{Ling34}.

[3] http://www.cs.cmu.edu/~lingwang/microtopia/
[4] www.annahar.com

We run several experiments with the two comparability measures described in Section 5.3 We calculated the classical Recall (R@1, R@5 and R@10). Results are presented in Table 8.

Table 8. LGT and MLGT results for parallel corpora.

Corpora	Method	R@1	R@5	R@10
$\mathrm{C_{ANN}}$	LGT	73	85	87
	MLGT	79.7	89.4	92
$\mathrm{C_{Ling34}}$	LGT	53	73	78
	MLGT	56	77	83

This table shows that the modified LGT achieves better results than LGT since it takes into account the treatment of proper names, dates and numbers. For Twitter corpus, the recall is 56% at R@1 and grows up to 83% at R@10. This result is interesting, it allows to retrieve in the top-10 the right target Tweet.

Concerning the newspaper corpus which is larger, the results are more crucial since at Top-10, we get a recall of 92%. In Table 9, we give some examples of matched tweets.

Table 9. Example of comparable tweets aligned by LGT .

Source: minister kerry: the diplomatic path is the only path that can isolate terrorist groups like daash and front victory 1 2 syria . **Target**: الوزير كيري المسار الدبلوماسي هو المسار الوحيد الذي يمكن ان يعزل الجامعات الارهابية مثل داعش وجبهة النصرة 1 2 **Translation**:minister kerry: the diplomatic path is the only path that can isolate terrorist groups like daash and front
Source: news: obama called putin on syria ceasefire: white house **Target**: البيت الابيض اوباما وبوتن يبحثان وقف اطلاق النار في سوريا **Translation**: white house: obama discusses with putin on syria ceasefire
Source: Syria president bashar al assad issues decree no 63 which sets wednesday 13/4/2016 **Target**: سورية اصدر الرئيس بشار الاسد المرسوم رقم 63 لعام 2016 القاضي بتحديد يوم الاربعاء 2016/4/13 موعدا **Translation**: Syria president bashar al assad issues decree no 63 which sets wednesday 13/4/2016

7. CONCLUSION

In order to obtain a parallel Twitter corpus for which further NLP process could be considered, we developed a method which allows to align the Tweets of a same topic. The experiments achieved showed that for a Tweet, the proposed method can retrieve the corresponding target Tweet with a recall of 83% at R@10. This result has been achieved by a series of preprocessing which permits to align more easily two bilingual Tweets (Arabic and English). Preprocessing is a crucial step, when the data are extracted from social networks and more particularly from those which are written in Arabic. Specific treatments of dates, numbers and transliteration of proper names have been proposed to overcome the issues relates to this specificity. This preprocessing allowed to improve the results by 5.6% in comparison to the baseline model. This result is very encouraging and will permit in a future work to consider the extraction of parallel phrases.

REFERENCES

[1] Sadaf Abdul-Rauf & Holger Schwenk (2009) “On the Use of Comparable Corpora to Improve {SMT} performance”, {EACL} 2009, 12th Conference of the European Chapter of the Association for Computational Linguistics, Proceedings of the Conference, Athens, Greece, March 30 - April 3, 2009.

[2] Karima Abidi & Kamel Smaili, (2016), "Measuring the comparability of multilingual corpora extracted from Twitter and others", The Tenth International Conference on Natural Language Processing, HrTAL2016, Croatia, 29 September – 1 October 2016.

[3] Syed Uzair Aqeel & Steven M. Beitzel & Eric C. Jensen & David A. Grossman & Ophir Frieder (2006) "On the development of name search techniques for Arabic", JASIST .

[4] Peter Christen (2006), "A Comparison of Personal Name Matching: Techniques and Practical Issues", Workshops Proceedings of the 6th {IEEE} International Conference on Data Mining {(ICDM} 2006), 18-22 December 2006, Hong Kong, China.

[5] Thierry Etchegoyhen & Andoni Azpeitia, (2016) "Set-Theoretic Alignment for Comparable Corpora", Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, {ACL} 2016, August 7-12, 2016, Berlin, Germany, Volume1: Long Papers.

[6] Degen Huang & Zhao, Lian & Li, Lishuang & Yu, Haitao(2010)," Mining Large-scale Comparable Corpora from Chinese-English News Collections", Proceedings of the 23rd International Conference on Computational Linguistics: Posters, COLING '10, Beijing, China

[7] Andrey Kutuzov & Mikhail Kopotev & Tatyana Sviridenko & Lyubov Ivanova, (2016), "Clustering Comparable Corpora of Russian and Ukrainian Academic Texts:Word Embeddings and Semantic Fingerprints", CoRR.

[8] Bo Li & Eric Gaussier (2010) , "Improving Corpus Comparability for Bilingual Lexicon Extraction from Comparable Corpora", COLING} 2010, 23rd International Conference on Computational Linguistics, Proceedings of the Conference, 23-27 August 2010, Beijing, China.

[9] Wang Ling & Guang Xiang & Chris Dyer & Alan W. Black & Isabel Trancoso (2013), "Microblogs as Parallel Corpora", Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, {ACL} 2013, 4-9 August 2013, Sofia, Bulgaria, Volume 1: Long Papers

[10] Dragos Stefan Munteanu & Daniel Marcu (2005)," Improving Machine Translation Performance by Exploiting Non-Parallel Corpora", Computational Linguistics .

[11] Franz Josef Och and Hermann Ney (2003), "A Systematic Comparison of Various Statistical Alignment Models", Computational Linguistics, Page: 19--51.

[12] Motaz Saad and David Langlois and Kamel Smaili (2014), "Cross-lingual semantic similarity measure for comparable articles", In Advances in Natural Language Processing - 9th International Conference on NLP, PolTAL 2014, Warsaw, Poland, September 17-19, 2014. Proceedings, pages 105-115. Springer International Publishing.

[13] Inguna Skadina & Ahmet Aker &Voula Giouli & Dan Tufis & Robert J. Gaizauskas & Madara Mierina &Nikos Mastropavlos (2010)), "A Collection of Comparable Corpora for Under-resourced Languages", Human Language Technologies - The Baltic Perspective - Proceedings of the Fourth International Conference Baltic {HLT} 2010, Riga, Latvia, October 7-8, 2010.

[14] Ivan Vulic & Marie Francine Moens (2015), "Monolingual and Cross-Lingual Information Retrieval Models Based on (Bilingual) Word Embeddings", Proceedings of the 38th International {ACM} {SIGIR} Conference on Research and Development in Information Retrieval, Santiago, Chile, August 9-13, 2015.

[15] Omnia H. Zayed and Samhaa R and El-Beltagy,(2015) "Hybrid Approach for Extracting Arabic Persons Names and Resolving their Ambiguity from Twitter", In 20th International Conference on Application of Natural Language to Information Systems (NLDB 2015), Passau, Germany, June. Springer.

[16] Zhu, Zede and Li, Miao and Chen, Lei and Yang, Zhenxin,(2013), “Building Comparable Corpora Based on Bilingual LDA Model.”, ACL (2) Page: 278--282.

AUTHOR INDEX