A Classification of Character Usage in Unique Addresses Employed for Accessing Yahoo! Groups Service

Jatinderkumar R. Saini[1], Apurva A. Desai[2]

ABSTRACT

A tremendous increase in the use of internet for online communication like message sending is witnessed worldwide. Yahoo! Inc. provides one such service in the form of Yahoo! Groups. Each such group is identified and accessed using a unique group address. The current paper presents an analysis of nearly 5000 Yahoo-group addresses. It presents a classification of characters employed by users in designing these addresses into 5 major sets. Our results show that around 90% characters used for designing the Yahoo-group addresses are alphabets whereas the remaining 10% constitute from the domain of digits and special characters. The paper also elaborates on the divisional values of these proportions highlighting the user's preference for selecting a particular character. To the best of our knowledge this is the first attempt to study online user behavior based on the classification of character usage for designing a unique online identifier.

Keywords : Character Usage, Digits, Lower-case Letters, Upper-case Letters, Yahoo! Groups, Yahoo-group address, Yahoo-group Identifier

[1]Associate Professor and Head, Department of Computer Science, S. P. College of Engineering, Visnagar, Mehsana, Gujarat, India. Mob. No. +91-9426861815. Email: saini_expert@yahoo.com

[2]Professor and Head, Department of Computer Science, Veer Narmad South Gujarat University, Surat, Gujarat, India. Email: desai_apu@hotmail.com

1. INTRODUCTION

The growth of internet has provided the users of 21st century with new means of communication. According to Jones and Fox [5], instant messaging, social networking, and blogging have gained ground as communications tools, but email remains the most popular online activity. Besides these, other online activities include sharing views and discussing various topics through groups and discussion forums. Yahoo! Inc. is one company that provides such internet services worldwide. The specific services provided by Yahoo! Inc. include mail, news, search, groups, video and maps, to name a few. Yahoo! Inc. was founded in 1995 and Yahoo! Groups which provides one of its services, came into existence in 2000 [9].

The Internet Marketing Definitions website [4] describes Yahoo! Groups as a service that operates as both electronic mailing list and Internet forum. Members can post and read messages either by receiving them in their email account or by going to the group's homepage. Since the inception of Yahoo! Groups there has been a tremendous increase in its user bank. This service allows the user to create a group address and provide group name. Technically speaking, the group address is the one that is used to uniquely identify a Yahoo-group. In the current work, the concentration has been only on group addresses and not on group names which can consist of any character combination and are meant to provide a brief introductory description line of the group.

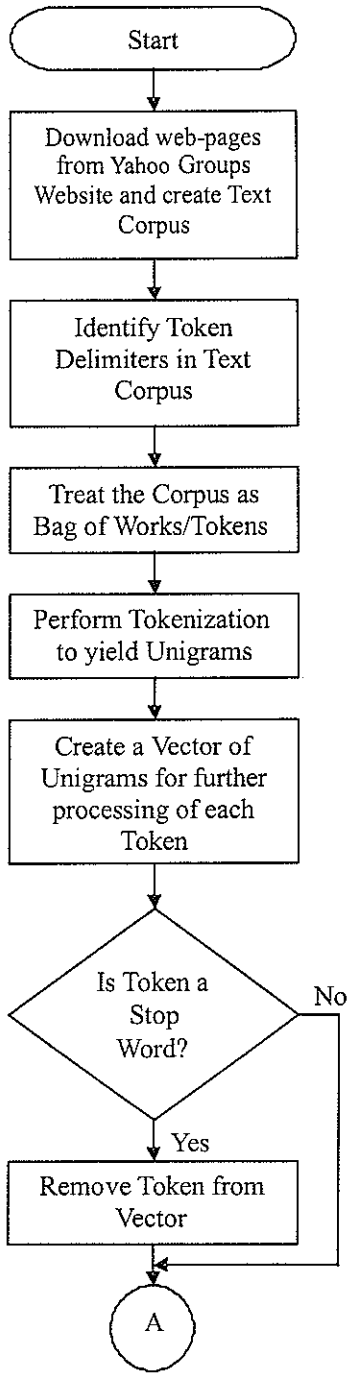
2. Related Work

There are quite a few research instances available with research community, that have thrown light on statistical analysis of character usage. Even for those available, most of the works have concentrated on usage of characters as a stylometric parameter. Aaronson [1] has listed a set of 15 features which are important for performing data-driven stylometric analysis. These features are Ampersand Sign, Apostrophe, Colon, Comma, Dash, Dollar Sign, Exclamation Mark, Forward Slash, Left Parenthesis, Percent Sign, Period, Question Mark, Right Double Quote, Right Parenthesis and Semi-colon.

An instance of analysis of character usage is provided by the work of Saini [6]. In his work, he has presented a detailed discussion on the usage of characters by spammers for sending Unsolicited Bulk Emails (UBE) commonly known as spam emails. Calix et al. [3] in their work which was targeted towards e-mail author identification and authentication have also employed the statistical analysis of characters used in the emails. They have provided a platform to identify the author of a given e-mail based on writing-style features like number of words, number of commas, number of times "well" appears, etc. To the best of our knowledge and survey of related research literature, this is the first formal attempt aimed towards classification of usage of characters used for designing the Yahoo-group addresses.

3. Methodology

During this phase, the goal of following a methodological sequence of steps was to obtain a list of Yahoo-group addresses which could be subjected to further analysis. For better comprehension of the process, the sequence of steps followed towards this end is presented in the form of a flow-chart in Figure 1. Towards the first step, a corpus, consisting of various web pages retrieved from Yahoo website, is created. Specifically, the target area in the Yahoo website is Yahoo Groups, which follows a hierarchical structure. This can be visualized as an inverted tree, wherein Yahoo itself acts as the root node at top-most level and other nodes reside at various lower levels



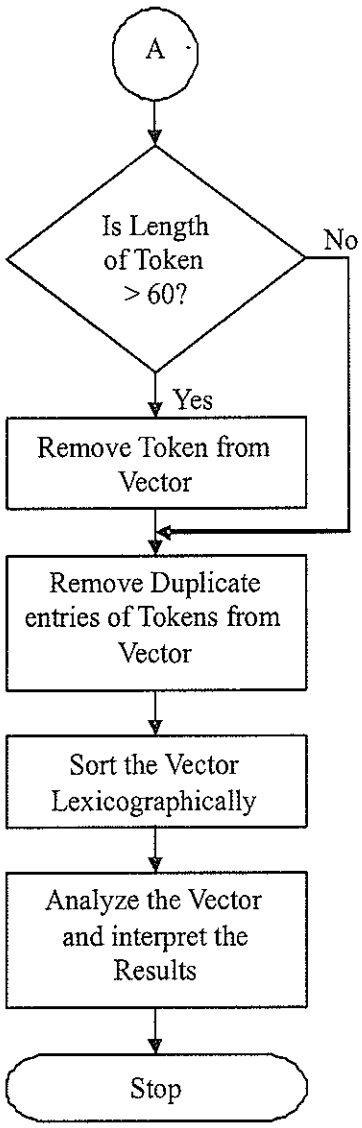


Figure 1: Classification and Usage of Constituent Characters of Yahoo-group Addresses

The relevant top-most level, just below the root node consists of 17 broad areas [10]. Each of these areas is further divided into various groups and these groups may be further recursively sub-divided into sub-groups up to any level. The leaf nodes resulting at the bottom-most level of the visualized inverted tree structure consist of actual Yahoo-group addresses. For the purpose of current work, the selection of Yahoo web pages constituting the corpus was done randomly from category, groups and sub-groups of any level.

The basic model of free text consists of documents which are sequences of basic units called Tokens. In English language the tokens are words [11] and the act of breaking the text into tokens is called Tokenization. In order to make it easier for analysis and further processing, Tokenization is performed on the corpus of Yahoo web pages. Here, the problem was to identify the token delimiters. This means to say that there was a need to identify the beginning and ending of a token in a sentence. As a de-facto standard, it is known that space character is not allowed in web technologies for a number of instances like creation of email addresses and (Uniform Resource Locator) URL addresses. Besides, Yahoo also has explicitly specified that spaces are not to be allowed for creation of a Yahoo-group address. We tried the creation of a Yahoo-group address by using non-printable space (ASCII Code 32) character. The submission of this request yielded us with the expected and desired message. The message read as, "The group address may only contain letters, numbers, and - or _ characters. It may not have more than one - or _ in a row." Consequently, it was possible to design the tokenization process to identify the words demarcated by any character other than letters, numbers, Dash (-) and Underscore (-). Tokenization, hence, was done in such a way that only unigrams could be focused.

The various text-processing activities completed so far yielded us with a text corpus in the Bag of Words (BOW) form. In BOW representation of a text document, terms or tokens in the document are identified with words in the document. Hence this representation is also called Set of Words (SOW) [7]. The BOW was further processed by removal of stop words from it. Sebastiani [8] has defined Stop Words as topic-neutral words such as articles and prepositions, which are eliminated in a pre-processing phase. Bharati et al. [2] have defined them as

few words which have high frequencies in all the categories, and hence are irrelevant for the classification exercise. The removal of stop words from the vector was helped by the fact that consecutive Dash (-) or consecutive Underscore (_) or consecutive combination of Dash and Underscore in any order was not allowed by Yahoo for the creation of a valid Yahoo-group address. Further, it was found that for creation of a Yahoo-group address, Yahoo explicitly specifies that it cannot have length more than 60 characters. Hence, tokens with length of more than 60 character length were also removed from the vector due to their irrelevance for the context of current work.

It is also to be noted that Yahoo allows the usage of only digits for creation of a Yahoo-group address. For instance an address formed of numbers like 12345 will be treated as a valid Yahoo Groups address by Yahoo website. Hence, the addresses consisting of only digits were prevented from removal as stop words. This processing of the BOW resulted in a vector consisting of 5129 word entries. As the vector was generated from the processing of a large text corpus, it was natural for the vector to contain duplicated entries. As a next step, the vector was refined by selecting only unique words. The removal of Stop Words was done before the removal of duplicated entries because the bulk of stop words was much more than the bulk of duplicated entries.

Table 1: Partial Snap-shots of Vector Containing Yahoo-group Addresses

Sr. No.	Vector Index No.	Yahoo-group Address	Sr. No.	Vector Index No.	Yahoo-group Address
1	37	2joel_kami	16	573	aykan8691
2	38	2ofy	17	574	aysegulunfistiklari
3	39	303kd_others	18	575	azalizan_ismail
4	40	317arazi27	19	576	azanimalrights
5	41	33_SSB	20	577	aziyar
6	42	3bru	21	578	azncrew
7	43	3e6-o7	22	579	b_vikas1986
8	44	3lionsroaras1	23	580	b122
9	45	407-thottigang	24	581	B29_pradan
10	46	43MirpurFriends	25	582	baarn
11	47	4asap	26	583	babajigroup
12	48	4-ever_friends	27	584	BABASEMENTgirls
13	49	4frenz	28	585	bablo_bablo40
14	50	4gats	29	586	baby-girl-4eva
15	51	4-i_green-mindedclub	30	587	babysugar

The reduced vector obtained after removal of duplicated entries consisted of 4940 entries and was ordered by sorting it in lexicographic ascending manner. The entries in the token vector of our experimental setup represented the Yahoo-group identifiers of the real world. Out of a total list of 4940 entries, two partial snap-shots of this vector from index positions 37 to 51 and 573 to 587 are selected randomly and presented in Table 1.

4. Results and Findings

The end of various activities of processing the text-corpus was a one-dimensional vector consisting of 4940 unique Yahoo-group addresses. By the analysis of the data available in the vector, it was found that Yahoo allows upper-case, lower-case as well as mixed-case characters to be used for creation of Yahoo-group address. Moreover, by attempting to create multiple Yahoo-group addresses by changing the case of letters, it was found that Yahoo does not treat its group addresses as case-sensitive. This means to say that if a group address called 'abc' is already created, then one cannot create a group address called 'ABC' or one with mix-case letters like 'Abc'. This is true also for addresses created in different categories. In other words, it can be said that a Yahoo-group address is unique irrespective of its belonging or categorization to any category in Yahoo groups. For the current work, the treatment of text corpus is considered to be case-sensitive. This was done to help meet our objective of studying the usage of lower-case and upper-case letters by users, in designing the Yahoo-group addresses. Each of the address in the vector partially depicted in Table 1, was a unigram and was made up of specific characters. These constituent characters used for creation of addresses into 5 categories were classified and the corresponding listing is presented in Table 2.

Table 2: Classification and Usage of Constituent Character Representations (CCR) in Yahoo-group Addresses

Sr. No.	Character Type	No. of CCR	Usage of CCR	CCR Usage in %
1	Lower-case Letters	26	52764	83.46
2	Upper-case Letters	26	4048	6.40
3	Digits	10	3945	6.24
4	Underscore	1	2159	3.41
5	Dash	1	306	0.48
6	Others	1	9880	-
Total		65	63222 (except others)	100

From Table 2, it can be seen that the Yahoo-group addresses could be created by either exclusive usage or any combinational usage of 26 English lower-case letters, 26 English upper-case letters, 10 digits from 0 (zero) to 9 (nine) and two special characters Dash and Underscore. Here it is to be noted that only consecutive usage of special characters is not allowed by Yahoo. Finally, it was also tried to analyze the vector for any other characters that could have been missed otherwise or that could have admitted in unknowingly.

It is noteworthy to see that the total number of other characters found by us was 9880. This value is equal to multiplication of 2 with 4940 which is the total number of tokens. This value comes into play due to counting of carriage-return and new-line characters during the processing of the token vector. Also since no other character was found either omitted for analysis or admitted unknowingly, for further discussion, value of 9880 is not given any consideration. This value has been treated as statistically insignificant from the viewpoint of objectives of current work. Hence, in Table 2, the number of Constituent Character Representations (CCR) of Yahoo-group addresses for character type of 'others' is also depicted as 1. Table 2 also depicts the usage of various representations of each character type by the designers of Yahoo-group addresses.

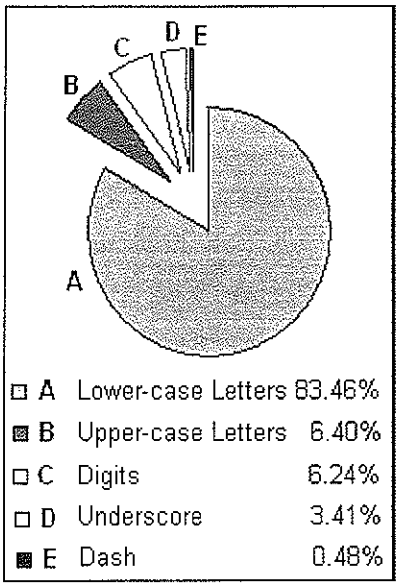


Figure 2: Classification and Usage of Constituent Characters of Yahoo-group Addresses

The data corresponding to this is presented graphically in Figure 2. For simplicity and better comprehension, the values of character usage in Figure 2 have been represented in terms of percentage of usage. From Figure 2, it can be seen that users have made almost 90% use of the English alphabet characters for designing the Yahoo-group addresses. Further, the lower-case letters constitute 83.46% of the total character usage whereas upper-case English letters constitute 6.40% of the total value. An interesting result to note here was that almost 10% of the characters used for designing Yahoo-group addresses are not from the domain of English alphabets. This value of 10% is further divided into three categories for ten digits and two special characters. The digits contribute 6.24% of the total character usage for designing Yahoo-group addresses. The underscore special character constitutes 3.41% followed by the dash special character constituting a meager amount of 0.48% of the total character usage.

5. Conclusion

The current paper is an attempt to analyze the addresses used for uniquely identifying the access of Yahoo! Groups service. The classification of usage of various characters, employed towards designing nearly 5000 Yahoo-group addresses by users, has been presented. It is concluded that even though Yahoo allows the design of group addresses with lower-case, upper-case and mixed case characters, it does not treat them as case-sensitive. Hence, the usage of different case characters in Yahoo-group addresses is meant either to make them visually pleasing or to let users incorporate different words in the identifier, each word demarcated by change in case of the letter.

It is further concluded that the character set used for designing Yahoo-group addresses could be divided into five categories for lower-case letters, upper-case letters, digits, underscore and dash. Additionally, of the total character usage, lower-case letters constitute the maximum amount of nearly 84% share for designing the

Yahoo-group addresses. This is followed by an almost equal usage of upper-case letters and digits with value of around 6% for both. The use of underscore is very less with value of around 3%. The preference of usage of dash character by users for designing Yahoo-group addresses is minimum, with a value of around 1% only.

Our results are best reported on the dataset used and we do not promote or discourage the use of any specific character for designing of Yahoo-group address. We just present the classification of usage of characters based on the preferences of users for designing the Yahoo-group addresses. The current work is not only an insight into the usage of characters for designing Yahoo-group identifiers but is also having a wide range of general applicability to other domains. On the sidelines of the current study, it is advocated that the paper has also provided an insight into usage and behavior of selection of characters by users for designing passwords as well as other unique identifiers in the online and offline world.

References

[1] Aaronson S. "*Stylometric Clustering*", Talk at University of Toronto, March 30, 2000, Available: http://www.scottaaronson.com/ talks/scslides.doc

[2] Bharati A., Varanasi K., Kamisetty C., Sangal R. and Bendre S. M. "*A Document Space Model for Automated Text Classification based on Frequency Distribution across Categories*" in the Proceedings of International Conference on Natural Language Processing (ICNLP – 2002), Recent Advances in Natural Language Processing, Vikas Publishing House, New Delhi, 2002

[3] Calix K., Connors M., Levy D., Manzar H., McCabe G. and Westcott S. "*Stylometry for E-mail Author Identification and Authentication*", in Proceedings of CSIS Research Day, Seidenberg School of CSIS, Pace University, New York. May 2008.

[4] Internet Marketing Definitions. "*Yahoo Groups*" Available: http://www.internetmarketingdefinitions.com/ YahooGroups

[5] Jones S. and Fox S. "*Generations Online in 2009*", Pew Internet and American Life Project of Pew Research Center, Washington, January 2009. Available: http://www.pewinternet.org/pdfs / PIP_Generations_2009.pdf

[6] Saini J. R. "*Self Learning Taxonomical Classification System using Vector Space Document Analysis Model for Web Text Mining in UBE*", Ph.D. Thesis guided by Desai Apurva A., accepted by Department of Computer Science, Veer Narmad South Gujarat University, Surat, Gujarat, India, September 2009

[7] Sebastiani F. "*Machine Learning in Automated Text Categorization*" in ACM Computing Surveys (CSUR), vol. 32, issue no. 1, pp. 1-47, March 2002. ISSN: 0360-0300

[8] Sebastiani F. "*Text Categorization*" in Text Mining and its Applications, Alessandro Zanasi (ed.), WIT Press, Southampton, UK, pp. 109-129, 2005

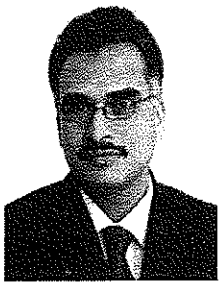
[9] Wikipedia, the free encyclopedia. "*Yahoo!*", Wikimedia Foundation Inc., Available: http://en.wikipedia.org/wiki/Yahoo!

[10] Yahoo! Inc. "*Yahoo Groups*", Available: http://groups.yahoo.com

[11] Zhang T. "*Predictive Methods for Text Mining*", Machine Learning Summer School - 2006, Taipei. Available:videolectures.net/mlss06tw_zhang_pmtm

Author's Biography

Jatinderkumar R. Saini is Ph.D. from Veer Narmad South Gujarat University, Surat, Gujarat, India. He secured First Rank in all three years of MCA at college and has been awarded Gold Medals for this. He is also recipient of Silver Medal for B.Sc. (Comp. Sci.). He is IBM Certified Database Associate – DB2 as well as IBM Certified Associate Developer – RAD. He has presented 12 papers in international and national conferences, all sponsored by either AICTE or ISTE. One of his papers has also won the 'Best Paper Award'. His 5 papers have been accepted for publication at international level and 8 papers have been accepted for national level publication. He is a member of many academic committees. He is also a member of various international and national professional bodies and scientific research academies and organizations.



Apurva A. Desai completed his post graduation from Veer Narmad South Gujarat University, securing First Rank in the University. He earned his Ph.D. in the year 1997. He has got a long teaching and research experience since 1990. Many students have completed Ph.D. and M.Phil. under his supervision. He has delivered many lectures and invited talks as resource person in various national and international events. He has 3 books and 13 research papers to his credit. He is chairman of Board of Studies (Computer Science) for last 6 years. He has attended many International and National conferences. He has also visited Canada and Italy for various academic programmes.