International Journal of Business, Human and Social Sciences
ISSN: 2517-9411
Vol:8, No:2, 2014

Structural Analysis of Username Segment in E-Mail Addresses of Engineering Institutes of Gujarat State of India

Jatinderkumar R. Saini

Abstract**—**E-mail has become a key mechanism of electronic communication. This is true for professional organizations that like to communicate with their subjects online and are slowly shifting to paper-less office. The current paper focuses specifically on academic institutions offering Engineering course in Gujarat state and attempts for textual analysis of the usernames of the institutional e-mail addresses. We found that the institutions tend to design the username segment of their e-mail addresses by choosing words or combination of words from specific categories. The paper also highlights the use of special characters, digits and random words in designing the usernames. On the sidelines, the paper lists the style of employing department names and designations for the design process. To the best of our knowledge, this is the first formal attempt to analyze the selection of words employed for designing username segment of e-mail addresses of engineering institutions.

Keywords**—**E-mail address, Institute, Engineering, Username.

I. Introduction

TECHNICALLY defined, e-mail, short for electronic mail and often abbreviated to e-mail or simply mail, is a store and forward method of composing, sending, receiving and storing messages over electronic communication systems [9]. Since e-mail is fast, cheap, and easy to send, it has gained enormous popularity not simply as a means for letting friends and colleagues exchange messages, but also as a medium for making paper-less office possible. From the point of view of institutions, it has become a chief source of communication and providing information and replying to the queries received through e-mail messages. From structural perspective, e-mail addresses have two parts. The part before the @ sign is the local-part of the address, often the username of the recipient, and the part after the @ sign is the domain to which the e-mail message will be sent [10]. The local-part of the address before the @ sign is also called username.

Situated on the west coast of India between 20.6 North to 24.42 North latitude and 68.10 East to 74.28 East longitude, Gujarat is one of most industrialized states in western India. It has a geographical area of 196,024 square kilometers [8]. Comparing with the countries of the world, only 87 countries have area more than Gujarat State of India [12]. It is the country's one of the wealthiest states, supporting modern industrial complexes as well as thriving village handicrafts. According to the census of 2001, the population of Gujarat was more than 5.06 crores (nearly 51 million). The glimpse of the prosperity of the state can just be obtained from the fact that it has the largest grass-root Petroleum Refinery in the world operational at Jamnagar. It has maximum capacity smelter in Asia for production of copper, gold and silver. The Gujarat Industrial Development Corporation (G.I.D.C.) is the biggest industrial area in Asia and among top 25 in the world. Gujarat is leader in various industrial sectors like chemicals, petrochemicals, drugs & pharmaceuticals, dairy, cement & ceramics, textiles, engineering and gems & jewellery. It has highest number of 12 Airports in India, including an international airport in Ahmedabad.

According to the 2001 census, around 70% (excluding children in the age group 0-6 years) of the population of the state was literate [8]. Speaking of education, Gujarat also has various world-class educational institutions like Indian Institute of Management (IIM), Ahmedabad, National Institute of Technology (NIT), Surat and Indian Institute of Technology (IIT), Gandhinagar. Institute of Rural Management, Anand imparts knowledge on rural management and is the only of its kind institute in Asia. Gujarat is 7th most literate state in India with average literacy rate higher than that of the country itself [2]. It has 17 universities [11] imparting education in varied spheres of arts, laws, sciences, engineering, management and medical fields. Specifically, the number of engineering institutions in Gujarat state of India is around 100.

This paper presents the analysis of usage of words selected for designing the username in the e-mail addresses for the engineering institutes of Gujarat state of India. The courses offered by such institutes comprise of two types of gradations. The first kind of gradation is called 'Diploma' or 'Diploma in Engineering' (DEngg) whereas the second kind of gradation is called 'Degree'. Further, for the 'Degree' gradation the award is either at the Graduation Level called the 'Bachelor of Engineering' (BE) or at the Post Graduation Level called the 'Master of Engineering' (ME). Some academic institutions also use the term 'Bachelor of Technology' (BTech) and 'Master of Technology' (MTech) respectively for the Graduation and Post Graduation Level Degrees. This paper focuses on the e-mail addresses chosen by the institutes that offer one or any number of courses from DEngg, BE, BTech, ME and MTech. Moreover, the term institute in the context of current paper means the various academic centers in Gujarat state offering the full-time courses for 3 year of DEngg, 4 year of BE or BTech and 2 year of ME or MTech. This includes the

J. R. Saini is with the Narmada College of Computer Application, Bharuch, Gujarat, India as Associate Professor and I/C Director. He is PhD from Veer Narmad South Gujarat University, Surat, Gujarat, India (phone: +91-9687689708; e-mail: saini_expert@yahoo.com).

university departments, colleges as well as college departments offering these courses. As the e-mail address is used to contact the institution, in general, we refer to this kind of address as 'institutional e-mail address'. Further, as this address is pertinent to the engineering institutions, we also refer to this address as that of 'engineering institutions'.

II. Related Work

The research world presents a mix of usage of e-mails for analysis. There are a large number of instances in the research world where the researchers have worked with the electronic mail messages with focus on classification of mail into spam and non-spam categories. Calix et al. [1] in their work which was targeted towards e-mail author identification and authentication have also employed the statistical analysis of characters used in the e-mails. 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. The notable thing here is that these kinds of research works make use of the text of mail message and not the e-mail address, per se. As another case, the researcher includes the e-mail address itself in the text corpus but the focus still is targeted towards the text of the mail message.

There is a definite dearth of research works targeting the analysis of e-mail addresses. Saini and Desai [4] have presented a digit-wise statistics and detailed analysis of the usage of digits based on study of nearly 1400 unique Yahoo-group addresses containing digits. They have also identified a number of areas that influence the selection of a digit or digits by the users in this design process. Saini and Desai in other works have presented a structural analysis of username segment in e-mail addresses of MCA Institutes [5] and MBA Institutes [6] of Gujarat State of India. To the best of our knowledge and survey of related research literature, this is the first formal attempt aimed towards username based structural analysis of the e-mail addresses of engineering institutions.

III. Methodology

In this section, we describe the detailed methodology followed by us for the structural analysis of the e-mail addresses of the Engineering institutes of Gujarat State. For the sake of simplicity and better understanding, the entire section is divided into three sub-sections for data collection, data cleaning and data analysis.

A. Data Collection

First of all, a crude corpus of text was created by collecting the text containing required data from various available sources. The websites of various Engineering institutions, home-pages of various Engineering departments of colleges and universities and portable document format (PDF) files listing Engineering institutions with their contact addresses provided the available sources to collect text containing pertinent data. Such PDF files contain contact details of various institutions and are available online owing to a number of reasons. For instance, this format is used by some universities to provide contact details of their affiliated institutions and by nodal agencies conducting the entrance tests to first list the centers where application forms are available and then to list the exam-centers. From our perspective, it was not important that the file had to be either in one format or the other. The availability of e-mail addresses of engineering institutions in the file was our key concern, irrespective of the format the file is in. Our target area of analysis was only e-mail addresses of Engineering institutions and hence the care was taken to see that e-mail addresses of other courses like 'Master of Computer Application' (MCA), 'Master of Business Administration' (MBA) and 'Bachelor of Pharmacy' (BPharm) do not creep inside our data. This, otherwise, could have biased the analysis of data by inclusion of impurities in the form of related but different data.

B. Data Cleaning

Next, we checked for institutional e-mail addresses provided in a split format, i.e. entire e-mail address not provided in the same line and a portion of the address broken and provided in the next line. We also checked for a similar case wherein an extra space or a character like comma, semi-colon, etc. has crawled into the e-mail address. It is noteworthy here that such characters are invalid characters for an e-mail address. The number of such instances found by us was 4. The created corpus was then refined by removing all text not related with the e-mail addresses. One of the Request For Comment (RFC) in which e-mail addresses are formally defined is RFC 5321 [3]. The refinement of the corpus was hence eased by the two typical structural characteristics of the e-mail addresses. First, an e-mail address cannot contain a space character (ASCII value 32) and second, it will necessarily contain @ character (ASCII value 64). The existing un-structured text corpus was analyzed syntactically by performing sentence splitting and Tokenization. The basic model of free text consists of documents which are sequences of basic units called Tokens. In English language, the tokens are words [13] and the act of breaking the text into tokens is called Tokenization. The un-broken e-mail address in its totality itself was treated as a token and the entire corpus was treated as a bag of words (BOW) containing various tokens. 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]. Those tokens, except e-mail addresses, were of no analytical relevance to our work and hence removed from the corpus. This final refinement of the corpus yielded us with one dimensional vector containing 71 valid e-mail addresses. It was natural for this vector to contain duplicated entries. Hence, as a final step, a new one dimensional vector containing 56 unique e-mail addresses was created. This vector was sorted in ascending order.

C. Data Analysis

The vector with 56 unique entries corresponding to the e-mail addresses provided the input for intended analysis. The e-mail addresses were analyzed based on the username segment of the addresses. Based on this analysis, we classified the addresses into 5 main categories. The pertinent statistical data is presented in Table I.

TABLE I
FREQUENCY OF USERNAMES OF E-MAIL ADDRESSES CORRESPONDING TO MAIN CATEGORIES

Sr. No.	Main Category	No. of Addresses
1	CITY	0
2	COLLEGE	17
3	DESIG	9
4	NAME	4
5	OTHER WORD	9
Total	5	39

CITY is a category that accounts for all e-mail addresses containing only the city's name in the username segment of the institutional e-mail address. Not much surprisingly, there was no address whose username consisted of just the city-name. COLLEGE is a category which counts all the e-mail addresses whose username consists only of the name of the college. We found that there were 17 such instances and actually full college name was found nowhere, instead the abbreviated college name was used. DESIG is the category consisting of words denoting the designations in the username of the e-mail address. The 9 entries corresponded to the designations from a list comprising of 'admin' (meaning administrator), 'principal', 'vc' (meaning vice-chancellor), 'director' and 'dir' (meaning director). NAME is a category which is used where the name of the Principal, Director, Head of Department (HoD), coordinator or concerned administrative head is directly used in the username part of the e-mail address. We found 4 instances in this category. OTHER WORD is the category where the username consists of words not befitting any of the above categories. This includes words like 'info', 'contact'/'contactus' and 'mail' used for the username segment of the e-mail address. The numbers of instances in this category were 9 and 5 of these used the word 'info' for the username part of the e-mail address. The words used for the remaining 4 instances were found not to have any apparent understandable meaning.

TABLE II
FREQUENCY OF USERNAMES OF E-MAIL ADDRESSES CORRESPONDING TO COMBINATIONAL CATEGORIES

Sr. No.	Combinational Category	No. of Addresses
1	COLLEGE_CITY	11
2	COLLEGE_DESIG	2
3	COLLEGE_OTHER WORD	2
4	COLLEGE_CITY_OTHER WORD	2
Total	4	17

After the description of the created categories, we now proceed to describe the steps followed by us for the creation of the categories. We first started with none existing category. Then the scanning of the vector from top to bottom was performed in such a way that when the first category was created, its corresponding count was initialized to 1. Next, if a new category is created, its count is similarly, initialized to 1. At any point of time, if a category is already found to exist, its count is incremented by 1. It needs to be mentioned here that the category in our context is nothing but occurrence of a word belonging to one of the special 5 classes proposed by us. For instance, DESIG is a category for which the count is incremented whenever any e-mail address containing a designation in the username is found. The execution of this procedure for all 56 e-mail addresses yielded us with the categories tabulated in Table I.

A specific thing to be noted here is that the categories in Table I are only the main categories and the corresponding counter is maintained only if the entire username is belonging to any of the 5 categories. This means to say that if username consisting of any combination of the 5 categories is found, it is not counted towards the data of Table I. In order to handle such circumstances, we created another 4 categories by combination of the main categories derived from Table I. This second set of categories is presented in Table II. To differentiate the categories of Tables I and II, we call them 'Main Categories' and 'Combinational Categories', respectively. Put simply, the 'Main Categories' comprise of single category words whereas 'Combinational Categories' comprise of two or three category words.

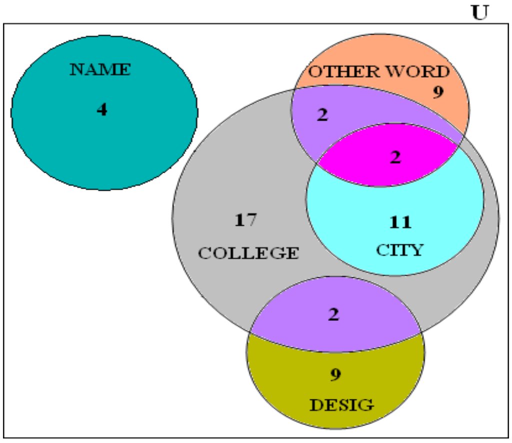

Fig. 1 Frequency distribution of usernames corresponding to main and combinational categories

The logic of creation of categories of Table II was also the same as for the categories in Table I, i.e., the first instance initializes the counter and second instance of the word belonging to a category, increments it. A category in Table II is identified by the combined name of categories of Table I. For instance, e-mail address in the combinational category COLLEGE_CITY means that the username is formed of abbreviated college name and city name. Similarly, COLLEGE_DESIG is abbreviated college or university name and the designation of the administrative head of the institute. In addition to the designations already listed, we also found a designation called 'princi' (meaning principal). The other combinational categories are similarly interpretable. COLLEGE_OTHER WORD combinational category needs special mention as it includes all addresses which contain abbreviated college name combined with words like ‘info’, year, random numbers and randomly selected characters. In this category, we were able to find only 2 instances. Finally, we created a category called ‘COLLEGE_CITY_OTHER WORD’ for which 1 instance of username of e-mail address was found to exist.

The inter-play of main categories and combinational categories is best represented diagrammatically in Fig. 1. The area under intersection of different main categories forms combinational categories. In Fig. 1, the space occupied by this area is not relevant to the number of instances depicted for that main or combinational category. It is notable here that from 5 main categories, 10 unique combinational pairs could be created, given that there is no difference between the combination and its reverse. For instance, COLLEGE_CITY and CITY_COLLEGE mean one and the same thing. Further, instead of creating only paired combinations, triplets (e.g. COLLEGE_CITY_DESIG), quadruplets (COLLEGE_CITY_DESIG_NAME), etc. combinational categories could also be created. Here it is notable that we neither created all possible 10 combinational pairs nor did we create more combinations of triplets and higher order combinations (beyond triplets). We contemplated only those combinations which we found to exist in the usernames available to us in the refined 56-rowed vector of unique e-mail addresses.

IV. Results and Findings

Our analysis was based on the institutional e-mail addresses for engineering institutes. The term institute in our context scoped for colleges, university departments and college departments. Based on the analysis of such 56 unique institutional e-mail addresses, we found that the main words dominating the selection of username for e-mail addresses belong to categories like abbreviated institute name, name of city where the institute is located, designation of the administrative head and name of the administrative head of the college or university department. We found that usernames also exist with un-ordered paired combinations of these categories as well. Other words are chosen with combination to categories like college name or without combination to mainly dominating categories. The inclusion of other words may range from general words like year, random numbers and random characters to specific words like ‘info’. Specifically, ‘info’ was found to be used 5 times, followed by 5 times usage of other most preferred word ‘principal’ (including one time usage of word ‘princi’). This was followed by 3 times usage of word ‘admin’. With an individual count of 5 instances, the word ‘info’ also happened to be the most frequently singly used word in the usernames of the e-mail addresses of the Engineering institutions. Apart from ‘OTHER WORD’, the single main category having maximum number of instances was that of ‘COLLEGE’ with a count of 17 instances of abbreviated college names. This was followed by ‘DESIG’ category with a count of 9 instances comprising of words like ‘admin’, ‘principal’, ‘princi’, ‘director’, ‘dir’ and ‘vc’. The category of ‘CITY’ was found to depict special behavior because in main category section, it had no instances.

In the combinational category, the abbreviated college name is found to have a maximum of 17 combinations with other categories with the highest number of 11 usernames recorded with a combination of city name. Also, there is no combinational category where abbreviated college name is not used. Further, ‘NAME’ is a category which is not having any combination with other categories.

Based on the analysis of the e-mail addresses, it was also found that the username with maximum length had 22 characters whereas the one with lowest length was only 2 characters long. The average length of usernames was found to be 7.66. On analyzing the institutional e-mail addresses on basis of usage of special characters in their designing, we found that underscore, i.e. character ‘_’ (ASCII value 95) was the most used special character with a frequency of 12. This was followed by the usage of digits from 0 to 9 (ASCII value 48 to 57) with a frequency of 8. The last position was occupied by special character period, i.e. ‘.’ (ASCII value 46) with a frequency of only 1.

V.Conclusion

Based on the structural analysis of usernames selected by engineering institutes of Gujarat for their institutional e-mail addresses, we conclude that the selection is mainly dominated by the words related to abbreviated college name, combination of abbreviated college name with its city name, the designation of the administrative head of the institute and usage of word ‘info’. We further conclude that these areas, in this order, pose the main attraction for selecting the usernames for institutional e-mail addresses. Hence, we strongly deduce that the current trend is a clear indication of selecting username for institutional e-mail address in such a way that it clearly identifies the college and city of the college among a given list of similar addresses. The trend is also to select a general word ‘info’ to indicate that the e-mail address can be used for sending queries to the institution and obtaining the information there from. It is concluded that the average length for selecting username for institutional address is just above 7.5. The tendency of selecting special characters in designing username for institutional e-mail address is very less. Underscore, numeric digits and period are used in the design process with underscore being the most used special character, followed by digits and period. Overall usage of special characters including numeric digits for designing the usernames of e-mail addresses of engineering institutions being nearly 5%.

Our results are best reported on the dataset used which is not necessarily exhaustive. We do not appreciate or criticize the selection of a word for username in institutional e-mail address. We just present the statistics on trend of selecting words for designing usernames for institutional e-mail addresses by the Engineering institutes of Gujarat state of India. Moving on these lines, we conclude that these inferences should be true not only for engineering institutes but also for other institutes, as well. A clearly vast variation is observed in the design of username for institutional e-mail addresses. Further, the world is digitizing fast and the need of electronic communication cannot be underestimated in the modern paper-less world. Keeping these two points in focus, we strongly advocate that there should be a common format for institutional e-mail address that not only helps its institute to be identified uniquely but is also easy to interpret and remember.

References

[1] Calix K, Connors M, Levy, D, Manzar H et al. (2008), "Stylometry for E-mail Author Identification and Authentication", in *Proceedings of CSIS Research Day*, Seidenberg School of CSIS, Pace University, New York. May 2008.

[2] Education website (2010), "Literacy Rates For States And Union Territories", Available: http://www.education.nic.in/cd50years/g/z/7G/0Z7G0501.htm.

[3] Klensin J (2008), "Simple Mail Transfer Protocol", in *RFC5321*, Network Working Group, Standards Track, October 2008. Available: http://tools.ietf.org/html/rfc5321.

[4] Saini J R and Desai A A (2010), "A Textual Analysis of Digits Used for Designing Yahoo-group Identifiers", in *The IUP National Journal of Information Technology*, The ICFAI University Press, Hyderabad, Andhra Pradesh, India, Vol. 6, No. 2, pp. 34-42, June 2010. ISSN: 0973-2896.

[5] Saini J R and Desai A A (2010), "Structural Analysis of Username Segment in e-mail Addresses of MCA Institutes of Gujarat State", in *The IUP National Journal of Information Technology*, The ICFAI University Press, Hyderabad, Andhra Pradesh, India, Vol. 6, No. 3, pp. 43-50, September 2010. ISSN: 0973-2896.

[6] Saini J R and Desai A A (2010), "Structural Analysis of Username Segment in e-mail Addresses of MBA Institutes of Gujarat State of India", in *International Journal of Human and Social Sciences*, The World Academy of Science, Engineering and Technology (WASET), France, Vol. 5, No. 6, pp. 356-360, October 2010. ISSN: 1307-6892.

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

[8] Whereincity website (2010), "Gujrat – Information", Available: http://www.whereincity.com/india/gujrat/.

[9] Wikipedia e-mail (2010), "e-mail", Wikimedia Foundation Inc., Available: http://en.wikipedia.org/wiki/e-mail.

[10] Wikipedia e-mail_Address (2010), "e-mail Address", Wikimedia Foundation Inc., Available: http://en.wikipedia.org/wiki/e-mail_address.

[11] Wikipedia Gujrat (2010), "Gujrat", Wikimedia Foundation Inc., Available: http://en.wikipedia.org/wiki/Gujrat.

[12] Wikipedia List_of_Countries (2010), "List of Countries and Outlying Territories by Total Area", Wikimedia Foundation Inc., Available: http://en.wikipedia.org/wiki/List_of_countries_and_outlying_territories_by_total_area.

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