A Fuzzy Based Approach to Stylometric Analysis of Blogger's Age and Gender

Sumit Goswami
Indian Institute of Technology- Kharagpur
Kharagpur, West Bengal, India- 721 302
E-mail: sumit_13@yahoo.com

Mayank Singh Shishodia
School of Computer Science and Engineering
Vellore Institute of Technology
Vellore, Tamil Nadu, India -632 014
E-mail: mayanksshishodia@gmail.com

Abstract **-Fuzzy logic deals with partial truth. A fuzzy based approach to blog analysis, on the basis of various feature words, allows us to determine the degree to which a blogger's style belongs to a particular age or gender group. Each blog was represented by a set of normalized word frequencies of selected feature words in it. Using membership values obtained from applying Fuzzy C-Means (FCM) algorithm to these blog representations, we can call the blogger's style to belong weakly, fairly, strongly or very strongly to a particular class. The advantage of using fuzzy logic for this problem is that a weak belonging to a particular class means that there is a decent belonging to the other class (es). Hence when a search or query is carried out, no useful blog will be left out of the results for that other class (es).**

Keywords- fuzzy logic; stylometrics; blog; age; gender; fuzzy c-means; clustering

I. INTRODUCTION

The rise in popularity of social-networking can be attributed to the fact that the cost of fast Internet access has seen a steep decline over the past few years. A lot of free blogging websites have come up where anyone can register. This has made publishing blogs rather easy. Every blogger has a tendency to choose one style of writing over other thereby giving him/her a unique style [1] [2]. These styles vary in the content of the blog, average length of sentences, use of emoticons, slang words, abbreviations etc. There is a wonderful opportunity to extract information from the analysis of the huge number of blogs available on the Internet.

Jonathan Schler, Moshe Koppel, Shlomo Argamon, James Pennebaker have shown that the blogs differ according to the age and gender of their authors. They have called these differences to be "style-based features" and "content-based features" [3]. They have exploited the knowledge of these differences for carrying out automated author profiling. Sumit Goswami, Sudeshna Sarkar and Mayur Rastogi have further improved this by also considering "Non-dictionary words" and "Average sentence length" [4].

However, there was a common issue in the approach used in those methods. If the analysis of a blog yields that its author's writing characteristics belong almost equally to multiple age group or gender categories, then still classification is carried out in a crisp sense. Hence when a search for blogs belonging to a particular age group or gender category is carried out, that blog may be left out [6].The solution to this is a fuzzy based approach which we will discuss in this paper. We chose 'n'-features as was done in [3] [4] [5] and represented each blog as a set of 'n'- values denoting the word frequencies (per total length of the blog per 10000 words) of those corresponding features. We then applied Fuzzy C-Means (FCM) algorithm to cluster the blogs and then used pre-known information about the word frequency variation between blogs of different age groups and gender to name those clusters as 'Male' and 'Female' or '10s','20s' and '30s' depending upon whether the clustering was done for two , or three clusters. By using this method, we could also ascertain the degree to which a blogger's writing characteristics matched those of a particular age group or gender.

We chose 18142 blogs from the blog corpus on Prof Moshe Koppel's website which has a collection of blogs collected in 2004 from a blogging website called blogger.com. There were total 19320 blogs in the downloaded corpus. It was found that some blogs had highly spurious information about the author's age and gender. Such blogs were excluded from our analysis. In the forthcoming sections we explain the features that vary between blogs of authors belonging to different age groups and genders, the algorithm we have used and the results we got.

II. FEATURES

In this section, we present the features that vary significantly between blogs belonging to authors of different genders and age groups. We used a total of 352 words for gender estimation and 427 for age estimation. The features we've used are style-based features, content-based features [3] and non-dictionary words [4].

A. Style-based Features

It was observed that there is a higher frequency of pronouns, blog words and assent/negation words in female authored blogs while male authored blogs display higher occurrences of articles, prepositions and hyperlinks [3]. It was also noted that the differences in male and female blogging style are

978-1-4673-5116-4/12/$31.00 ©2012 IEEE

similar to older and younger blogging styles, or in other words, as people get older; their style of blogging becomes manlier [3][7].

B. Content Based Features

Female bloggers write more about "personal" stuff compared to male bloggers. Blogs of teenagers are about friends. As people get older, their interests change and this is observed in their blogs where they start talking about financial issues and politics. As noted in [3] usage of most words in blogs either increases or decreases monotonically.

C. Non-Dictionary Words

Slangs, emoticons, chat abbreviations etc. have found widespread usage in blogs as the blogs have no grammatical checks and almost all sort of language is acceptable. To express emotions, bloggers sometime even misspell words or extend/shorten them to impact readers differently [4].

Table 1 and Table 2 show the words more frequently used in female and male authored blogs respectively. Table 3, Table 4 and Table 5 show the words more frequently by blogger's in their 10's, 20's and 30's respectively.

TABLE 1: PARTIAL LIST OF WORDS USED MORE FREQUENTLY IN FEMALE AUTHORED BLOGS

know	people	think	Friends	Talk	Feelings	Care	Thinking	Friends	Relationship
Fat	Color	Wearing	Clothes	Body	Minutes	Shirt	Green	Coffee	Store
Shopping	Gone	Face	Hair	Times	Love	i	You	Yay	Wat
thats	Nite	Na	Hmmm	Hehe	Didn't	Cuz	Lol	Pink	Cute
Cried	Freaked	Gosh	Kisses	Yummy	Mommy	Boyfriend	Skirt	Adorable	Husband
Hubby	Fuckin	Don't	age	Im	Ish	Everytime	loved	Love	fat

TABLE 2: PARTIAL LIST OF WORDS USED MORE FREQUENTLY IN MALE AUTHORED BLOGS

Game	Games	Team	Win	Play	Won	Season	Beat	Final	Two
Club	Big	Straight	Site	Email	Page	Please	Website	Web	Post
Link	Urllink	Blog	Mail	Information	Free	Send	Comment	Internet	Online
Name	Services	List	Computer	Thanks	Update	Message	Bush	President	Iraq
An	The	Linux	Microsoft	Gaming	Server	Software	Gb	Programming	Google
Data	Graphics	India	Nation	Users	Economic	Money	Job	Tv	table

TABLE 3: PARTIAL LIST OF WORDS USED MORE FREQUENTLY BY BLOGGERS IN THEIR 10's

Tired	Wake	Eat	Watch	Dinner	Ate	Bed	Day	House	Early
Ran	Tried	Picked	Left	Im	Cool	Summer	Awesome	Lol	Stuff
Loved	Shit	Fuck	Sucks	Hate	Stupid	Drunk	Crap	Kill	Guy
Ass	Damn	Kid	Crazy	Music	Songs	Band	Cd	Rock	Listen
Show	Favorite	Radio	School	Teacher	Class	Study	Test	Finish	Students
Period	Paper	Pass	Maths	Homework	Bored	Sis	Boring	Awesome	Mum

TABLE 4: PARTIAL LIST OF WORDS USED MORE FREQUENTLY BY BLOGGERS IN THEIR 20's

Tomorrow	Tonight	Evening	Days	Afternoon	Weeks	Hours	July	Busy	Meeting
Hour	Month	Work	Working	Job	Trying	Right	Met	Figure	Meet
Start	Better	Try	Street	Place	college	Road	City	Walking	Trip
Headed	Front	Car	Beer	Apartment	Bus	Area	Park	Apartment	Building
dating	Area	Walk	Small	Places	Ride	Driving	Looking	Local	Sitting
Bar	Bad	Standing	Floor	Weather	Beach	View	Semester	Drunk	album

TABLE 5: PARTIAL LIST OF WORDS USED MORE FREQUENTLY BY BLOGGERS IN THEIR 30's

Years	Father	Mother	Children	Family	Kids	Parents	Old	Year	Child
Web	Post	Link	Check	Blog	Mail	Information	Free	Send	Commence
Using	Internet	Online	Name	Service	List	Computer	Add	Thanks	Update
Message	God	Jesus	Lord	Church	Earth	World	Word	Lives	Power
Human	Believe	Evil	Own	Truth	Thank	Peace	Speak	Bring	Truly
President	Iraq	American	States	America	Country	Government	National	News	State

III. ALGORITHM USED FOR AUTOMATED AUTHOR PROFILING

We used the FCM Algorithm [8] to cluster our blog vectors. We did this twice, once according to gender and then according to age.

For each blog i, b_i is a weight vector $<b_i^1, b_i^2, b_i^3..,b_i^n>$, where n is the size of feature set. The entries $b_i^1,..b_i^n$ represents the frequency of the various features normalized by the document length. The membership value of a blog in the i^{th} class has the following notation:

$$\mu_{ik} \in [0, 1]$$

with the restriction that the sum of all membership values for a single blog in all of the classes (n) has to be unity.

A. Gender Profiling

The various blogs are clustered into two. The degree of belonging to a particular cluster for a given blog is given by the corresponding entry in the fuzzy c-partition matrix (U). We already know author's gender and age from the author provided information. The two clusters are determined to be representing male or female gender according to whether they have more male authored blogs in them or female authored blogs. Once the clusters are known, the blogs' membership value in each of the clusters can be seen from U. If the blog has higher membership value for the cluster representing male authored blogs, then the author of the blog is predicted to be a male and vice-versa. These membership values are used for finding the accuracy of gender prediction of the authors of those blogs.

TABLE 6: CONFUSION MATRIX FOR GENDER ANALYSIS (USING MEMBERSHIP VALUE>0.50) [WEAKLY BELONG]

	Male	**Female**
Male	6002	2771
Female	3102	6267
Accuracy=67.62%		

TABLE 7: CONFUSION MATRIX FOR GENDER ANALYSIS (USING MEMBERSHIP VALUE>0.60) [MEDIOCRELY BELONG]

	Male	**Female**
Male	4804	1674
Female	2005	4902
Accuracy=72.51%		

TABLE 8: CONFUSION MATRIX FOR GENDER ANALYSIS (USING MEMBERSHIP VALUE>0.70) [FAIRLY BELONG]

	Male	**Female**
Male	3067	683
Female	673	2643
Accuracy=80.80%		

B. Age Profiling

The same approach is carried out as in Gender Profiling. Only difference is that here three clusters used represent the three age groups -10s ,20s and 30s.

IV. RESULTS AND ANALYSIS

One advantage of using fuzzy logic in our paper is that we can determine to which extent the blogger's style belongs to a particular gender. If we simply classify the blog into a class on basis of its greater membership value in it, we get an accuracy of 67.62%. This is shown in Table 6. Doing this on basis of membership value >0.60, we get an accuracy of 72.51% (Table 7). For 0.70 or higher membership value, the accuracy is 80.80% (Table 8) and for 0.80 or higher, its 83.96% (Table 9). The accuracy peaks for classification done for membership values of 0.84 or higher (Table 10), after this, the accuracy decreases and reaches 82.5% for 0.90. The blogs with membership values between 0.50 and 0.60 indicate that they belong almost equally to both classes. This is shown via an accuracy of 53.87% (Table 11).

For our age estimation, we get peak accuracy of 80.66% (Table 16) when we classify on basis of membership value 0.70 or higher. Without using any criteria, we get an accuracy of 55.39% (Table 12). As we increase the membership value criteria, the accuracy increases as in seen in Table 13 where we get 59.56% accuracy considering membership values of only over 0.40, Table 14 (66.86% for criteria of over 0.50), Table 15 (78.71% for criteria of over 0.60).

TABLE 9: CONFUSION MATRIX FOR GENDER ANALYSIS (USING MEMBERSHIP VALUE>0.80) [STRONGLY BELONG]

	Male	**Female**
Male	940	196
Female	94	579
Accuracy=83.96%		

TABLE 10: CONFUSION MATRIX FOR GENDER ANALYSIS (USING MEMBERSHIP VALUE>0.84) [VERY STRONGLY BELONG]

	Male	**Female**
Male	427	91
Female	30	221
Accuracy=84.26%		

TABLE 11: CONFUSION MATRIX FOR GENDER ANALYSIS (USING MEMBERSHIP VALUE BETWEEN 0.50 AND 0.60)

	Male	**Female**
Male	1198	1097
Female	1097	1365
Accuracy=53.87%		

TABLE 12: CONFUSION MATRIX FOR AGE GROUP ANALYSIS (USING MEMBERSHIP VALUE>0.30) [WEAKLY BELONG]

	10s	**20s**	**30s**
10s	4650	1984	1006
20s	1570	3832	2108
30s	206	1219	1567
Accuracy=55.39%			

TABLE 13: CONFUSION MATRIX FOR AGE GROUP ANALYSIS (USING MEMBERSHIP VALUE>0.40) [MEDIOCRELY BELONG]

	10s	**20s**	**30s**
10s	4523	1652	858
20s	1124	3701	1827
30s	127	1028	1521
Accuracy=59.56%			

TABLE 14: CONFUSION MATRIX FOR AGE GROUP ANALYSIS (USING MEMBERSHIP VALUE>0.50) [FAIRLY BELONG]

	10s	**20s**	**30s**
10s	3650	943	531
20s	682	2895	1020
30s	52	614	1208
Accuracy=66.86%			

TABLE 15: CONFUSION MATRIX FOR AGE GROUP ANALYSIS (USING MEMBERSHIP VALUE>0.60) [STRONGLY BELONG]

	10s	**20s**	**30s**
10s	2346	267	132
20s	236	1499	354
30s	11	198	586
Accuracy=78.71%			

TABLE 16: CONFUSION MATRIX FOR AGE GROUP ANALYSIS (USING MEMBERSHIP VALUE>0.70) [VERY STRONGLY BELONG]

	10s	**20s**	**30s**
10s	1312	108	62
20s	197	1001	147
30s	4	108	298
Accuracy=80.66%			

TABLE 17: CONFUSION MATRIX FOR AGE GROUP ANALYSIS (USING MEMBERSHIP VALUE>0.75)

	10s	**20s**	**30s**
10s	688	96	48
20s	99	544	118
30s	2	101	171
Accuracy=75.14%			

We achieved a peak accuracy of 84.26% (Figure 1) for gender estimation (considering membership values of only over 0.84) and 80.66% (Figure 2) for age estimation (considering membership values of only over 0.70). Our results were better than [3]. This can be attributed to our use of fuzzy logic as well as non-dictionary words. Our results were a bit poorer than [4]. This can be attributed to us using fewer words and not considering average sentence length in our paper.

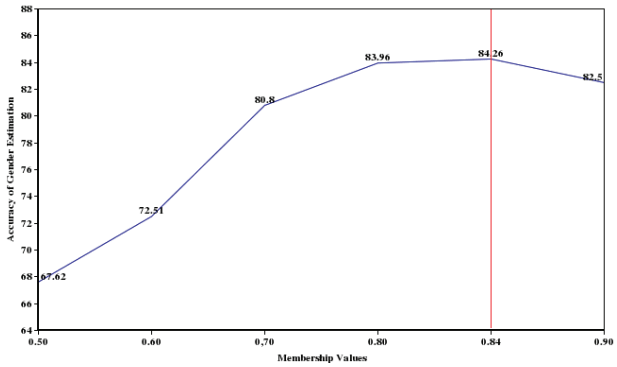

FIGURE 1: ANALYSIS OF GENDER ESTIMATION RESULTS

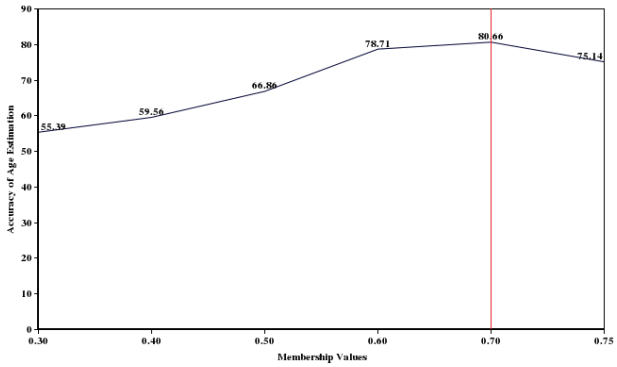

FIGURE 2: ANALYSIS OF AGE ESTIMATION RESULTS

V. CONCLUSION AND FUTURE WORK

Using fuzzy logic, we could determine the degree to which a blogger's style belonged to particular gender or age group. Using these degrees, we analyzed the accuracy of our results for various membership values. It was observed that the accuracy of our age and gender estimation was poor for low membership values but it increased as our membership value criteria rose to higher values like 0.80 (Gender) and 0.70 (Age).

To further improve our algorithm, we can add some fuzzy inference rules. Instead of just using the normalized word frequencies as such, we can use their respective membership values. The reason for this is that if blog belonging to a given gender or age group has an abnormally high word frequency for a feature of that same gender or age group, then its distance from the cluster center will be large. This will cause it to have a lower membership value for that class. Suppose, average Normalized Word Frequency (NWF) values of blogs in class C1 are 1.43 for a word X. For the other class C2, this value is 0.045. If one of the blogs, b, has a NWF value of 10.50 for the same word X, then it will have a lower membership value in that class C1. This will be fixed if we use membership functions where values of greater than 1.43 for the word X are assigned the value of 1. Moreover, use of a larger feature list,

average sentence length [4], author provided self-identification [3], themes and colors used [3], emoticons, blogger's blogging name will give better results.

REFERENCES

[1]. Gilad Mishne; "Information access challenges in the blogspace";IIIA 2006

[2]. James W. Pennebaker and Lori D. Stone. 2003. "Words of wisdom: language use over the lifespan"; Journal of Personality and Social Psychology, 85:291-301.

[3]. Jonathan Schler, Moshe Koppel, Shlomo Argamon, James Pennebaker; "Effects of age and gender on blogging"; AAAI Spring Symposium: Computational Approaches to Analysing Weblogs 2006: 199-205

[4]. Sumit Goswami, Sudeshna Sarkar, Mayur Rustagi: "Stylometric analysis of bloggers' age and gender"; ICWSM 2009

[5]. Koppel M, Argamon S, Shimoni A; "Automatically categorizing written texts by author gender", Literary and Linguistic Computing, 17(4), 401-412. (2003)

[6]. Vishal Gupta, Gurpreet S. Lehal; "A survey of text mining techniques and applications"; Journal of Emerging Technologies in Web Intelligence, Vol.1, No.1, August 2009

[7]. Argamon S, Koppel M, Pennebaker W and Schler J; "Mining the blogosphere: Age, gender and the varieties of self-expression."; First Monday. 12, 9 (September 2007).

[8]. Ross, T. J. (2010); "Fuzzy logic with engineering applications", Third Edition, John Wiley & Sons, Ltd, Chichester, UK