Proceeding Series of the Brazilian Society of Applied and Computational Mathematics, Vol. 5, N. 1, 2017.

Trabalho apresentado no CNMAC, Gramado - RS, 2016.

Proceeding Series of the Brazilian Society of Computational and Applied Mathematics

A Fuzzy Approach to Measure the Similarity Between Web Streaming Users

Sidnei Pereira Jr.[1]
Centro de Ciências Computacionais, Universidade Federal do Rio Grande, Rio Grande, Brazil
Graçaliz P. Dimuro[2]
Centro de Ciências Computacionais, Universidade Federal do Rio Grande, Rio Grande, Brazil
Eduardo N. Borges[3]
Centro de Ciências Computacionais, Universidade Federal do Rio Grande, Rio Grande, Brazil
Paula Fernanda Schiavo[4]
Centro de Ciências Computacionais, Universidade Federal do Rio Grande, Rio Grande, Brazil
Alex Camargo[5]
Centro de Ciências Computacionais, Universidade Federal do Rio Grande, Rio Grande, Brazil

Abstract. The increase of the Internet access and the popularity of mobile devices have influenced the consumption of radio/TV programs on the Web. An alternative for customization programming is the use of recommendation systems to adapt the content transmitted based on the preference of the listeners. The behavior of users accessing content on the Web is highly uncertain and naturally diffuse. In this paper, we propose an approach based on fuzzy set theory to analyze the similarity between users of Web radio programs, capturing similar interests from streaming data available in log files.

Keywords. Fuzzy Sets, Similarity Measure, Web Streaming

1 Introduction

The increase of Internet access and the popularity of mobile devices have influenced the consumption of radio and TV programs on the Web. This behavior created a new scenario for broadcasters because it allows to custom programming models. Usage profiles can be used for recommendation, for helping to structure web sites and for improving the user experience [9]. In this context, the similarity between users is extremely important. It can provide the foundation for organizing and identifying new user patterns [11].

An alternative for customization programming is using recommendation systems to adapt the content transmitted based on the preference of the listeners. The first step for implementing such systems is to identify the profile of the audience and the degree of

[1] sidnei.pereira@furg.br
[2] gracalizdimuro@furg.br
[3] eduardoborges@furg.br
[4] pfschiavo@furg.br
[5] alexcamargo@furg.br

DOI: 10.5540/03.2017.005.01.0410

interest in the content. For instance, a common feature used as the audience measure is the time that listeners remain connected to a particular program. Since this is a naturally vague information, it can be qualified in linguistic form as: little interest, some interest or much interest of the listener in the radio program. The theory of fuzzy sets is the ideal tool to deal with this kind of linguist variable. Since the introduction of fuzzy sets by Zadeh [12], measures of similarity between fuzzy sets have gained attention in many fields [10], such as image processing, pattern recognition, fuzzy reasoning [4,5]. Similarity concepts are a common term in classical set theory, as well as in statistics. [2]

In this paper, we propose a new approach to measure the similarity of Web streaming users, based on fuzzy set theory, capturing similar interests from streaming data available in log files. The time spent by users listening to a particular program were considered to calculate our similarity. The paper is organized as follows. In Section 2, we briefly present some related work on fuzzy similarity and introduce the proposed approach. The results are discussed in Section 3. Section 4 is the Conclusion.

2 A Fuzzy Similarity Measure Between Web Streaming Users

Ambiguity is the greatest challenge faced by systems that handle natural language. According to [3], identifying the real meaning of a given word can be so complicated, that it is sometimes only possible by consulting the user. Computing with vagueness and ambiguity is natural done by fuzzy systems. A fuzzy set is a set containing elements with membership grades varying in the unit interval $[0, 1]$. A fuzzy subset A of a universe X is determined by a membership function $\mu_A : X \to [0, 1]$, where $\mu_A(x)$ is the membership grade of the element $x \in X$ to the fuzzy subset A.

In recent decades, fuzzy similarity classification has got much attention. It has been considered as an important research area and different techniques are searched to design better categorization systems [6]. An interesting work is [1], where the proposal is a measure employed in a relational fuzzy clustering algorithm to discover clusters capturing the semantic information incorporated in website usage data. It was proposed a technique for mining Web usage profiles based on subtractive clustering that scales to large datasets.

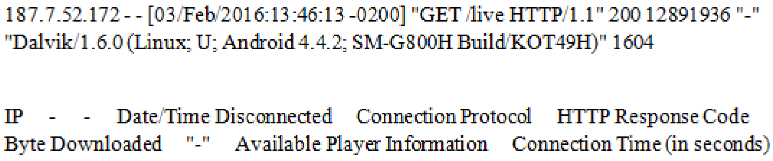
In the following, we introduce a fuzzy similarity measure, taking into account the time the users spend listening to a particular program. For this work we analyzed the logs of the Icecast[6], a free server software for streaming multimedia. Broadcasters divide their content at events with determined starting and ending time. Figure 1 shows an example of user access information collected from the streaming server log.

In the first phase, we use a technique to remove all the log content that is not meaningful (e.g., incomplete sessions, short connections). The interrupted and restored connections within the 5-minute interval is considered as a single session. An user session consists of an access originated from the same IP within a predefined time period [9].

After the preparation phase, the unique accesses for each user were separated, applied to a filter and organized in a matrix that relates each user u to the program p that he/she has been connected to (i.e., an user session), given by $T = [t_{u,p}]$. Each matrix entry $t_{u,p}$

[6] Stream Structure per Mount Point, available at http://www.icecast.org/docs.php.

DOI: 10.5540/03.2017.005.01.0410



```
187.7.52.172 - - [03/Feb/2016:13:46:13 -0200] "GET /live HTTP/1.1" 200 12891936 "-"
"Dalvik/1.6.0 (Linux; U; Android 4.4.2; SM-G800H Build/KOT49H)" 1604

IP  -  -  Date/Time Disconnected   Connection Protocol   HTTP Response Code
Byte Downloaded   "-"   Available Player Information   Connection Time (in seconds)
```

Figure 1: Icecast's access log with description of each information.

indicates the time duration a user has been connected to program p. The time duration of all programs have been converted to the base 3600 seconds (1 h) in order to maintain a single scale among all sessions. Algorithm 1 illustrates the pseudocode for filtering the sessions in each program.

Algorithm 1 Filtering the sessions in each program

```
if user_hr_start < prog_hr_start then                ▷ Did the user connect before the program started?
    if user_hr_end < prog_hr_end then                ▷ Did the user disconnect before the program ended?
        t = user_hr_end − user_hr_start
    else t = prog_hr_end − user_hr_start       ▷ Was the user still listening to the program after the end?
    end if
else
    if user_hr_end < prog_hr_end then            ▷ Did the user start listening after the program ended?
        t = user_hr_end − prog_hr_start
    else t = prog_hr_end − prog_hr_start            ▷ Was the user still listening the program after the end?
    end if
end if
```

In order to analyse the similarity between two users, we proposed an output based on fuzzy sets. Let U be a set of $\#U$ users and P a set of $\#P$ programs. It was assumed that the interest of a user $u \in U$ in a program $p \in P$ is directly related to the time T he/she stays connected to p. In this work, time was normalized to base 100 in $TN = [\frac{t_u * 100}{3600}, p]$ to calculate the interest I. Thus, the data of the matrix TN are processed in the matrix $I = [I_{little_{u,p}}, I_{some_{u,p}}, I_{much_{u,p}}]$, where I is characterized by linguistic terms that qualifies TN as equations (1), (2) and (3).

$$I_{little} = \begin{cases} \dfrac{50 - tn_{u,p}}{50} & \text{if } tn_{u,p} \in [0, 50] \\ 0 & \text{if } tn_{u,p} > 50 \end{cases} \tag{1}$$

$$I_{some} = \begin{cases} \dfrac{tn_{u,p}}{50} & \text{if } tn_{u,p} \in [0, 50[\\ \dfrac{100 - tn_{u,p}}{50} & \text{if } tn_{u,p} \in [50, 100] \end{cases} \tag{2}$$

$$I_{much} = \begin{cases} 0 & \text{if } tn_{u,p} < 50 \\ \dfrac{tn_{u,p} - 50}{50} & \text{if } tn_{u,p} \in [50, 100] \end{cases} \tag{3}$$

The matrix $I_{u,p}$ converts the user crisp *connection time* into fuzzy values, according to its membership grade to three fuzzy sets that qualify the fuzzy linguist variable *user interest*. We consider that the higher the value I_{little}, the lower the interest that a user can have in the content of a program (rule 1). However, if the value I_{some} is high, then it is assumed that the user has some interest in the program content (rule 2). At last, the higher the value I_{much}, the higher the user preference to the program content (rule 3). See Figure 2.

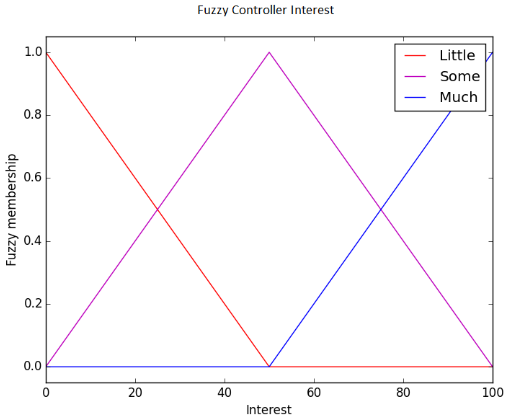


Figure 2: User interest in a particular program.

The output of the interest function is tuple $I_{u,p}$ composed by the 3 membership grades that form the profile of the user's interest, which is defined by equation (4).

$$\forall p = 1, \ldots, P : I_{u,p} = [I_{little_{u,p}}, I_{some_{u,p}}, I_{much_{u,p}}]. \tag{4}$$

We use the defuzzification based on centroid point to obtain the (crisp) numerical value after applying the T-conorm [8] of the maximum value in $I_{u,p}$. The similarity $S(u_1, u_2)$ between two users u_1 and u_2 varies in the closed interval $[0, 1]$ and it is computed by using the fuzzy similarity $S(u_1, u_2)$, given by equation (5) [9], where the returned values closer to 1 represent user profiles with very similar interests, and values closer to 0 represent little or no similarity.

$$S(u_1, u_2) = \frac{\Sigma_{p=1}^{\#P} \min \{I_{u_1,p}, I_{u_2,p}\}}{\Sigma_{p=1}^{\#P} \max \{I_{u_1,p}, I_{u_2,p}\}}, \tag{5}$$

3 Analysis of the Results

In order to evaluate our approach to measure the similarity between Web streaming users, we conducted an experiment using 800 unique users sessions distributed among 11 radio programs. The dataset was extracted from access logs in a range of 24 hours. We derived the degree of interest I for each user in each program and computed the similarity between all pair of users. In this paper, we present the case of three users u_A, u_B, u_C to validate the membership function and the fuzzy aggregation.

User u_A spent $t = 450$ seconds in the program p_1, enabling the rules 1 and 2, i.e., he or she belongs to the fuzzy subset I_{little} and I_{some}. Figure 3 shows the output of interest I on the membership function I_{llitle}, I_{some} and I_{much}. The fuzzy aggregation is presented

DOI: 10.5540/03.2017.005.01.0410

in Figure 4. We used the T-conorm of $max[I_{little} = 0.76, I_{some} = 0.24, I_{much} = 0.0]$, and defuzzification process based on centroid to reach the numerical value crisp $I = 34.47$.

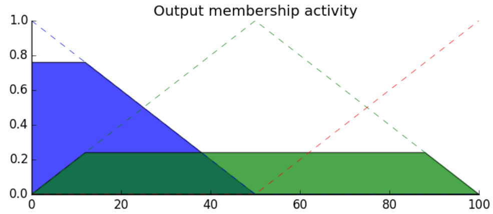


Figure 3: Connection time × interest (low)

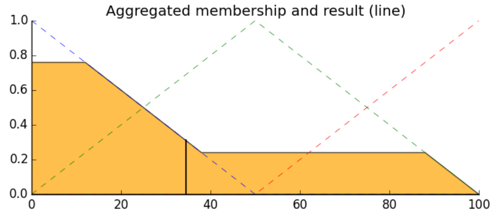


Figure 4: Aggregation relevance (low)

Figures 5 and 6 show the results for another user u_B, in a corresponding way to the figures presented above. $t_{u_B}, p_1 = 1400$, enabling rules 1 and 2 ($max[I_{little} = 0.23, I_{some} = 0.77, I_{much} = 0.0]$). $I = 48.51$. Lastly, $t_{u_C,p_1} = 3250$, enabling rules 2 and 3 ($max[I_{little} = 0.0, I_{some} = 0.19, I_{much} = 0.81$). $I = 68.77$ (Figures 7 and 8).

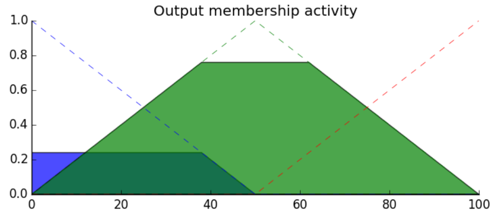


Figure 5: Access time × interest (medium)

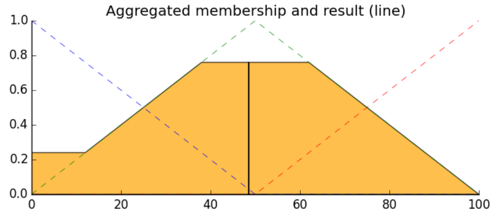


Figure 6: Aggregation relevance (medium)

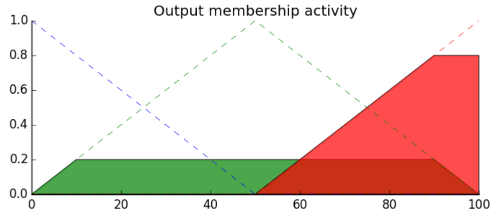


Figure 7: Access time × interest (high)

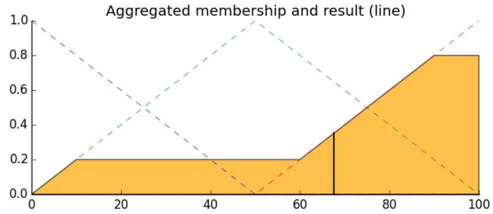
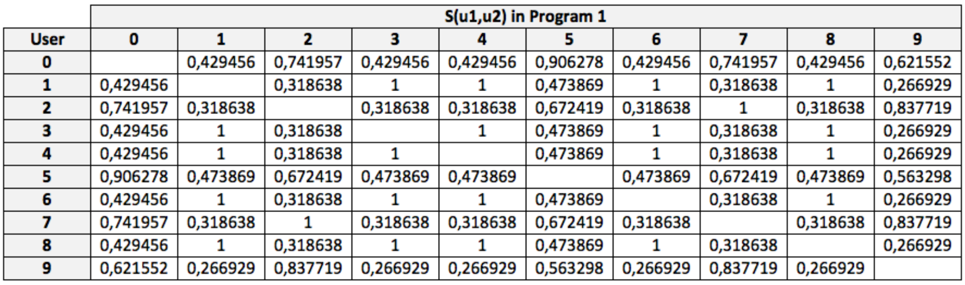


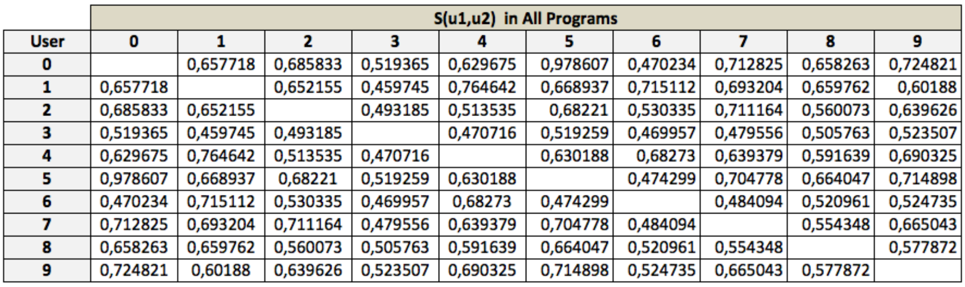
Figure 8: Aggregation relevance (high)

Finally, Figure 9 presents the matrix of the computed similarities between a sample of ten users listening to the program p_1. Users who did not listen or missed the program achieved maximum similarity equal to 1. These users have similar profile when compared to u_0 and u_5 as they heard only less than 600 seconds. However, $sim(u_2, u_7) = 1$ because they have the same connection time. Figure 10 refers to all day schedule, highlighting the heterogeneity in the users profile.



	S(u1,u2) in Program 1									
User	0	1	2	3	4	5	6	7	8	9
0		0,429456	0,741957	0,429456	0,429456	0,906278	0,429456	0,741957	0,429456	0,621552
1	0,429456		0,318638	1	1	0,473869	1	0,318638	1	0,266929
2	0,741957	0,318638		0,318638	0,318638	0,672419	0,318638	1	0,318638	0,837719
3	0,429456	1	0,318638		1	0,473869	1	0,318638	1	0,266929
4	0,429456	1	0,318638	1		0,473869	1	0,318638	1	0,266929
5	0,906278	0,473869	0,672419	0,473869	0,473869		0,473869	0,672419	0,473869	0,563298
6	0,429456	1	0,318638	1	1	0,473869		0,318638	1	0,266929
7	0,741957	0,318638	1	0,318638	0,318638	0,672419	0,318638		0,318638	0,837719
8	0,429456	1	0,318638	1	1	0,473869	1	0,318638		0,266929
9	0,621552	0,266929	0,837719	0,266929	0,266929	0,563298	0,266929	0,837719	0,266929	

Figure 9: Matrix of similarities (program p_1)



	S(u1,u2) in All Programs									
User	0	1	2	3	4	5	6	7	8	9
0		0,657718	0,685833	0,519365	0,629675	0,978607	0,470234	0,712825	0,658263	0,724821
1	0,657718		0,652155	0,459745	0,764642	0,668937	0,715112	0,693204	0,659762	0,60188
2	0,685833	0,652155		0,493185	0,513535	0,68221	0,530335	0,711164	0,560073	0,639626
3	0,519365	0,459745	0,493185		0,470716	0,519259	0,469957	0,479556	0,505763	0,523507
4	0,629675	0,764642	0,513535	0,470716		0,630188	0,68273	0,639379	0,591639	0,690325
5	0,978607	0,668937	0,68221	0,519259	0,630188		0,474299	0,704778	0,664047	0,714898
6	0,470234	0,715112	0,530335	0,469957	0,68273	0,474299		0,484094	0,520961	0,524735
7	0,712825	0,693204	0,711164	0,479556	0,639379	0,704778	0,484094		0,554348	0,665043
8	0,658263	0,659762	0,560073	0,505763	0,591639	0,664047	0,520961	0,554348		0,577872
9	0,724821	0,60188	0,639626	0,523507	0,690325	0,714898	0,524735	0,665043	0,577872	

Figure 10: Matrix of similarities (all programs in the 24 hours)

4 Conclusion

According to [1], Web server access logs contain substantial data about the accesses of users to a website. Hence, if properly exploited, the log data can reveal useful information. According to [7], the extraction of knowledge from a set of measured data will face both problems: the attributes of this data can be presented in linguistic terms, which will usually be subjective and vague, and a rough description of the data elements may be a cause for indiscernibility among some of them. The behavior of users accessing content on the Web is highly uncertain and naturally diffuse. Therefore, the fuzzy sets have been widely used in order to obtain similar groups.

This paper presented a approach of similarity measures based on access logs from a streaming server using fuzzy sets. As future work, we propose better qualify the interest adding the frequency that the listener returns to the stream, the program size as linguistic variable, the aggregation of the results of similarity by type of program and the possibility to identify the device that the user hears the streaming as a variable of interest.

Acknowledgment

Supported by CAPES and CNPQ (481283/2013-7, 306970/2013-9, 232827/2014-1).

References

[1] G. Castellano, A. M. Fanelli, C. Mencar, and M. A. Torsello. Similarity-based fuzzy clustering for user profiling. In *IEEE Int. Conf. on Web Intelligence and Intelligent Agent Technology Workshops*, pages 75–78, 2007. DOI: 10.1109/WI-IATW.2007.32.

[2] G. Deng, Y. Jiang, and J. Fu. Monotonic similarity measures between fuzzy sets and their relationship with entropy and inclusion measure. *Fuzzy Sets and Systems*, 287:97–118, 2016. DOI:10.1016/j.fss.2015.03.008.

[3] M. V. C. Guelpeli, A. C. B. Garcia, and F. C. Bernardini. *Emergent Web Intelligence: Advanced Semantic Technologies*, chapter An Analysis of Constructed Categories for Textual Classification Using Fuzzy Similarity and Agglomerative Hierarchical Methods, pages 277–306. Springer, 2010. DOI:10.1007/978-1-84996-077-9_11.

[4] G. Lucca, G. P. Dimuro, V. Mattos, B. Bedregal, H. Bustince, and J. A. Sanz. A family of choquet-based non-associative aggregation functions for application in fuzzy rule-based classification systems. In *IEEE International Conference on Fuzzy Systems*, pages 1–8, 2015. DOI: 10.1109/FUZZ-IEEE.2015.7337911.

[5] G. Lucca, J. Sanz, G. P. Dimuro, B. Bedregal, R. Mesiar, A. Kolesarova, and H. Bustince. Pre-aggregation functions: construction and an application. *IEEE Transactions on Fuzzy Systems*, PP(99):1–1, 2015. DOI: 10.1109/TFUZZ.2015.2453020.

[6] S. Puri and S. Kaushik. A technical study and analysis on fuzzy similarity based models for text classification. *International Journal of Data Mining & Knowledge Management Process*, 2(2):1–15, 2012. DOI: 10.5121/ijdkp.2012.2201.

[7] J. M. F. Salido and S. Murakami. Rough set analysis of a general type of fuzzy data using transitive aggregations of fuzzy similarity relations. *Fuzzy sets and systems*, 139(3):635–660, 2003. DOI: 10.1016/S0165-0114(03)00124-6.

[8] E. P. Klement, R. Mesiar, and E. Pap. *Triangular norms.* Kluwer Academic Publisher, Dordrecht, 2000.

[9] B. S. Suryavanshi, N. Shiri, and S. P. Mudur. An efficient technique for mining usage profiles using relational fuzzy subtractive clustering. In *Proceedings of the International Workshop on Challenges in Web Information Retrieval and Integration*, pages 23–29, 2005. DOI: 10.1109/WIRI.2005.7.

[10] Y. Tang and J. Zheng. Linguistic modelling based on semantic similarity relation among linguistic labels. *Fuzzy Sets and Systems*, 157(12):1662–1673, 2006. DOI: 10.1016/j.fss.2006.02.014.

[11] A. Tversky. Features of similarity. *Psychological review*, 84(4):327–352, 1977. DOI: 10.1037/0033-295X.84.4.327.

[12] L. A. Zadeh. Fuzzy sets as a basis for a theory of possibility. *Fuzzy sets and systems*, 1(1):3–28, 1978. DOI: 10.1016/0165-0114(78)90029-5.