E-ISSN 2581 –7957
P-ISSN 2277 – 3916

CVR Journal of Science and Technology, Volume 21, December 2021
DOI: 10.32377/cvrjst2105

Web Page Genre Identification and Categorization using Single-Label and Multi-Label Corpuses in English and Telugu Languages

K. Pranitha Kumari[1] and K. Srinivasa Reddy[2]
[1]Assoc. Professor, CVR College of Engineering/CSE Department, Hyderabad, India
Email: k.pranithakumari@cvr.ac.in
[2]Assoc. Professor, CVR College of Engineering/CSE Department, Hyderabad, India
Email: k.srinivasareddy@cvr.ac.in

Abstract: **As web is fluid, new web page genres emerge, and these genres are known as emerging genres. Genre based searches can yield better search results than topic based searches for the user. In this paper, Refined Adjustable Centroid Classification (RACC) algorithm is proposed to classify web page genres including emerging genres in multiple languages. Seven Telugu web page genres TELPOETRY, TELENTERTAINMENT, TELFAQ, TELCHILDREN, TELSOCIAL, TELETIQUETTE and TELE-GENRE are identified using the method of annotation by objective sources. Telugu web page genre corpus(10-genre) is developed which contains newly identified seven Telugu web page genres and three existing Telugu web page genres. The 7-genre, 10-genre, 20-genre, 23-genre and newly formed Telugu 10-genre corpuses are classified using RACC algorithm. The classification results obtained show that RACC algorithm gave better results when compared with existing classification techniques on all five corpuses. The experimental results obtained are statistically significant ($p<0.05$).**

Index Terms: **Telugu web page genres, Emerging genres, syllable extraction, genre threshold, web page genre classification, genre corpus, identification of genres**

I. Introduction

As the web is mutable, it is very hard to find required information from web pages retrieved by a search engine. To solve this problem, it is better to classify the web pages based on their genre instead of topic. Genre is a non-topical descriptor. It can be characterized by its type, functionality, content, form, discourse functionality, communicative purpose, layout and social acceptance. Genres normally have established names and are acknowledged by specific discourse communities.

To search the information using genres, it is necessary to collect web page genres. As web is fluid, new genres will come into existence called as emerging genres. Contemporary web page genres are fewer in number and do not include emerging genres [23]. Therefore, it is necessary to identify emerging genres from the web pages. So far, no proper methodology is developed to identify emerging genres.

Web page genre classification is an essential technique that can contribute to the progress of search results. Generally, classification techniques are used to find the correctness of the genre corpora [23]. Currently many classification techniques are available, but not all classification techniques are useful to classify the web page genre corpuses accurately. The accuracy of a classifier depends on the type of genre corpuses. Hence, a scalable technique is required to classify the web page genre corpuses that can accommodate emerging genres.

Usage of the internet in India is increasing day by day. Therefore, information in Indian languages on the web is also growing rapidly. So, there is a need to adopt genre based search for querying Indian language web pages. Until now, no one has identified a web page genre in the Indian language. Most of the Indian languages are syllabic languages in which each syllable is a glyph or a combination of glyphs and each glyph is represented with a unique code called Unicode [22].

This paper proposes a refined adjustable centroid classification technique which distinguishes genres correctly when compared with the well-known state of the art classification techniques on both single-label and multi-label corpuses in multiple languages, and a Telugu web page genre corpus with ten newly identified genres. The new 10-genreTelugu web page corpus is tested using different classification techniques.

The structure of the paper is as follows. Section 2 discusses the related work. Section 3 describes proposed Refined Adjustable Centroid Classification algorithm and experiments on web page genre classification. Section 4 discusses the process of identification of genres in Telugu web pages and classification results. Finally, Section 5 concludes the paper.

II. Related Work

A brief review of the reported work in literature towards the identification and classification of web page genres are presented in this section. With reference to a comprehensive survey of web page genres, there are two important issues in the area of research on web page genres: 1) genre evolution on the web and 2) the application of genre for classification of web pages on the web.

The Orators since Aristotle have initiated genres in the communication by considering subject, form, and purpose. The authors in genre analysis have given various definitions for genre such as: genre as a set of rules of discursive properties [32], genre is a social action [20], genre as a persuasive classifying statement [25], genre as a class of communicative events [31], genre as a typified

Received on 04-09-2021, Revised on 30-09-2021, Accepted on 23-11-2021.

CVR Journal of Science and Technology, Volume 21, December 2021
DOI: 10.32377/cvrjst2105

communicative action [36], genre is related to topic or subject [14]. In [16], genre is stated as a widely acknowledged class of texts determined by communicative purpose. The observations similar to [16] are given in [18] and they observed that most definitions of genre include the notions of document form, expected content, and intended communicative purpose, and the notion of social acceptance.

The study on web page genres is initiated in [4][5] by considering the objectives such as: the development of genres on the web and significant impact of genres on the web. A survey is carried out on genre evolution on the web in [29] and gave a special name to a genre on the web called as cybergenre. Web page genre identification mainly considers four factors such as granularity of web page genres, labels associated with each web page, assignment of genre labels and the characterization of genres. Among all four factors, "the characterization of genres" has more influence on the evolution of genres in various media. Three home page genres such as personal, corporate, and organizational home page genres from a collection of 321 websites are identified in [15]. Online encyclopedias web page genre is identified in [7]. BLOG is identified as an emerging genre in [2]. Corporate blogs is identified as an emerging genre [3] of computer-mediated communication from CBC/Corporati corpus by considering pragmatic aspects and linguistic aspects. Most recently in [17], a specialized genre named as Touristic Websites genre, which is a hybrid genre is identified. Rehm et al. constructed reference benchmark web page genres and are given in [24]. The limitations of existing work on genre identification are, there is no automatic framework for identification of emerging genres in web pages and no researcher focused on Indian language web page genre identification (i.e. genres are available in English, Spanish, German, Italian, Arabic and Russian language web pages, and are also very limited in number).

The classifier performance depends on type of the genre (single-label or multi-label), feature set to represent web page genre, and level of granularity of genre. The necessity of multi-label genre to a web page is described in [9][10][19][34]. In [28], they performed cross-testing experiments in web page genre classification on different web page genre corpuses such as: HGC, I-EN, KI-04, KRYS-I, MGC and 7-genre corpus [27][37][38][39] by considering a variety of features. In [33], a work on web page genre classification is performed on a set of 1539 web pages labeled manually and named as 20-genre corpus. In [27], three small sets of carefully chosen features are used to represent the web pages and introduced a new classification model called as an inferential model for classifying web pages by genre. In [12][13], they investigated cross classification testing on three web page genre corpuses (7-genre, KI-04 and 20-genre corpuses). The usefulness of combinations of content, form, and functionality features in web page genre classification is examined in [6]. As observed from the literature, there is no scalable classification method for emerging genres in web page genre classification which is applicable for both single-label and multi-label. A feature extraction method to extract genre specific features and language-independent features is required.

The features are extracted from Telugu web pages either directly from the text or by using OCR system. To work with machine learning techniques, Telugu web page features can be denoted in English. WX notation specifies a transliteration scheme to symbolize Indian Language scripts in English. As given in [1], the purpose of WX notation is to furnish a unique representation of Indian languages in English alphabets. A transliteration model for Indian languages to convert UTF to WX notation and WX notation to UTF was implemented in [8]. In [8], they conducted a study on Hindi, Telugu, Bengali, Punjabi, Malayalam and Kannada language documents. In [21], to locate and extract Telugu script in a document image, OCR system is built by taking into account the circular nature of Telugu script. In [11], a dictation system for Telugu speech recognition is built. During this process, they considered CIIL Mysore Telugu text corpus of 3 million words of running texts in Telugu for training. They also made rules for syllabification based on the canonical structure of Telugu script using WX notations but not directly from Telugu script.

III. Web Page Genre Classification

A. Refined Adjustable Centroid Classification

Classification is a method used to find the effectiveness of web page genre corpuses. A scalable classification technique is essential to classify web page genres. A Refined Adjustable Centroid Classification(RACC) algorithm is proposed in this paper to classify web page genres. RACC algorithm is the modification of ACC algorithm[23]. RACC algorithm is used to classify both single-label and multi-label web page genre corpuses which better represent a real world environment. The main idea of RACC algorithm is to find the similarity between each web page in test set and each web page genre profile in training set. RACC algorithm is described in Fig. 1 and it consists of two phases. In the training phase, for each web page genre, web page genre profile (centroid) is created. Web page profile contains n-gram features with their corresponding normalized weights. Web page genre profile is created by combining all unique n-gram features of all web page profiles belonging to that particular genre and the weight of each n-gram 'x' is the sum of the weights of 'x' in all the web pages. Genre threshold is also calculated in RACC. In test phase, cosine similarity between each web page and each web page genre profile is calculated and compared with genre threshold.

In case of single-label classification, among all genre similarities which are calculated, if the similarity between web page and web page genre is greater than that particular web page genre threshold among all highest the web page genre threshold then that web page genre is predicted as actual genre for that web page. In case of multi-label classification, if the similarity between web page and web page genre is greater than that particular web page genre threshold then that web page genre is predicted as actual genre for that web page. Here there is a possibility of predicting multiple genres for each web page. Setting genre threshold is important in multi-label classification. In RACC

CVR Journal of Science and Technology, Volume 21, December 2021
DOI: 10.32377/cvrjst2105

Input: Genre corpus

Output: Predicted genres

Notations: G: genre, Gc: genre corpus, exG: existing genre, WP: web page, Th: threshold, Nc: new corpus, Sim: cosine similarity, Tf: n-gram frequency, df: n-gram document frequency, Tw: n-gram weight, SimSum: sum of all the similarities, df: document frequency, MTh: modified threshold.

Training phase:

- For each genre exG ∈ Gc, Create genre profile exG_i
- For each web page WP ∈ exG_i, create web page profile WP_j
- For each n-$gram_{k,i}$ ∈ exG_i
  - $Tf_{k,i} \leftarrow Tf_{k,i} + Tf_{k,j}$
- For each n-$gram_{k,i}$ ∈ exG_i
  - Find $df_{k,i}$
  - if n-$gram_{k,i}$ exists in exG_i
  - $df_{k,i} \leftarrow df_{k,i} + 1$
  - $TW_{k,i} \leftarrow Tf_{k,i} * \log(|G_c|/df_{k,i})$
  - $Sim(WP, G) = \frac{\sum_{i=1}^{n} WP(tw_i) \times G(tw_i)}{\sqrt{\sum_{i=1}^{n} (WP(tw_i))^2} \times \sqrt{\sum_{i=1}^{n} (G(tw_i))^2}}$
  - $SimSum_i = SimSum_i + Sim(WP,G)$
- Find genre threshold

$$Th_i = \frac{Sim(WP_j, exG_i) + SimSum_i}{|G_i|}$$

Test phase:

- For each web page WP ∈ exG_i, create web page profile WP_j
- For each n-$gram_{k,i}$ ∈ exG_i,
  - For each n-$gram_{k,i}$ ∈ WP_j ,
  - Compute $Sim(WP_j, exG_i) \leftarrow \sum$n-gram∈(i∩j) $Tw_{k,i} * Tw_{k,j} / (\| Tw_{k,i} \| * \| Tw_{k,i} \|)$
- if $Sim(WP_j, exG_i) > Th_i$, p≠i (for single-label corpus)
  - WP_j predicted as actual genre exG_i with highest Th_i genre
- if $Sim(WP_j, exG_i) > Th_i$ (for multi-label corpus)
  - WP_j predicted as actual genre

Figure 1. RACC Algorithm

algorithm, the genre thresholds are genre specific thresholds instead of a fixed constant value.

B. Classification Experiments

The classification experiments are carried out on four corpuses: single-label 7-genre corpus and 10-genre corpus, multi-label 20-genre corpus and 23-genre corpus[23]. The performance evaluation measures used for single-label and multi-label classification experiments are described in [35] and [30] respectively. Performance measures such as precision, recall, F-measure and accuracy are used to evaluate the classification results on single-label genre corpuses Tan et al. [41] and Witten and Frank [35]. In this paper, macro-precision, macro-recall and macro F-measure[40] are used to evaluate the multi-label classifiers because macro-averages gives equal weight to each genre, whereas micro-averages gives equal weight to each web page. Binary relevance (BR) approach is followed during the multi-label classification. The advantage of the BR approach is its low computational complexity compared with other multi-label methods. Weka[42] and Meka [43] tools are used in the experimentation and implementations of SVM and Naive Bayes algorithms were developed by Witten and Frank [35]

The 7-genre corpus is classified using RACC, the results obtained are better when compared with ACC in terms of precision, recall and F-measure and are reported in Table 1. ANOVA test results show that the classification performance of RACC algorithm on single label 7-genre corpus is statistically significant in terms of F-measure ($p < 0.05$). The 10-genre corpus is evaluated using RACC, ACC and SVM classification algorithms and the results are reported in Table 2. All the three classification algorithms gave good results on 10-genre corpus but among the three, RACC algorithm performance is better. Based on ANOVA and paired t-test results, the performance of RACC algorithm is statistically significant than that of ACC and SVM in terms of precision, recall and F-measure ($p < 0.05$).

The RACC algorithm is used to classify multi-label 20-genre corpus. This multi-label classification task is performed by using twenty binary classifiers, each for a genre. The classification performance of RACC algorithm on the 20-genre corpus is compared with the existing work done on the same corpus and is shown in Table 3. The results obtained are statistically significant in terms of precision, recall and F-measure ($p < 0.05$).

TABLE I.
COMPARISON OF RACC ALGORITHM PERFORMANCE ON SINGLE-LABEL 7-GENRE CORPUS WITH ACC IN TERMS OF PRECISION, RECALL AND F-MEASURE

Genre	ACC			RACC		
	Precision	Recall	F-Measure	Precision	Recall	F-Measure
BLOG	1.0	0.97	0.9847	1.0	0.98	0.989
ESHOP	0.99	0.99	0.99	0.99	0.99	0.99
FAQS	0.99	1.0	0.9949	0.992	1.0	0.995
FRONTPAGE	1.0	1.0	1.0	1.0	1.0	1.0
LISTING	0.989	0.98	0.9845	0.989	0.99	0.99
PHP	0.966	0.995	0.98	0.97	0.995	0.985
SPAGE	0.985	0.985	0.985	0.989	0.988	0.989
Average	0.988	0.988	0.988	0.99	0.991	0.991

CVR Journal of Science and Technology, Volume 21, December 2021
DOI: 10.32377/cvrjst2105

TABLE II.
COMPARISON OF RACC ALGORITHM RESULTS WITH ACC AND SVM CLASSIFIERS ON SINGLE LABEL 10-GENRE CORPUS

Classifiers	A	P	R	FM
RACC	0.966	0.96	0.959	0.96
SVM	0.923	0.924	0.923	0.922
ACC	0.961	0.952	0.966	0.955

A = Accuracy, P = Precision, R = Recall, FM = F-measure.

TABLE III.
COMPARISON OF EXISTING WORK ON 20-GENRE CORPUS WITH RACC ALGORITHM RESULTS

Classifiers	A	P	R	FM
RACC (BR)	0.971	0.97	0.968	0.969
ACC(BR)	0.965	0.968	0.963	0.966
Mason	0.843	0.99	0.754	0.846

A = Accuracy, P = Precision, R = Recall, FM = F-measure.

TABLE IV.
COMPARISON OF RACC ALGORITHM RESULTS WITH ACC (BR) AND SVM (BR) CLASSIFIERS ON MULTI-LABEL 23-GENRE CORPUS

Classifiers	A	P	R	FM
RACC (BR)	0.955	0.969	0.948	0.957
SVM (BR)	0.817	0.871	0.953	0.91
ACC (BR)	0.947	0.964	0.948	0.953

A = Accuracy, P = Precision, R = Recall, FM = F-measure.

The classification experiment is also performed on multi-label 23-genre corpus. This multi-label classification task is performed by using twenty three binary classifiers, each for a genre. The classification performance of RACC algorithm on 23-genre corpus is compared with ACC and SVM (BR) in terms of precision, recall and F-measure and is shown in Table 4. The results obtained are statistically significant in terms of precision, recall and F-measure ($p < 0.05$).

IV. INTRODUCING AND CLASSIFYING GENRES IN TELUGU WEB PAGES

A. Syllable Extraction

The alphabets of different languages have unique codes that are present in the Unicode set. Unicode from 0C00 to 0C7F are allocated for Telugu syllables. Unicode Transformation Format (UTF) is the universal character code standard to represent character sets. UTF-8 is an alternative coded representation and maintains compatibility with ASCII. The web pages in Telugu are transliterated into an orthographic form using one of the standard forms which the machine can understand. WX notation is used to transliterate the Telugu text in UTF-8 to roman script and each Telugu syllable is represented using this notation. Character n-gram method is used to represent web pages. Character n-gram of length 'k' has 'k' syllables. To extract n-gram features of Telugu web pages it is necessary to extract syllables from web pages.

As Telugu is a syllabic language, syllables of Telugu web pages are extracted using Syllable extraction algorithm[22]. A Unicode based syllable extraction method is proposed[22] to extract syllables from Telugu web pages. Unicode sets defined for Telugu letters are used in Syllable extraction algorithm as follows: Vowel Unicode V = {0c05, 0c06, 0c07, 0c08, 0c09, 0c0a, 0c0b, 0c0c, 0c0e, 0c0f, 0c10, 0c12, 0c13, 0c14}. Consonant unicode set C = {0c15, 0c16, 0c17, 0c18, 0c19, 0c1a, 0c1b, 0c1c, 0c1d, 0c1e, 0c1f, 0c20, 0c21, 0c22, 0c23, 0c24, 0c25, 0c26, 0c27, 0c28, 0c2a, 0c2b, 0c2c, 0c2d, 0c2e, 0c2f, 0c30, 0c31, 0c32, 0c33, 0c35, 0c36, 0c37, 0c38, 0c39}. Halant h=0c4d. Special signs S={0c01, 0c02, 0c03}.Vowel signs VS={0C66, 0C67, 0C68, 0C69, 0C6A, 0C6B, 0C6C, 0C6D, 0C6E, 0C6F,0C78, 0C79, 0C7A, 0C7B, 0C7C, 0C7D, 0C7E, 0C7F}.

B. Introducing Telugu Web Page Genres

This paper focused on identifying web page genres in Telugu and introducing Telugu web page genre corpus. Telugu web page genre identification process consists of the following steps:

- If existing Telugu web page genre corpus is available then predict genre labels and apply IEG algorithm[23]to identify new Telugu web page genres.
- If there is no existing Telugu web page genre corpus then predict genre labels, collect web pages and based on the method of annotation by objective sources assign genre labels to the web pages.

During the identification process, a web page collection consisting of 100 web pages was downloaded in the year 2020 from different Telugu web sites. Telugu web page 3-genre corpus[22] is used and applied to IEG algorithm. Web pages are labeled in this corpus with a single-label using annotation by objective sources because it is faster and lesser complex compared to human annotation. Sevn genres TELPOETRY, TELENTERTAINMENT, TELFAQ, TELCHILDREN, TELSOCIAL, TELETIQUETTE and TELE-GENRE are identified from the Telugu web pages by considering the characteristics of web page genres such as communicative purpose and discourse functionality. These characteristics are considered because the users of the internet would share and communicate with each other through web pages. Table 5 shows the characterization of Telugu web page genres identified in this paper. A new Telugu web page genre corpus named as 10-genre corpus is shown in Table 6 is formed with 3-genre corpus and newly identified Telugu web page genres.

CVR Journal of Science and Technology, Volume 21, December 2021
DOI: 10.32377/cvrjst2105

TABLE V.
CHARACTERIZATION OF TELUGU WEB PAGE GENRES

Genre	Characteristics	References
TELPOETRY	Communicative purpose and discourse functionality	http://en.wikipedia.org/wiki/
TELENTERTAINMENT	Communicative purpose and discourse functionality	http://en.wikipedia.org/wiki/
TELFAQ	Communicative purpose and discourse functionality	http://en.wikipedia.org/wiki/
TELCHILDREN	Communicative purpose and discourse functionality	http://en.wikipedia.org/wiki/
TELSOCIAL	Communicative purpose and discourse functionality	http://en.wikipedia.org/wiki/
TELETIQUETTE	Communicative purpose and discourse functionality	http://en.wikipedia.org/wiki/
TELE-GENRE	Communicative purpose and discourse functionality	http://en.wikipedia.org/wiki/

TABLE VI.
SINGLE-LABEL TELUGU 10-GENRE CORPUS

Genre Name	Total web pages
TELARTICLE	50
TELBLOG	54
TELNEWS	51
TELPOETRY	98
TELENTERTAINMENT	100
TELFAQ	100
TELCHILDREN	100
TELSOCIAL	99
TELETIQUETTE	100
TELE-GENRE	98

C. N-gram Reresentation

Syllable n-grams are used to represent each web page. An example for extraction of syllable n-grams from the Telugu word ఉస్మానియా is shown in Fig. 2. The 3-gram features are extracted from all the web pages of 10-genre corpus. For each n-gram, n-gram weight is calculated and then CFS is applied to reduce the number of features.

E.g.: ఉస్మానియా
- Unicodes
  - 0c09 0c38 0c4d 0c2e 0c3e 0c28 0c3f 0c2f 0c3e
- Syllables
  - 0c09 0c380c4d0c2e0c3e 0c280c3f 0c2f0c3e
- N-grams
  - 0c090c380c4d0c2e0c3e0c280c3f
  - 0c380c4d0c2e0c3e0c280c3f0c2f0c3e
  - 0c280c3f0c2f0c3e0020
- ఉస్మాని
- స్మానియా
- నియా-

Figure 2. Extraction of syllable n-grams from Telugu word

D. Classification Experiments N Telugu Genre Corpus

The newly formed Telugu web page genre corpus is tested using different classification techniques. The classification results obtained in terms of accuracy, precision, recall and F-measure are shown in Table 7. Classification performance of RACC algorithm on Telugu web page genre corpus is tested using ANOVA, and it is observed that there is no statistically significant difference between the RACC, ACC, SVM and Naive Bayes classification techniques ($p < 0.05$). The reason is that most of the n-grams in the Telugu web page genres are unique.

TABLE VII.
COMPARISON OF RACC ALGORITHM RESULTS ON TELUGU 10-GENRE CORPUS WITH EXISTING CLASSIFICATION ALGORITHMS IN TERMS ACCURACY, PRECISION, RECALL AND F-MEASURE

Classifiers	A	P	R	FM
SVM	0.912	0.911	0.912	0.912
Naïve Bayes	0.90	0.93	0.92	0.925
ACC	0.98	0.972	0.978	0.974
RACC	0.987	0.99	0.986	0.985

A = Accuracy, P = Precision, R = Recall, FM = F-measure.

V. CONCLUSIONS

A scalable classification algorithm called Refined Adjustable Centroid Classification (RACC) is proposed for web page genre classification. RACC is the modification of ACC algorithm. The proposed algorithm is scalable because adding new genre to the classification model needs only the creation of a centroid for the new genre in the training phase. The proposed algorithm is used to classify different kinds of web page genre corpuses. On single-label 7-genre corpus, RACC algorithm performance in terms of F-measure is 99.1% when compared with ACC (98.8%). On single-label 10-genre corpus, RACC algorithm performance in terms of F-measure is 96.0% when compared with SVM (92.2%) and ACC (95.5%). On multi-label 20-genre corpus,

RACC algorithm performance in terms of F-measure is 96.9% when compared with ACC (96.6%) and Mason (84.6%). On multi-label 23-genre corpus, RACC algorithm performance in terms of F-measure is 95.7% when compared with SVM (91.0%) and ACC (95.3%). The experimental results suggest that RACC algorithm performance is better when compared with other classification algorithms on all the four corpuses. Statistical analysis is done using both ANOVA and paired t-test. The results obtained are found to be statistically significant ($p < 0.05$).

Syllable extraction algorithm is used to extract n-gram features from Telugu web pages. Seven new Telugu web page genres named as TELPOETRY, TELENTERTAINMENT, TELFAQ, TELCHILDREN, TELSOCIAL, TELETIQUETTE and TELE-GENRE are identified from randomly collected web pages based on annotation by objective sources and applying IEG algorithm with the use of existing 3-genre Telugu web page genre corpus. RACC algorithm performance in terms of F-measure is 98.5% on the newly formed 10-genre Telugu corpus when compared with SVM (91.2%), Naïve Bayes (92.5%) and ACC (97.4%). Statistical analysis results show that there is no statistically significant difference between the results of all the four algorithms. The reason is that the web pages in these genres are having unique 3-gram features. Overall, the ACC algorithm results are better when compared to the remaining three algorithms.

The work presented in this paper provides many directions for future extensions. This paper constructed ten Telugu web page genres, but still many genres are available in Telugu web pages. Constructing genre corpuses for Telugu web pages that contain more genres could be considered as a future work. A further possible extension could be to expand this work of identifying genres to other Indian languages. Future work could also include a study to develop an efficient and effective genre based search engine that supports genres in various orthographic spoken languages.

REFERENCES

[1] Bharati A., Chaitanya V., Sangal R. and Ramakrishnamacharyulu K.V., "Natural Language Processing: A Paninian Perspective", Prentice Hall of India, 1995.

[2] Blood R., "Weblogs: A History and Perspective", Rebecca's Pocket,07September2000.http://www.rebeccablood.net/essays/weblog_history.html.

[3] Cornelius P., "The corporate blog as an emerging genre of computer-mediated communication: features, constraints, discourse situation", Ph.D. thesis, 2010.

[4] Crowston K. and Williams M., "Reproduced and Emergent Genres of Communication on the World-Wide Web", In Proceedings of the 30th Hawaii International Conference on System Sciences (HICSS-30), IEEE Computer Society, pp. 30-39, 1997.

[5] Crowston K. and Williams M., "Reproduced and Emergent Genres of Communication on the World-Wide Web", The Information Society, vol. 16(3), pp. 201-216, 2000.

[6] Dong L., Watters C., Duffy J. and Shepherd M., "An Examination of Genre Attributes for Web Page Classification", In Proceedings of the 41st Hawaii International Conference on System Sciences (HICSS-41), IEEE Computer Society, 2008.

[7] Dott A. E., "An analysis of Wikipedia digital writing", Conference of the European Chapter of the Association for Computational Linguistics, 2006.

[8] Gupta R., Goyal P. and Diwakar S., "Transliteration among Indian Languages using WX Notation", In Proceedings of Semantic Approaches in Natural Language Processing, KONVENS 2010, Germany, September 2010.

[9] Jebari C. and ArifWani A., "A Multi-label and Adapative Genre Classification of Web Pages", In Proceedings of 11th International Conference on Machine Learning and Applications (ICMLA), 2012.

[10] Jebari C., "Enhancing the identification of web genres by combining internal and external structures", Pattern Recognition Letters, vol. 146, pp. 83-89, 2021.

[11] Kalyani N. and Sunitha K.V.N., "Syllable analysis to build a dictation system in Telugu Language", International Journal of Computer Science and Information Security (IJCSIS), vol. 6(3), 2009.

[12] Kanaris I. and Stamatatos E., "Webpage genre identification using variable-length character n-grams", In Proceedings of the 19th IEEE International Conference on Tools with Artificial Intelligence (ICTAI 2007), vol. 2, pp. 3-10, 2007.

[13] Kanaris I. and Stamatatos E., "Learning to Recognize Webpage Genres", Information Processing & Management, vol. 45(5), pp. 499-512, 2009.

[14] Karlgren J. and Cutting D., "Recognizing Text Genre with Simple Metrics Using Discriminant Analysis", Proceedings of the 15th International Conference on Computational Linguistics (COLING), Kyoto, Japan, 1994.

[15] Kennedy A. and Shepherd M., "Automatic Identification of Home Pages on the Web", In Proceedings of the 38th Hawaii International Conference on System Sciences (HICSS-38), IEEE Computer Society, 2005.

[16] Kessler B., Numberg G. and Shutze H., "Automatic Detection of Text Genre". Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and 8th Conference of the European Chapter of the Association for Computational Linguistics, Madrid, Spain, 1997.

[17] Koskensalo A., "Problems of LSP-Didactics Shown in Case of Specialized Genre of Touristic Websites", US-China Foreign Language, ISSN 1539-8080, Vol. 12, No. 7, 543-550, July 2014.

[18] Kwasnik B.H. and Crowston K., "Introduction to the Special Issue: Genres of Digital Documents", Information Technology & People, vol. 18(2), pp. 76-88, 2005.

[19] Mason J.E., "An n-gram Based Approach to the Automatic Classification of Web Page by Genre", Ph.D. thesis, Dalhousie University, Nova Scotia, 2010.

[20] Miller C., "Genre as Social Action". Quarterly Journal of Speech, vol. 70, pp.151-167, 1984.

[21] Negi A., Shanker K.N. and Chereddi C.K., "Localization, Extraction and Recognition of Text in Telugu Document Images", In Proceedings of the IEEE Seventh International Conference on Document Analysis and Recognition (ICDAR - 2003), 2003.

[22] Pranitha Kumari K., Venugopal Reddy A., "Syllable n-gram approach for Identification and Classification of genres in

Telugu language", in Proceedings of International Conference on Networks & Soft Computing, pp. 141-145, 19-20 August, 2014.

[23] Pranitha Kumari K., Venugopal Reddy A., "Identification and Classification of Emerging Genres in Web Pages", in Proceedings of International Conference on Computing and Communication Technologies, pp. 1-6, 11-13 December, 2014.

[24] Rehm G., Santini M., Mehler A., Braslavski P., Gleim R., Stubbe A., Symonenko S., Tavosanis M., and Vidulin V., "Towards a Reference Corpus of Web Genres for the Evaluation of Genre Identification Systems", In Proceedings of the 6th Language Resources and Evaluation Conference (LREC 2008), 2008.

[25] Rosmarin A., "The Power of Genre", University of Minnesota Press, 1985.

[26] Rosso M., "Using Genre to Improve Web Search", Ph.D. thesis, University of North Carolina, Chapel Hill (USA), 2005.

[27] Santini M., "Automatic Identification of Genre in Web Pages", Ph.D thesis, University of Brighton, 2007.

[28] Sharoff S., Zhili W. and Katja M., "The Web Library of Babel: evaluating genre collections", LREC, 2010.

[29] Shepherd M. and Watters C., "The Evolution of Cybergenres", In Proceedings of the 31st Hawaii International Conference on System Sciences (HICSS-31), IEEE Computer Society, vol. 2, pp. 97-109, Hawaii, USA, 1998.

[30] Sorower M.S., "A Literature Survey on Algorithms for Multi-label Learning", Oregon State University, Corvallis, vol. 18, pp. 1-25, 2010.

[31] Swales J., "Genre analysis", English in Academic and Research Settings. Cambridge University Press, Cambridge, UK, 1990.

[32] Todorov T., "Genres in Discourse", Cambridge University Press, Cambridge (UK) (translated by Catherine Porter from French, Editions du Seuil, 1978), 1990.

[33] Vidulin V., Lustrek M., and Gams M., "Training the Genre Classifier for Automatic Classification of Web Pages", In Proceedings of the 29th International Conference on Information Technology Interfaces, Journal of Computing and Information Technology, vol. 15(4), pp. 305-311, 2007.

[34] A. Onan, ''An ensemble scheme based on language function analysis and feature engineering for text genre classification,'' J. Inf. Sci., vol. 44(1), pp. 28–47, 2018.

[35] Witten I.H. and Frank E., "Data Mining: Practical Machine Learning Tools and Techniques", Morgan Kaufmann Publishers, Amsterdam, The Netherlands, 2nd edition, 2005.

[36] Yates J. and Orlikowski W., "Genres of organizational communication: A structural approach to studying communications and media", Academy of Management Review, vol. 17(2), pp. 229-326, 1992.

[37] K. Pranitha Kumari and A. Venugopal Reddy, "Performance improvement of web page genre classification," International Journal of Computer Applications, vol. 53(10), September 2012.

[38] K. Pranitha Kumari, A. Venugopal Reddy and S. Sameen Fatima, "Web page genre classification: Impact of n-gram lengths," International Journal of Computer Applications, vol. 88(13), February 2014.

[39] M. Wan, A. C. Fang, and C.-R. Huang, ''The discriminativeness of internal syntactic representations in automatic genre classification," Journal of Quantitative Linguistics, vol. 28(2), pp. 138-171, 2021.

[40] Godbole S. and Sarawagi S., "Discriminative Methods for Multi-labeled Classification", In Proceedings of the 8th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD 2004), Springer, pp. 22-30, 2004.

[41] Tan P.T., Vipin Kumar and Michael S., "Introduction to Data Mining", 2006.

[42] Meka machine learning workbench download site. http://sourceforge.net/projects/meka/.

[43] Weka machine learning workbench download site. www.cs.waikato.ac.nz/ml/weka/downloading.html.