Sentiment Progression based Searching and Indexing of Literary Textual Artefacts

Hrishikesh Kulkarni[1] and Bradly Alicea[2]

[1] Georgetown University, Washington DC, USA
[2] Orthogonal Research & Educational Laboratory, USA
hrishikeshparag@ieee.org

Abstract. Literary artefacts are generally indexed and searched based on titles, meta data and keywords over the years. This searching and indexing works well when user/reader already knows about that particular creative textual artefact or document. This indexing and search hardly takes into account interest and emotional makeup of readers and its mapping to books. In case of literary artefacts, progression of emotions across the key events could prove to be the key for indexing and searching. In this paper, we establish clusters among literary artefacts based on computational relationships among sentiment progressions using intelligent text analysis. We have created a database of 1076 English titles + 20 Marathi titles and also used database http://www.cs.cmu.edu/~dbamman/booksummaries.html with 16559 titles and their summaries. We have proposed Sentiment Progression based Search and Indexing (SPbSI) for locating and recommending books. This can be used to create personalized clusters of book titles of interest to readers. The analysis clearly suggests better searching and indexing when we are targeting book lovers looking for a particular type of books or creative artefacts.

Keywords: Literature, Creative Artifacts, Searching, NLP, Text Analysis, Machine Learning, Information Retrieval, Sentiment Mining

1 Introduction and Related Work

1.1 Searching and Indexing literature

Searching, recommending and indexing literary artefacts is generally driven by names of authors, topics, and keywords. This is very effective but very primitive method and cannot cope up with uncertainty and variations associated with user interests. It comes with its own advantages and challenges. But when we take into account millions of unknown titles; does this indexing in true sense gives us book titles best suited to our interest and emotional makeup? Actually, these primitive indexing mechanisms create bias while making some titles best seller and do injustice to many classic literary creations by unknown champions. This even refrains newcomers from creating novel literary experiments. This makes many worthy literary creations even sometimes vanish behind the curtains of brands created by this name and author-based system. While digital libraries are taking care of availability of books, we here propose an algorithm

for fair indexing and recommending literary artefacts. The joy and satisfaction of reading has more to do with plot, theme, progression of emotional upheavals than title or name of author. Indexing based on sentiment progression and thematic changes could help in dealing with bias of indexing and making justice to many titles those could not reach to book lovers. Availability, reachability and personalized indexing can help in solving this problem. This will reduce the bias, nurture creativity and bust the monopolies in the literary world.

1.2 Related Work

Indexing, listing and searching of books help readers in selecting the title of their interest. When we were focused on physical books – book catalogues are maintained alphabetically. These catalogues in digital platform were very useful for searching and locating a book if the title is known to a reader. These catalogues were extended with same paradigm of indexing for digital books. You can search even contents in these digital artefacts. But this basic paradigm of catalogues, indexing and searching has many limitations. You need to know a book if you are searching one. Keyword based search works effectively for scientific books but for fictions it fails miserably. The concepts of decoding character relationships for indexing and recommending is the core idea proposed in this paper. There are many attempts to decode relationships among characters. Decoding relationships among characters in narratives [1] can be considered as one of the core aspects while analyzing it. Relationship among characters at various places in a narrative is indicative of sentiment progression. These relationships can be modelled in different ways using semi-supervised machine learning [1]. Narrative structure is core to this analysis [2]. The role of sentiment in this pattern mapping is crucial to such relationship models [3]. Topic transition is generally determined using consistency analysis and coherence [4].

Linguistic perspective and contextual event analysis can play a vital role in narrative assessment. Surprises bring unexpected changes in relationships and event progressions, differentiating adorable events [5]. The overall narrative can be viewed as an emotional journey with variations in interestedness. This journey progresses as various relationships in the given narrative unfold. In the concept journey, different concepts are battling for existence and key concept prove to be ultimate winner. Emotional aspects blended in very personalized culture are at the helm of this journey. These emotional aspects are associated with part of stories or creative textual artefacts depicted through different impacting sentences [6]. In any of such scenarios decoding personality and culture with personality vector analysis prove to be effective for mapping [7, 8]. Researchers also used text-based analysis for clustering books [9]. The progression of relationships among characters in a narrative can be used for searching and indexing of books. This can take book catalogues beyond traditional limitations and hence searching can be possible based on progression of emotions, and interestedness of readers. This paper proposes ‘Sentiment Progression based Search and Indexing’ (SPbSI) to overcome limitations of traditional indexing approaches.

2 Sentiment Progression based Search and Indexing (SPbSI)

The proposed method is divided into four important phases:

1. Keyword-based core character identification and selection of pivot points
2. Sentiment progression analysis across pivot points
3. Derive similarity using 'Sentiment Progression Similarity Indicator' (SPSI)
4. Indexing and preparing catalogue for sentiment progression search

3 Data Analysis

A database of (1076 English + 20 Marathi) book titles from different genre is prepared and used for testing and learning. Another database used is http://www.cs.cmu.edu/~dbamman/booksummaries.html. One of the sample books we used as an example - titled "*Rage of Angels*" is a part of both of these datasets. The analysis on this data is preformed using SPbSI to determine relevance for indexing.

4 Mathematical Model

4.1 Core character identification and Pivot Point Selection:

Core characters are crucial to narrative and the story cannot progress without them. They are identified based on their frequency and relationships with other characters. To explain the concept, we have chosen two interesting fictions those were read by 50 out of 150+ book lovers from the BDB book club[1]: First one is '*Rage of Angels*' (https://en.wikipedia.org/wiki/Rage_of_Angels) published by Sidney Sheldon in 1980 titled & the second one is Marathi Classic *KraunchVadh* by V.S. Khandekar (https://en.wikipedia.org/wiki/Vishnu_Sakharam_Khandekar).

The algorithm to identify core characters is developed around core words and pivot points. Here core word is defined as a word that belongs to keyword set and has highest frequency of occurrence across the text space of interest. This word acts as a reference while creating cluster of words. Similarly, Core Character (CC) is one of the prime characters in narrative and is defined based on its presence and association with other prime characters. Equation 1 gives mathematical definition of CC.

$$\forall c \in c | C \in [CC] \; and \; c \rightarrow [CC] \; where \; [CC] \neq \Phi \qquad (1)$$

Going through narrative in an iterative fashion, the core characters are identified. The characters Jenifer, Michael and Adam are identified as core characters in *Rage of Angels*, while Sulu, Dilip, and Bhagvantrao in *Kraunch-Vadh*. Pivot point is a location in a narrative marked by intense interaction where we perform sentiment analysis.

[1] BDB Book Club is a major book club run by BDB India Pvt Ltd in Pune https://bdbipl.com/index.php/bdb-book-club/

4.2 Sentiment Progression Analysis across Pivot Points (PP)

The sentiment and emotional index at a pivot point using expressive word distribution is used to derive sentiment index. The progression of sentiment across these pivot points represents the behavior and nature of narrative. PP detection algorithm has identified 10 pivot points across the novel for characters Sulu and Dilip. Similarly, there are 8 pivot points for Sulu and Bhagwantrao. The detail algorithm SPbSI for indexing based on pivot point determination and sentiment association is given in algorithm 1.

Algorithm 1. Indexing for effective sentiment progression search

Task 1: Initialization (Determine Pivot Points)

1. $CC_{NAR} = \forall c \in c | C \in [CC]\ and\ c \rightarrow [CC]\ where\ [CC] \neq \Phi$

Task 2: Identify Pivot Point based on frequency, presence and Sentiment

2. $for\ \forall\ SJ_i,\ CC_i\ \ where\ CC\ ] \neq \Phi\ \&\&\ j \leq PP\_\max\ \ do$
3. $SI_i[SJ_iCC_i] = Sentiment(CC_i) + \alpha$
4. $With\ gradient\ decent\ determine\ local\ sentiment\ maxima$
5. $PP_j = Local\ Maxima$
6. $j++$
7. end

Task 3: Determine Sentiment Progression (based on Sentiment Value at Pivot Point (PP))

8. $for\ \forall\ PP\ do$
7. $if([SV@PP] \neq \Phi)$
8. $Insert\ SV\ in\ to\ series$

Task 4: Cluster formation of Sentiment Progression

9. $Sentiment\ Distanc\ SD = \frac{\sum_{i=1}^{n} CF(i)^2 \times N(i)}{\sum_{i=1}^{n} N(i)}$
10. $Sentiment\ Progression\ Similarity\ Indicator\ SPSI = \frac{1}{(1+ln\ (1+SD))}$
11. $for\ \forall\ Series\ \ IF\ \left[(Series_{Length} <> M)\right]$
12. $Make\ length\ of\ series = M$
13. $for\ \forall\ Series\ \ where\ [Series] \neq \Phi\ \&\&\ Series\ length = M$
14. $Create\ SPSI\ Matrix$
15. $for\ \forall\ Series\ \ where\ [Series] \neq \Phi\ \&\&\ \exists(\ SPSI) > DT$
16. $IF\ [(SPSI(i,j) > Dynamic\ Threshold\ DT)]$
17. $Combine\ Seires\ (i,J)$
19. $endif$
20. end

End of Algorithm

The extraction of sentiment (emotional positivity and negativity in this case) with reference to context of story includes pivot point identification, extracting sentiment at a particular event. These sentiments are progressed from one pivot point to next one. Thus, Model' (SPbSI) identifies sentiment progression from one pivot point to another.

4.3 Derive Similarity Using Sentiment Progression Similarity Indicator

Statistically Sentiment Progression Similarity Indicator (SPSI) gives behavioral similarity between two sentiment progression patterns. Every pivot point has a sentiment value. Hence every book has a sentiment progression series and can be represented as

a data series. Let's take two creative textual artefacts at a time and get corresponding two sentiment progression series. It is highly likely that these two series will have different number of pivot points. Hence, we add supporting points so that both series have equal number of elements. Hence series will look like:

$$RS \ni S(i) = S_1(i) + S_2(i) \tag{5}$$

Here, RS is a series derived by summing corresponding pivot point sentiment values. This is used to calculate the probable value for corresponding series. The probable sentiment value (PS) in accordance with sentiment progression is determined using Eq 6.

$$PS = \frac{\sum_{i=1}^{n} S_1(i)}{\sum_{i=1}^{n} S(i)} \tag{6}$$

PS is used to derive expected sentiment value with assumption that sentiment progression in second series is same. It us used to calculate correction factor CF.

$$CF(i) = \frac{PS \times RS(i) - S_1(i)}{\sqrt{RS(i) \times PS \times (1-PS)}} \tag{7}$$

The sentiment distance SD between two text artefacts is given by Eq 8.

$$SD = \frac{\sum_{i=1}^{n} CF(i)^2 \times N(i)}{\sum_{i=1}^{n} N(i)} \tag{8}$$

Here N is normalization factor $N(i) = \sqrt{R(i)}$. The SPSI is calculated using Eq 9

$$SPSI = \frac{1}{(1+ln\,(1+SD))} \tag{9}$$

SPSI will drop slowly with increase in sentiment distance. It is close to 1 for patterns those look alike & approaches to zero for completely different patterns.

4.4 Indexing and Preparing Catalogue for Effective Sentiment Progression Search

Iteratively similarity between sentiment progression of every pair of creative textual artifacts is calculated. This leads to SPSI matrix. The diagonal of this matrix is always 1. Two series with maximum similarity are combined to reduce the (n X n) matrix to (n-1 X n-1) and so on. This process continues till the similarity between all representative patterns is less than dynamic threshold. This process results in getting clusters with representative sentiment progression patterns. A Progression Similarity Matrix for nine books is depicted in Fig 1.

Here out of these 9 series during first iteration series (7, 9) are combined. Further the resultant series is combined with series 3, then with series 6 and later with series 5. Thus a representative series is formed for cluster made up of series (3, 5, 6, 7, 9). Similarly, series (1, 4) are combined and that series is combined with series 8. Thus, a cluster is formed of series (1, 4, 8). Thus, at the end of iteration 1 the matrix will be of size 3 X 3, with members representing cluster (3, 5, 6, 7, 9), cluster (1, 4, 8) and (2).

Series	1	2	3	4	5	6	7	8	9
1	1	0.32	0.11	0.76	0.36	0.16	0.15	0.62	0.11
2	0.32	1	0.18	0.22	0.31	0.28	0.14	1	0.23
3	0.11	0.18	1	0.16	0.58	0.54	0.73	0.25	0.41
4	0.76	0.22	0.16	1	0.37	0.26	0.39	0.57	0.25
5	0.36	0.31	0.58	0.37	1	0.49	0.52	0.16	0.53
6	0.16	0.28	0.54	0.26	0.49	1	0.66	0.11	0.50
7	0.15	0.32	0.73	0.39	0.52	0.66	1	0.29	0.86
8	0.62	0.14	0.25	0.57	0.16	0.11	0.29	1	0.15
9	0.11	0.23	0.41	0.25	0.53	0.50	0.86	0.15	1

Figure 1: Progression Similarity Matrix.
Thus, each representative cluster pattern is converted in to an index point.

4.5 Handling Unequal Length Pivot Point Data Sets

Handling unequal length data series is the most challenging aspect of this method. To deal with this, we distributed pivot points based on its distribution across the book for the shorter length data series. The biggest gap is filled with interpolation first. This process is continued till the length of two data series becomes same. The 30% length difference can be handled with this method.

5 Experimentation

5.1 Baselines

Creative artifacts are generally indexed and catalogued using titles or author names. In some very special cases support is provided using metadata and keywords. Thus, indexing in the past performed using two different approaches [17]. In the first approach it is based on metadata, author names, genre and titles. In the second approach textual similarity is used across the complete text or on the summary. The first approach is developed as the baseline-1 while the second one is developed as baseline-2.

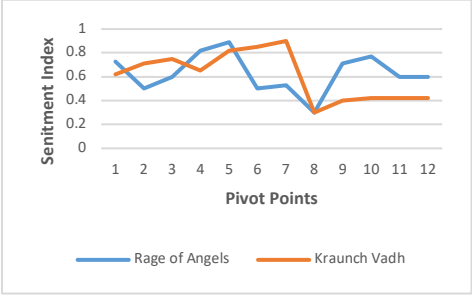


Fig. 2. Comparison of Sentiment progression

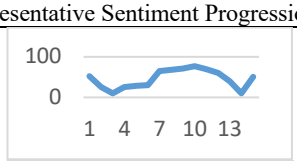
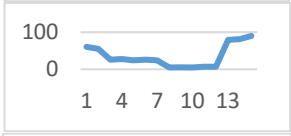
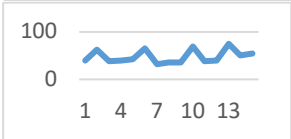
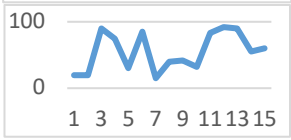
5.2 Results

Table 1 gives sentiment indices (SI) at normalized pivot points for fictions *Rage of Angels* and *KraunchVadh*. The Sentiment Progression Similarity Indicator (SPSI) between these two fictions is 0.649149. Fig. 2 depicts sentiment progression.

Table 1. Pivot Point Mapping and Ranking across the fiction (Normalized values)

Pivot Points	SI *RageofAngels*	SI *KraunchVadh*	Pivot Points	SI *RageofAngels*	SI *KraunchVadh*
1	0.73	0.62	7	0.53	0.9
2	0.5	0.71	8	0.3	0.3
3	0.6	0.75	9	0.71	0.4
4	0.82	0.65	10	0.77	0.42
5	0.89	0.82	11	0.6	0.42
6	0.5	0.85	12	0.6	0.42

Table 2. Sentiment Progression based Indexing

Sr. No.	Representative Sentiment Progression	Example Books	Remarks
1		*Rage of Angels, KraunchVadh,* etc	Middle portion creates higher sustained positive emotions with peak at pivot point 11
2		*Nothing lasts forever, Amrutvel*, etc	Slowly leads to lower point and there are surprises and excitement towards end
3		*Alchemist, Rikama Devhara*, etc.	Distributed surprises excitement leading to multiple spikes across the narrative
4		*Kite Runner, Five Point Someone,* etc.	Begins with excitements and multiple spikes and followed by sudden negativity with some positive resolve towards end

Books from database (http://www.cs.cmu.edu/~dbamman/booksummaries.html) are used for experimentation. We needed complete text of the book, hence the number of samples used for experimentation are kept limited. The response of 25 book lovers is compiled for analysis of outcome. Total 100 top books are indexed and catalogued using SPbSI. This outcome is compared with results from baseline algorithm where book lovers look for a book of a particular type. The representative behavioral patterns are depicted in Table 2. 100 books are used for Indexing. This indexing based on clusters is verified using inputs from book lovers. Out of these 100 books for 92 all the book

lovers were in agreement of indexing. On the other side for baseline-1 only 55 book-lovers endorsed the outcome. Baseline-2 based on text similarity could find 66 books indexed as per expectations of the book lovers. The behavioral pattern and indexing are depicted in Table 2. Around 38% improvement could be obtained for the given set of data using SPbSI over the baseline-1 and 26% over baseline-2. Thus, readers look for sentiment progression rather than metadata related to book. Though the sample size is small one, it is representative of overall similarity.

6 Conclusion

Indexing and searching narratives and creative textual artefacts using author names and titles comes with its own challenges. While searching narrative based on features and behaviors, sentiment progression could prove to be a valid alternative to traditional way of indexing. Book similarity in terms of reader preferences depends on progression of sentiment. This paper proposed an approach of indexing and searching of books based on 'Sentiment Progression based Searching and Indexing' (SPbSI). The results are analyzed with reference to data collected from 25 book lovers, but the method may be scaled to the analysis of thousands of candidates. The proposed algorithm gives around 26% improvement over the base line algorithm. The algorithm SPbSI can further be improved using moving window-based similarity approach which can make possible even to recommend certain part of a narrative or creative artifact to readers.

References

1. Iyyer, M., Guha, A., Chaturvedi, S., Boyd-Graber, J., & Daume, H. Feuding families and former friends: unsupervised learning for dynamic fictional relationships. ACL, San Diego, California (2016)
2. Chaturvedi, S., Srivastava, S., Daume, H., & Dyer, C. Modeling evolving relationships between characters in literary novels. *AAAI*, Phoenix, Arizona. (2016)
3. Li, J., Jia, R., He, H., & Liang, P. Delete, retrieve, generate: a simple approach to sentiment and style transfer. ACL, New Orleans, Louisiana. (2018)
4. Lund, J. Fine-grained Topic Models Using Anchor Words. Dissertation. Brigham Young University, Provo, Utah. (2018)
5. Oard, D.W. & Carpuat, M. et.al Surprise Languages: Rapid-Response Cross-Language IR. *ACM NTCIR-14 Conference*, June 10, 2019 Tokyo Japan. (2019)
6. Quan, C. and Ren, F. Selecting clause emotion for sentence emotion recognition. *International Conference on Natural Language Processing and Knowledge Engineering*, Tokushima, Japan. (2011)
7. Kulkarni, H. & Alicea, B. Cultural association based on machine learning for team formation. arXiv, 1908.00234 (2019)
8. Kulkarni, H. & Marathe, M. Machine Learning Based Cultural Suitability Index (CSI) for Right Task Allocation. *IEEE International Conference on Electrical, Computer and Communication Technologies* (IEEE ICECCT), Coimbatore, India. (2019)
9. N. Spasojevic and G. Poncin, Large Scale Page-Based Book Similarity Clustering, *2011 International Conference on Document Analysis and Recognition*, Beijing, 119-125, doi: 10.1109/ICDAR.2011.33 (2011)