International Journal of Computer Science Engineering & Technology
Approved by
National Science Library (NSL)
National Institute of Science Communication and Information Resources (NISCAIR)
Council of Scientific and Industrial Research (CSIR)- New Delhi. India .
ISSN :2455-9091

# DISCOVERING USAGE PATTERN AND LATENT FACTOR WITH PROBABILISTIC LATENT SEMANTIC ANALYSIS

[1] PADMA PRIYA .G, [2] Dr. M. HEMALATHA
[1] PhD Research Scholar, [2] Research Supervisor
[1& 2] Bharathiar university,
[1, 2] Coimbatore. Tamil Nadu India.

**ABSTRACT:** With regards to Web utilization mining, one imperative undertaking is to uncover inherent client navigational patterns and a latent assignment space. Such sort of use information can be found by a wide scope of factual techniques, AI and information mining algorithms. Among these procedures, LSA dependent on a likelihood derivation approach is a promising worldview which can't just uncover the basic relationships covered up in Web co-event perceptions, yet in addition recognize the latent assignment factor related with use information. In this paper we plan to present a Probabilistic Latent Semantic Analysis (PLSA) model to produce Web client gatherings and Web page groups dependent on latent use analysis.

**Keywords:** [Semantic, Pattern, Probabilistic, Latent Factor.]

## 1. INTRODUCTION

The PLSA show has been right off the bat exhibited and effectively connected in content mining rather than standard LSI algorithms, which use the Frobenius standard as an advancement paradigm, PLSA demonstrate depends on a most extreme probability guideline, which is gotten from the vulnerability hypothesis in insights. Fundamentally, the PLSA display depends on a measurement show called viewpoint demonstrate, which can be used to recognize the covered up semantic connections among general co-event exercises. The PLSA demonstrate has been right off the bat exhibited and effectively connected in content mining rather than standard LSI algorithms, which use the Frobenius standard as an enhancement foundation, PLSA display depends on a greatest probability rule, which is gotten from the vulnerability hypothesis in insights. Fundamentally, the PLSA display depends on a measurement show called angle demonstrate, which can be used to recognize the covered up semantic connections among general co-event exercises. Hypothetically, we can adroitly see the client sessions over Web pages space as co-event exercises with regards to Web utilization mining, to construe the latent use pattern. Given the viewpoint show over the client get to pattern with regards to Web use mining, it is first expected that there is a latent factor space every co-event perception information session is related with the factor(for example the visit of a page by a shifting degree to k the perspective of angle show, it tends to be induced that there do exist distinctive connections among Web clients or pages comparing to various factors. Moreover, the distinctive factors can be considered to speak to the comparing client get to patterns. For instance, amid a Web utilization mining process on an internet business site, we can characterize that there exist k latent factors related with k sorts of navigational standards of conduct, for example, z factor representing having interests in games explicit item class, 1sale item intrigue. As indicated by z for perusing through an assortment of item pages in various z … and so on,. Thusly, every co-event perception information may pass on client navigational enthusiasm by mapping the perception information into a dimensional latent factor space. The degree, to which such relationship is "clarified" by each factor, is determined by a contingent likelihood circulation related with the Web use information. Therefore, the objective of utilizing the PLSA demonstrate is to decide the restrictive likelihood circulation, thus, to uncover the natural connections among Web clients or pages dependent on a likelihood induction approach. In single word, the PLSA show is to demonstrate and derive client navigational practices in a latent semantic space, and distinguish the latent factors related. Before we propose the PLSA based algorithm for Web utilization mining, it is important to present the numerical foundation of the PLSA show, and the algorithm which is utilized to gauge the restrictive likelihood appropriation.

## 2. PROPOSED WORK

### 2.1 Latent Factor with PLSA

As such, for each factor, there might exist an undertaking focused client get to pattern relating to it. We, in this way, can use the class-restrictive likelihood gauges created by the PLSA model to deliver the amassed client profiles for portraying client navigational practices. Thoughtfully, each amassed client profile will be communicated as a gathering of pages, which are joined by their comparing loads demonstrating the commitments to such client bunch made by those pages. Besides, dissecting the created client profile can prompt uncovering normal client get to interests, for example, prevailing or optional "topic" by arranging the page loads. We allot client sessions into the comparing bunches which can be considered to speak to client navigational patterns dependent on the determined restrictive likelihood conveyances from the PLSA demonstrate and describe the portrayals of the client profiles regarding weighted page vector too. As examined above, it very well may be seen that a specific client session does have a place with only one bunch, yet additionally to other diverse groups related with various latent factors. For instance, a client session may show diverse interests (with various probabilities) on two perspectives can be "clarified" as that a client may, in fact, perform distinctive assignments amid a similar session and truly mirror the idea of client get to patterns in genuine world. It very well may be suggested, thus, the PLSA display parcels client session-page sets, which is unique in relation to bunching either client sessions or pages or both. At the end of the day, the client session-page probabilities in the PLSA demonstrate reflect "overlay" of latent factors, while the customary bunching model expect there is only one group explicit dissemination contributed by all client sessions in the cluster.

[Algorithm]: Characterizing latent semantic factor

[Input]: A set of conditional probabilities, P p z, a predefined thresholdμ. ( )j k

[Output]: A set of latent semantic factors represented by a set of dominant pages.

Step 1: Set PCL PCL PCL f= = = =_ ,
Step 2: For each then construct1 2 kz , choose all Web pages such that k PCL p PCL= ∪ ,
Step 3: If there are still pages to be classified, go back to step 2, PCL PCL= .
Step 4: Output { }k

## 3. EXPERIMENTAL RESULTS

### INTENSITY RATIO

| Existing 1 | Existing 2 | Proposed |
|---|---|---|
| 22 | 7 | 35 |
| 27 | 15 | 39 |
| 35 | 20 | 44 |
| 44 | 22 | 56 |
| 51 | 26 | 60 |

**Table 1: Comparison table of Intensity Ratio**

The examination table 1 clarifies about the power proportion of existing technique and proposed strategy. The power proportion of existing 1 is least 22 and most extreme 51, the force proportion of existing 2 is least 7 and greatest 26 and the power proportion of proposed technique is least 35 and greatest 60. It is accepted that the power proportion of proposed strategy is greatly improved to deal with information when contrasted with existing proportion.

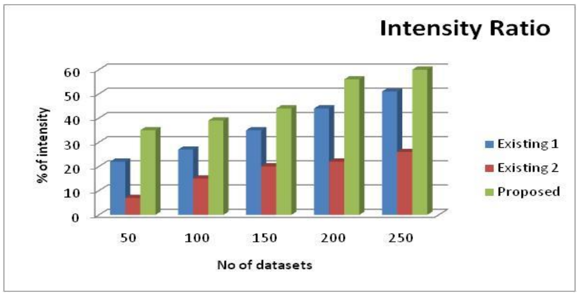

**Figure 1: Comparison graph of Intensity Ratio**

The figure 1 clarifies the correlation of power proportion in rate and ascertaining number of datasets dealt with. The force proportion of existing 1 is least 22 of every 50 datasets and most extreme 51 of every 250 number of datasets, the power proportion of existing 2 is least 7 out of 50 datasets and greatest 26 out of 250 number of datasets and the force proportion of proposed strategy is least 35 out of 50 datasets and greatest 60 out of 250 datasets. It is expected that the power proportion of proposed technique is greatly improved to deal with information when contrasted with existing proportion.

### Accuracy

| Existing 1 | Existing 2 | Proposed |
|---|---|---|
| 55 | 74.7 | 86.7 |
| 60 | 76.1 | 88.6 |
| 65 | 77.7 | 89.8 |
| 70 | 79.1 | 90 |

**Table 2: Comparison table of Accuracy**

The correlation table 2 clarifies the precision proportion of existing technique and proposed strategy. The proportion of existing strategy 1 is 55 to 70 while taking care of 20 to 80 number of datasets. The proportion of existing technique 2 is 74.7 to 79.1 while dealing with 20 to 80 number of datasets. In any case, in proposed strategy the proportion of exactness level is 86.7 to 90 while dealing with 20 to 80 number of datasets. While looking at the proportion of existing strategies and proposed technique the precision of proposed strategy has high proportion in taking care of information precisely.

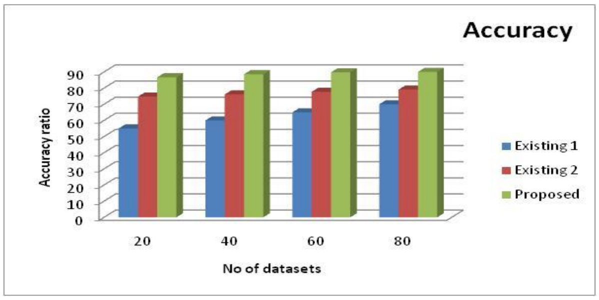

**Figure 2: Comparison graph of accuracy**

The examination figure 2 clarifies the precision proportion of existing and proposed strategy. The chart demonstrates the precision of the two techniques by the proportion it handles by number of datasets. The proportion of existing technique 1 is 55 to 70 while dealing with 20 to 80 number of datasets. The proportion of existing technique 2 is 74.7 to

79.1 while dealing with 20 to 80 number of datasets. Yet, in proposed technique the proportion of exactness level is 86.7 to 90 while taking care of 20 to 80 number of datasets. While contrasting the proportion of existing techniques and proposed strategy the exactness of proposed strategy has high proportion in taking care of information precisely.

## Cost Evaluation

| Existing 1 | Existing 2 | Proposed |
|---|---|---|
| 39 | 26.77 | 66 |
| 45 | 31.98 | 72 |
| 49 | 34.56 | 76.5 |
| 55 | 38.92 | 79.8 |
| 58 | 44.56 | 85 |

**Table 3: Comparison table of processing cost evaluation**

The examination table 3 clarifies the preparing cost assessment of existing techniques and proposed strategy. The cost assessment proportion of existing 1 is from 39 to 58 and cost assessment proportion of existing 2 is from 26.77 to 44.56. In proposed technique the cost assessment preparing proportion is from 66 to 85. It is expected that the procedure of cost assessment proportion is more in proposed technique while looking at other existing strategy.

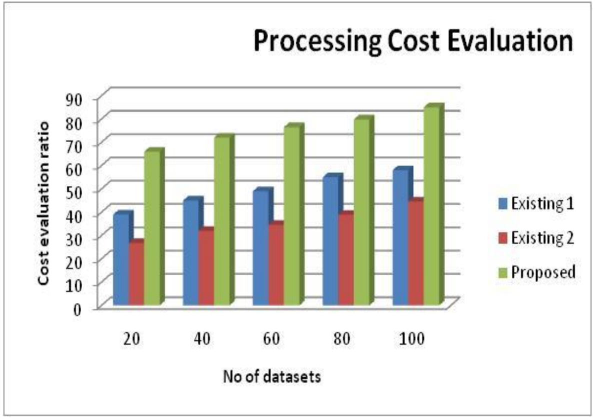

**Figure 3: Comparison graph of processing cost evaluation**

The correlation figure 3 clarifies the preparing cost assessment of existing techniques and proposed strategy by their proportion while taking care of number of datasets. The cost assessment proportion of existing 1 is from 39 to 58 while dealing with 20 to 100 datasets and cost assessment proportion of existing 2 is from 26.77 to 44.56 when taking care of 20 to 100 datasets. In proposed strategy the cost assessment preparing proportion is from 66 to 85 when taking care of 20 to 100 datasets. It is expected that the procedure of cost assessment proportion is more in proposed technique while looking at other existing strategy.

## Processing time evaluation

| Existing 1 | Existing 2 | Proposed |
|---|---|---|
| 57 | 69.5 | 83 |
| 59 | 69.9 | 84.8 |
| 62 | 69.5 | 87.9 |
| 66 | 70.9 | 90.2 |
| 69 | 72 | 93.6 |

**Table 4: Comparison table of processing time evaluation**

The examination table 4 clarifies the preparing time assessment of existing strategies and proposed strategy. The assessment time proportion of existing 1 is 57 to 69 when taking care of 20 to 100 datasets, the assessment time proportion of existing 2 is 69.5 to 72 when dealing with 20 to 100 datasets. Be that as it may, in proposed strategy the assessment time proportion is 83 to 93.6 when dealing with 20 to 100 datasets. It is clearly demonstrated that the assessment time of preparing information is more in proposed strategy when contrasted with other existing technique.

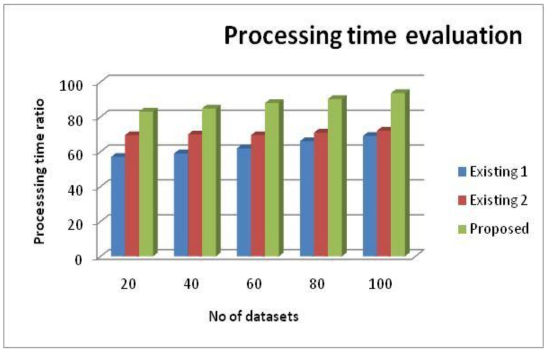

**Figure 4: Comparison graph of processing time evaluation**

The examination diagram 4 clarifies the handling time assessment of existing technique and proposed strategy. The proportion of time assessment of two techniques have been accepted by dealing with number of datasets. . The assessment time proportion of existing 1 is 57 to 69 when dealing with 20 to 100 datasets, the assessment time proportion of existing 2 is 69.5 to 72 when taking care of 20 to 100 datasets. Be that as it may, in proposed technique the assessment time proportion is 83 to 93.6 when taking care of 20 to 100 datasets. It is clearly demonstrated that the assessment time of preparing information is more in proposed strategy when contrasted with existing strategies.

## Absolute Effectiveness

| Existing 1 | Existing 2 | Proposed |
|---|---|---|
| **0.09** | 0.04 | 0.13 |
| **0.14** | 0.08 | 0.2 |
| **0.19** | 0.13 | 0.28 |
| **0.25** | 0.19 | 0.39 |
| **0.3** | 0.22 | 0.45 |

**Table 5: Comparison table of absolute effectiveness**

The correlation table 5 clarifies the total adequacy of existing strategies and proposed technique. The supreme viability proportion of existing 1 is 0.09 to 0.3 and the outright adequacy proportion of existing 2 is 0.04 to 0.22, the total adequacy of proposed strategy is 0.13 to 0.45. The supreme proportion of proposed strategy is high when contrasted with existing technique.

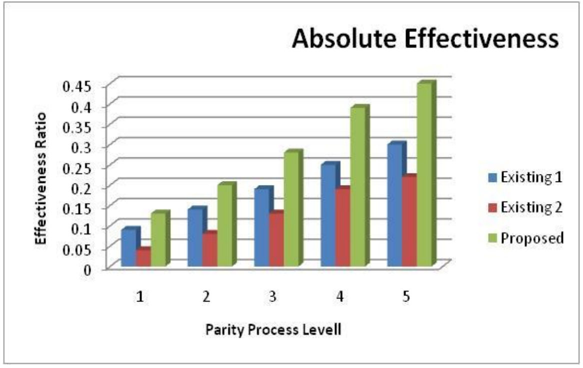

**Figure 5: Comparison graph of absolute effectiveness**

The correlation diagram 5 clarifies the outright viability proportion of existing strategies and proposed technique. The supreme adequacy proportion of existing 1 is 0.09 to 0.3 in 1 to 5 equality procedure level and the total viability proportion of existing 2 is 0.04 to 0.22 in 1 to 5 equality procedure level, the outright adequacy of proposed strategy is 0.13 to 0.45 in 1 to 5 equality procedure level. The outright proportion of proposed technique is high when contrasted with existing strategy.

## Survial Probability

| Existing 1 | Existing 2 | Proposed |
|---|---|---|
| **0.73** | 0.41 | 0.8 |
| **0.73** | 0.47 | 0.83 |
| **0.83** | 0.53 | 0.85 |
| **0.78** | 0.55 | 0.89 |

**Table 6: Comparison table of survival probability**

The examination table 6 clarifies the survival likelihood of existing strategy and proposed technique. The survival likelihood of existing 1 is 0.73 to 0.78 in 1 to 4 test dimension , in existing 2 the survival likelihood proportion is 0.41 to 0.55 in 1 to 4 test dimension. In any case, in proposed level the survival likelihood proportion is 0.8 to 0.89 in 1 to 4 test dimension. It is accepted that the survival likelihood proportion of proposed strategy appears to be high and better when contrasted with existing techniques.

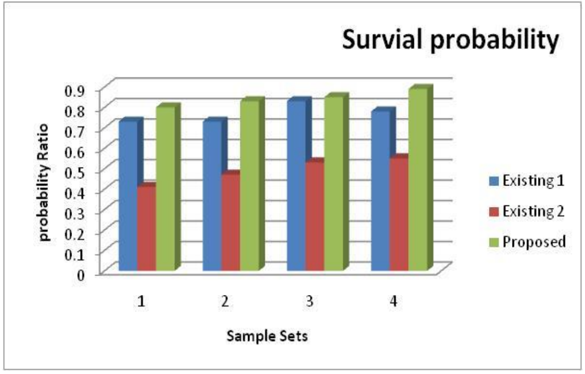

**Figure 6: Comparison graph of survival probability**

The correlation diagram 6 clarifies the survival likelihood of existing technique and proposed strategy. It is recognized by the proportion of likelihood and test sets of both the techniques. The survival likelihood of existing 1 is 0.73 to 0.78 in 1 to 4 test dimension , in existing 2 the survival likelihood proportion is 0.41 to 0.55 in 1 to 4 test dimension. In any case, in proposed level the survival likelihood proportion is 0.8 to 0.89 in 1 to 4 test dimension. It is expected that the survival likelihood proportion of proposed strategy appears to be high and better when contrasted with existing techniques.

## CONCLUSION

Proposed a Probabilistic Latent Semantic Analysis (PLSA) demonstrate, which can construe the covered up semantic factors and reveal client get to patterns from the session-page perception information. We began with presenting the hypothetical foundation of PLSA display. The inspiration driving of this model is on a premise of a presumption that every co-event perception is related with a lot of latent viewpoints or undertakings, whose degrees could be resolved from a likelihood deduction process. We have proposed a LSI-based methodology, named LUI, for gathering Web exchanges and creating client profiles. By utilizing PLSA demonstrate, the latent factor space and client profiles have been effectively uncovered by utilizing algorithms of bunching pages and client sessions dependent on the evaluations of contingent likelihood circulation. The tests on two true informational collections have been led to assess the viability of the proposed strategy. The test results have demonstrated that the latent factors can be literarily construed dependent on the translation of the semantic factor space. Furthermore, the client get to patterns have additionally been described by the client profiles, which are communicated in the types of weighted page sets.

## REFERENCES

[1] ShahrukhTeli, PrashastiKanikar " A Survey on Decision Tree Based Approaches in Data Mining", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 5, Issue 4, 2015.

[2] Jehad Ali, Rehanullah Khan, Nasir Ahmad, Imran Maqsood," Random Forests and Decision Trees", IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 5, No 3, September 2012.

[3] Amit Gupte, Sourabh Joshi, Pratik Gadgul, Akshay Kadam," Comparative Study of Classification Algorithms used in Sentiment Analysis", International Journal of Computer Science and Information Technologies, Vol. 5 (5), 2014.

[4] UCI Machine Learning Repository, " http://archive.ics.uci.edu/ml/machine-learning-databases/credit-screening/ ".

[5] Liu, S., Xia, C., and Jiang, X., ―Efficient Probabilistic Latent Semantic Analysis with Sparsity Control‖, IEEE International Conference on Data Mining, 2010, 905-910.

[6] Bassiou, N., and Kotropoulos C. ―RPLSA: A novel updating scheme for Probabilistic Latent Semantic Analysis‖,

Department of Informatics, Aristotle University of Thessaloniki, Box 451 Thessaloniki 541 24, Greece Received 14 April 2010.

[7] Romberg, S., Hörster, E., and Lienhart, R., ―Multimodal pLSA on visual features and tags‖, The Institute of Electrical and Electronics Engineers Inc., 2009, 414-417.

[8] Wu, H., Wang, Y., and Cheng, X., ―Incremental probabilistic latent semantic analysis for automatic question recommendation‖, ACM New York, NY, USA, 2008, 99-106.

[9] Blei, D.M., Ng, A.Y., and Jordan, M.I., ―Latent Dirichlet Allocation‖, Journal of Machine Learning Research, 3, 2003, 993-1022.

[10] Ahmed,A., Xing,E.P., and William W. ―Joint Latent Topic Models for Text and Citations‖, ACM New York, NY, USA, 2008.

[11] Zhi-Yong Shen,Z.Y., Sun,J., and Yi-Dong Shen,Y.D., ―Collective Latent Dirichlet Allocation‖, Eighth IEEE International Conference on Data Mining, pages 1019–1025, 2008.

[12] Porteous, L.,Newman,D., Ihler, A., Asuncion, A., Smyth, P., and Welling, M., ―Fast Collapsed Gibbs Sampling For Latent Dirichlet Allocation‖, ACM New York, NY, USA, 2008.

[13] McCallum, A., Wang, X., and Corrada-Emmanuel, A., ―Topic and role discovery in social networks with experiments on enron and academic email‖, Journal of Artificial Intelligence Research, 30 (1), 2007, 249- 272.

[14] Bao, S., Xu, S., Zhang, L., Yan, R., Su, Z., Han, D., and Yu, Y., ―Joint Emotion-Topic Modeling for Social Affective Text Mining‖, Data Mining, 2009. ICDM _09. Ninth IEEE International Conference, 2009, 699-704.

[15] Kakkonen, T., Myller, N., and Sutinen, E., ―Applying latent Dirichlet allocation to automatic essay grading‖, Lecture Notes in Computer Science, 4139, 2006, 110-120.