भारतीय वैज्ञानिक एवं औद्योगिक अनुसंधान पत्रिका
वर्ष 28 अंक (1) जून 2020 पृ. 19-24

लेम्पेल-ज़िव कंप्रेशन तथा लॉन्गेस्ट कॉमन सबसीक्वेन्स एल्गोरिद्म के संयोजन से यात्रा समय के अनुक्रम हेतु पूर्वानुमान की माप

अजय कुमार एवं रवीन्द्र कुमार
सीएसआईआर-केंद्रीय सड़क अनुसंधान संस्थान, नई दिल्ली 110 025

सारांश : यह आमतौर पर माना जाता है कि यात्रा समय की विश्वसनीयता एक महत्वपूर्ण कारक है, क्योंकि यह न केवल यात्री के व्यवहार पर महत्वपूर्ण प्रभाव डालता है बल्कि परिवहन प्रणालियों की सामान्य दक्षता पर भी काफी प्रभाव डालता है। यात्रा के समय अनुक्रम के लिए पूर्वानुमान की माप को संभावना एक दृष्टिकोण है जो ऐतिहासिक डेटा पर निर्भर करता है। पूर्वानुमानशीलता के दृष्टिकोण से यात्रा समय अनुक्रम की विश्वसनीयता को मापने के लिए एंट्रॉपी एक महत्वपूर्ण भूमिका निभाता है। यात्रा समय अनुक्रम की एंट्रॉपी का अनुमान लगाने के लिए लेम्पेल-ज़िव डेटा कम्प्रेशन एल्गोरिद्म का उपयोग किया जा सकता है, लेकिन लेम्पेल-ज़िव एल्गोरिद्म बिल्कुल समान उप-स्ट्रिंग के लिए पूरे ऐतिहासिक स्ट्रिंग को स्कैन करेगा, इसलिए एंट्रोपी आकलन प्रक्रिया में कुछ आंशिक रूप से समान उप-स्ट्रिंग या उप अनुक्रम की अनदेखी हो जाती है। गहन विश्लेषण के लिए, लॉन्गेस्ट कॉमन सबसीक्वेन्स (एलसीएस) एल्गोरिद्म बिल्कुल समान उप-स्ट्रिंग और समान उप-अनुक्रम को भी गणना में जोड़ने का बेहतर तरीका है। पूर्वानुमान गणना में एलसीएस एल्गोरिद्म, कुछ अतिरिक्त उप-अनुक्रम को छोड़ देता है जो आगे एंट्रॉपी को कम करता है। इससे यात्रा समय अनुक्रम के पूर्वानुमान की माप में वृद्धि होती है। इस प्रकार, एलजेड और एलसीएस एल्गोरिद्म का संयोजन पूर्वानुमान की माप में एलजेड एल्गोरिद्म से अधिक सटीक गणना करता है और दिए गए यात्रा समय अनुक्रम के आधार पर अगली यात्रा समय के लिए पूर्वानुमान दर की गणना करने के लिए आसानी से आईटीएस (ITS) सम्बन्धित उपकरण के लिए भी प्रयोग में लाया जा सकता है।

Predictability measurement for the sequence of travel time by combining the Lempel-Ziv compression and the Longest Common Subsequence algorithm

Ajay Kumar & Ravindra Kumar
CSIR-Central Road Research Institute, New Delhi 110 025

Abstract

It is commonly recognized that reliability in travel time is a significant factor to consider, as it not only wields a significant impact on traveller behaviour, but also substantially impacts the general efficiency of transport systems. Predictability measure for Travel time Sequence is probabilistic approach that depends on historical data. To measure the reliability of travel time sequence from the perspective of predictability, entropy plays an important role. To estimate the entropy of such a travel time sequence Lempel-Ziv data compression algorithm can be used but Lempel-Ziv scanned the whole historical string for the exactly identical substring. So, in entropy estimation process some partially similar substring or subsequence gets overlooked. For deep analysis, longest common subsequence (LCS) algorithm is employed to calculate exactly identical substring and most similar subsequence. LCS algorithm discarded redundant subsequence's which further reduce entropy. This lead to increase in predictability of travel time sequence. Thus, the combination of LZ and LCS give more accurate predictability and easily can be implemented for travel time forecasting in ITS to calculate prediction rate for next travel time based on given travel time sequence.

प्रस्तावना

पूर्वानुमानशीलता के दृष्टिकोण से यात्रा समय अनुक्रम की विश्वसनीयता को मापने के लिए कई गणिति, और वैज्ञानिक मॉडल्स प्रस्तुत किए गए हैं। इस अध्ययन में एक गणितीय, और एल्गोरिद्मिक मॉडल की चर्चा की गयी है तथा एक मॉडल के साथ तुलना भी की गयी है। अनुक्रम की विशेषता के आधार पर विशेष रूप से चर्चित आम संख्ययिकी की प्रणालियाँ जैसे कि औसत, मानक विचलन इत्यादि की गणना, इन वैज्ञानिक तथा गणितीय, मॉडल से अच्छी नहीं है जिसकी चर्चा[1] में दिये गए शोध पत्र में की गयी है।

• **एन्ट्रॉपी :** सीधे एन्ट्रॉपी में जाने से पहले, इससे संबंधित अवधारणा के बारे में बात करते हैं, मान लो किसी चर में कुछ प्रयोग (यादृच्छिक कोड- a, b, c, d) है, जो एक सतत यादृच्छिक चर X उत्पन्न करता है। जहां ए x={ a, b, c, d } तथा a, b, c, d कुछ विशिष्ट वास्तविक संख्याएँ हैं जिनकी प्रायिकता क्रमशः 0.50, 0.25, 0.125, 0.125 है तथा x एक सतत रैन्डम चर है, जिसकी एन्ट्रॉपी की सही गणना निम्न प्रकार से की जाती है-

$$\text{Entropy} = -\sum_{i=1}^{n} p_i(x) \log_2 p_i(x)$$

जहां $p_i(x)$, x की i स्टेट में प्रायिकता है। हालाँकि, यात्रा समय अनुक्रम एक अस्थायी रूप से सहसंबद्ध स्टोकेस्टिक प्रक्रिया है क्योंकि एक ऐतिहासिक यात्रा समय, अगली यात्रा में लगने वाले समय के परीकलित मान को दृढ़ता से प्रभावित करता है। इसलिए, इस तरह के एक स्ट्रिंग अनुक्रम की वास्तविक एन्ट्रॉपी का अनुमान लेम्पेल-ज़िव एन्ट्रॉपी द्वारा लेम्पेल-ज़िव डेटा संपीड़न एल्गोरिद्म के साथ लगाया गया है। समय अनुक्रम की अनुमानित एन्ट्रॉपी की गणना के लिए लेम्पेल-ज़िव एन्ट्रॉपी एस्टीमेटिंग विधि वास्तविक एन्ट्रॉपी के बराबर मान तो नहीं देती, पर यदि समय अनुक्रम की लंबाई n हो और इसका मान लगभग अनंत हो, तो अनुमानित एन्ट्रॉपी उस समय अनुक्रम की वास्तविक एन्ट्रॉपी में परिवर्तित हो जाती है।

• **ऊपरी बाध्य पूर्वामान (UBP)** : प्रायिकता किसी भी पूर्वानुमान को परिकलित करने के लिए महत्वपूर्ण साधन है। प्रायिकता के परिकलित मान के उपयोग से किसी भी उपयुक्त प्रीडिक्टिव एल्गोरिद्म के द्वारा लगभग सटीक यात्रा समय की गणना कर सकते हैं। उदाहरण के लिए, यदि UBP का परिकलित मान 0.9 है, तो इसका मतलब है कि 90% टाइम्स हम यात्रा समय की पूर्वानुमान करने में सही हैं। अन्य शब्दों में, चाहे हमारा पूर्वानुमान कितना भी अच्छा हो, हम 90% से अधिक सटीकता के साथ अगली यात्रा में लगने वाले समय का पूर्वानुमान नहीं कर सकते। इसलिए UBP प्रत्येक यात्रा समय की अनुक्रम के पूर्वानुमान के लिए मौलिक सीमा का प्रतिनिधित्व करता है।

• **एन्ट्रॉपी गणना में लेम्पेल-ज़िव में समस्या :** लेम्पेल-ज़िव एन्ट्रॉपी की गणना समय अनुक्रम के डिस्क्रिटाइजेशन पर आधारित है। एन्ट्रॉपी के गणनात्मक प्रोसेस में लेम्पेल-ज़िव सम्पूर्ण ऐतिहासिक स्ट्रिंग को समान रूप से दिखने वाली उप-स्ट्रिंग के लिए स्कैन करता है (उदाहरण के लिए उप-स्ट्रिंग "ABB", स्ट्रिंग "CABBCBC" में उपस्थित है)। निरंतर समय अनुक्रम तथा लेम्पेल-ज़िव एल्गोरिथ्म के एन्ट्रॉपी परिकलन के गुणों के आधार पर, यात्रा के समय की विविधताओं को शामिल करने के लिए, आगे ‘‘एलजेड-एलसीएस एन्ट्रॉपी” नाम के एल्गोरिद्म का विवरण किया गया है। ऐतिहासिक समय अनुक्रम में एक विशिष्ट पैटर्न की खोज करते समयए निरंतर समय अनुक्रम में ‘‘स्ट्रिंग समानता” के साथ ‘‘बिल्कुल समान” स्थिति को प्रतिस्थापित किया है। एलज़ेड-एलसीएस की परिभाषा में, उनकी स्ट्रिंग की समानता को दर्शाने के लिए दो स्ट्रिंग के एलसीएस (लॉन्गेस्ट कॉमन सबसीक्वेन्स) का उपयोग करते हैं। यदि एलसीएस का परिकलित मान परिभाषित मानक से छोटा है, तो दो स्ट्रिंग को एक समान ही माना जाता है।

• **एलजेड-एलसीएस एन्ट्रॉपी :** एलजेड-एलसीएस, लेम्पेल-ज़िव डेटा एल्गोरिद्म तथा लॉन्गेस्ट कॉमन सबसीक्वेन्स एल्गोरिद्म का संयोजन है, जिसमें हम एलसीएस एल्गोरिद्म का उपयोग करते हुए एलसीएस के लंबाई के मान के साथ, एन्ट्रॉपी के मान की गणना करते हैं।

उदाहरण के लिए- "abc", "abg", "bdf", "aeg", "acefg",.. ..आदि "abcdefg" के सबसीक्वन्स हैं।

जैसे कि - इनपुट अनुक्रम "ABCDGH" और "AEDFHR" के लिए एलसीएस "ADH" है, तथा एलसीएस की लंबाई 3 है।

सामग्री एवं विधि

मान लीजिए कि एक यात्रा समय अनुक्रम अंतराल [x, y] से घिरा हुआ है जहाँ y अधिकतम यात्रा समय है तथा x न्यूनतम यात्रा समय है। इस अंतराल को समान रूप से N उप-अंतरालों में विभाजित किया गया है तथा प्रत्येक उप-अंतराल को एक विशिष्ट वर्ण (A,B,C,...) प्रदान किया गया है। इस परिवर्तन के बाद, यात्रा समय अनुक्रम सीमित स्टेट्स के साथ एक स्टोकेस्टिक प्रोसेस बन जाता है। हालाँकि, यात्रा समय अनुक्रम एक अस्थायी सहसंबद्ध स्टोकैस्टिक प्रोसेस है, क्योंकि एक ऐतिहासिक यात्रा

समय अनुक्रम अगली यात्रा के समय के परिकलित मान को प्रभावित करता है, इसलिए, इस तरह के स्ट्रिंग अनुक्रम की एन्ट्रॉपी के अनुमानित मान की गणना लेम्पेल-ज़िव डेटा कम्प्रेशन एल्गोरिद्म से की गई है।

यह फ़ंक्शन टाइम अनुक्रम की अनुमानित एन्ट्रॉपी P की गणना करने के लिए है जो कि लेम्पेल-ज़िव कम्प्रेशन एल्गोरिथ्म पर आधारित है[4]।

$$P = (\frac{1}{n}\sum_{i=1}^{n} l_i)^{-1}\log_e n \quad \ldots (2)$$

जहां द स्ट्रिंग की लंबाई है तथा li पद से शुरू होने वाले सबसे छोटे सब-स्ट्रिंग की लंबाई है जो पद 1 से पद i-1 से पहले की हो तथा लिस्ट में उपस्थित नहीं हो।

यदि यात्रा समय अनुक्रम की वास्तविक एन्ट्रॉपी Pmax है, तथा इस अनुक्रम में N स्टेट्स है, तो उस अनुक्रम की अनुमानित एन्ट्रॉपी $P <= P_{max}(S, N)$ होगी[2]।

जहाँ -

$$S = H(P_{max}) + (1 - P_{max})\log_2(N - 1) \quad \ldots (3)$$

$H(P_{max})$ - एन्ट्रॉपी फंक्शन तथा S ऊपरी बाध्य पूर्वामान अर्थात् UBP को दर्शाता है।

$$H(P_{max}) = - \{P_{max}\log_2(P_{max}) + (1 - P_{max})\log_2(1-P_{max})\} \quad \ldots (4)$$

समीकरण (2) में अनुमानित एन्ट्रॉपी P की गणना करने के लिए लेम्ब्डा li की गणना (कलन विधि अनुभाग 1 तथा 2 में है) दो तरह से की गयी है, जिसे एलज़ेड एन्ट्रॉपी P_{lz} (अनुभाग-1) तथा एलज़ेड-एलसीएस एन्ट्रॉपी Plzlcs (अनुभाग-2) कहा गया है। एलज़ेड एन्ट्रॉपी से प्राप्त UBP को UBPlz तथा एलज़ेड एन्ट्रॉपी से प्राप्त UBP को UBP_{lzlcs} कहा गया है।

अनुभाग-1

माना एक स्ट्रिंग्स X="ABBC....." है तथा यहाँ Xpre पिछली सब-स्ट्रिंग को दर्शाता है[1]।

1. यहाँ X_{pre} =" है तथा A, Xpre में अभी तक नहीं आया है, इसलिए l_1= 1,
2. यहाँ X_{pre} =A है तथा B, X_{pre} में अभी तक नहीं आया है, इसलिए l_2= 2,
3. यहाँ X_{pre} =AB है तथा एक बार आ चुका है, लेकिन BC, X_{pre} में अभी तक नहीं आया है, इसलिए l_3= 2,
4. यहाँ X_{pre}= ABB है तथा C, X_{pre} में अभी तक नहीं आया है, इसलिए l_4= 1।

एलज़ेड एन्ट्रॉपी समीकरण संख्या-2 यह दर्शाती है, कि यदि l_i का मान अधिक होगा तो एन्ट्रॉपी का मान कम हो जाएगा।

अनुभाग-2

यहाँ X = BACDB तथा Y = BDCB के एलसीएस की लंबाई के लिए सारणी दिखाई गयी है[3] (संदर्भ-3 में दिये गए LCS एल्गोरिद्म के आधार पर)-

Let c[i,j] be the length of an LCS of X[1...i] and Y[1...j].
c[i,j] can be computed as follows:

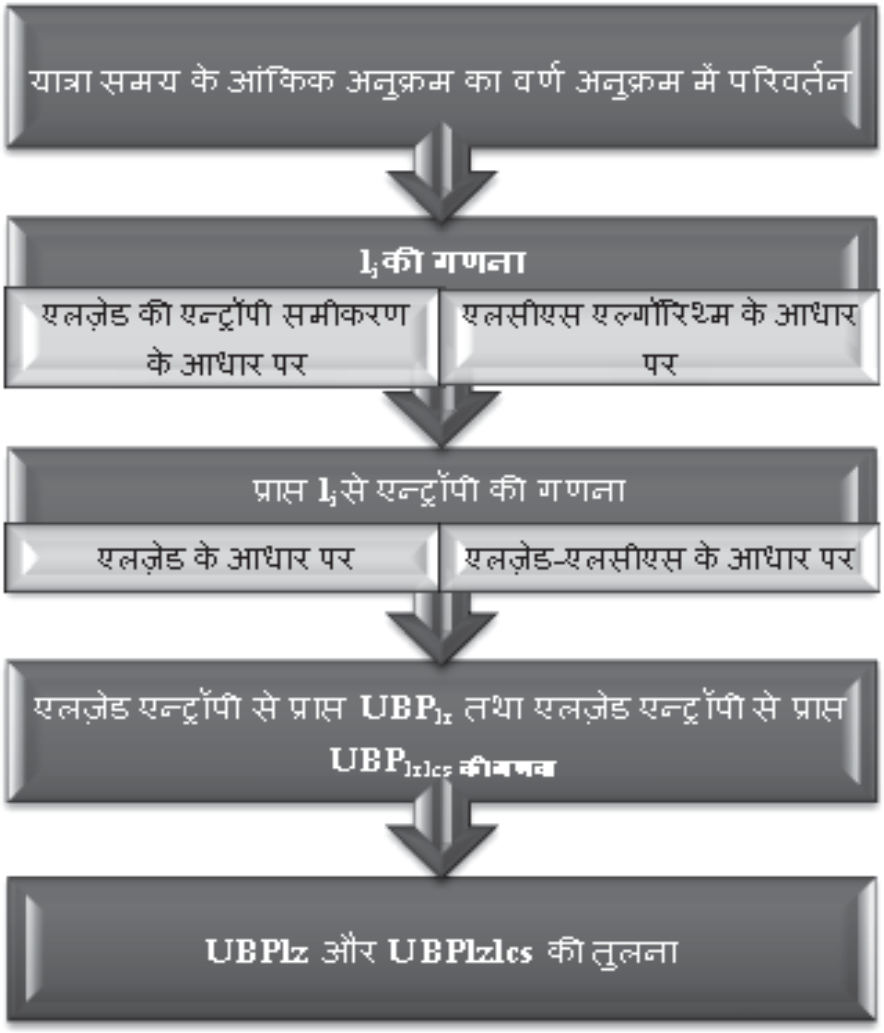
$$c[i,j] = \begin{cases} 0 & \text{if } i=0 \text{ or } j=0, \\ c[i-1,j-1]+1 & \text{if } i,j>0 \text{ and } x_i=y_j, \\ \max\{c[i,j-1],c[i-1,j]\} & \text{if } i,j>0 \text{ and } x_i \neq y_j. \end{cases}$$

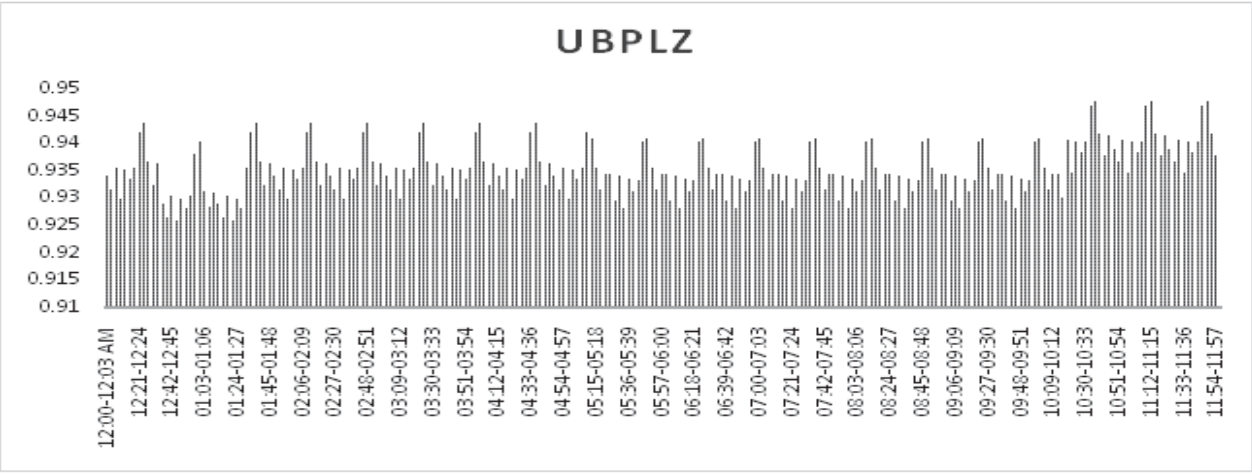
		0	1	2	3	4
			B	D	C	B
0		0	0	0	0	0
1	B	0	1	1	1	1
2	A	0	1	1	1	1
3	C	0	1	1	2	2
4	D	0	1	2	2	2
5	B	0	1	2	2	3

UBP_{lz} तथा UBP_{lzlcs} की गणना प्रक्रिया को प्रवाह आरेख से दर्शाया गया है-

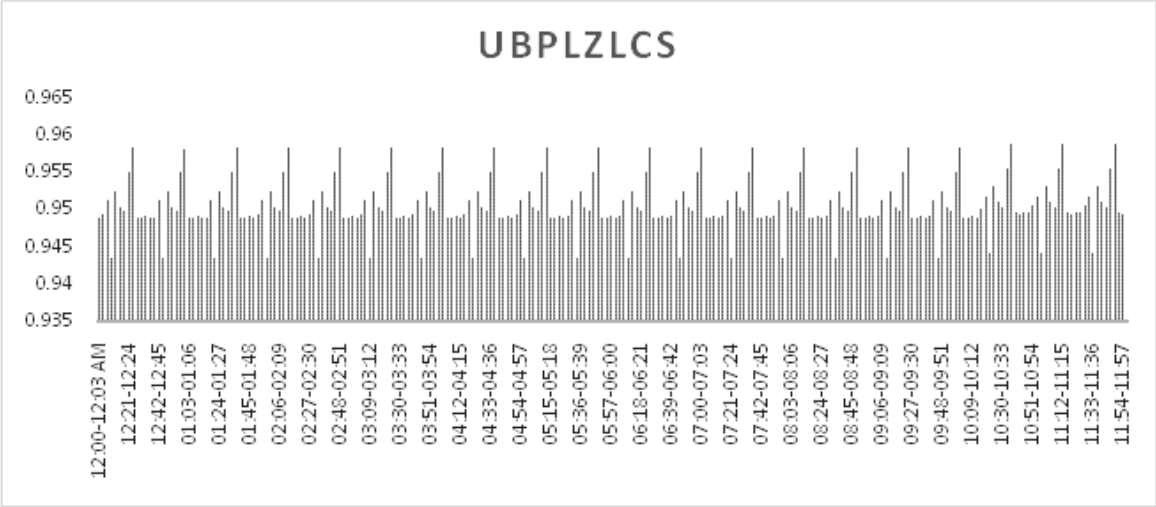
परिणाम एवं विवेचना

चित्र-1 UBP_{lz} तथा चित्र-2 UBP_{lzlcs} 12 घण्टे में प्रत्येक 3 मिनट के लिए UBP प्रदर्शित करता है तथा UBP की गणना, समय अनुक्रम (180 दिन) की एन्ट्रॉपी से की जाती है।

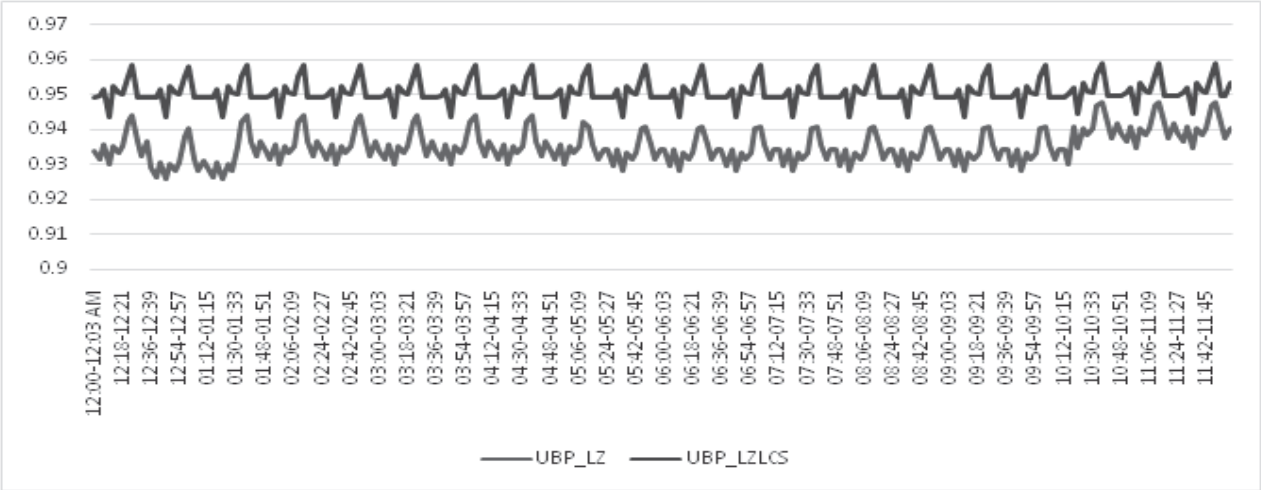

यात्रा समय के आंकिक अनुक्रम का वर्ण अनुक्रम में परिवर्तन
l_i की गणना
एलज़ेड की एन्ट्रॉपी समीकरण के आधार पर
एलसीएस एल्गोरिथ्म के आधार पर
प्राप्त l_i से एन्ट्रॉपी की गणना
एलज़ेड के आधार पर
एलज़ेड-एलसीएस के आधार पर
एलज़ेड एन्ट्रॉपी से प्राप्त UBP_{lz} तथा एलज़ेड एन्ट्रॉपी से प्राप्त UBP_{lzlcs} की तुलना
UBPlz और UBPlzlcs की तुलना


UBPLZ
0.95
0.945
0.94
0.935
0.93
0.925
0.92
0.915
0.91
12:00-12:03 AM
12:21-12:24
12:42-12:45
01:03-01:06
01:24-01:27
01:45-01:48
02:06-02:09
02:27-02:30
02:48-02:51
03:09-03:12
03:30-03:33
03:51-03:54
04:12-04:15
04:33-04:36
04:54-04:57
05:15-05:18
05:36-05:39
05:57-06:00
06:18-06:21
06:39-06:42
07:00-07:03
07:21-07:24
07:42-07:45
08:03-08:06
08:24-08:27
08:45-08:48
09:06-09:09
09:27-09:30
09:48-09:51
10:09-10:12
10:30-10:33
10:51-10:54
11:12-11:15
11:33-11:36
11:54-11:57

चित्र 1 – UBPlz



चित्र 2 – UBPlzlcs



चित्र 3 – UBPcmp

निम्न ग्राफ़ UBPlz और UBPlzlcs के बीच तुलना दिखाता है-

निष्कर्ष

इस अध्ययन में हमने एक सड़क के लिए डमी डेटा के लिए यूबीपी की गणना एलज़ेड तथा एलज़ेड-एलसीएसए दोनों ही एल्गॉरिद्म का प्रयोग किया है, जिसके आधार पर, हम यूबीपी तथा एन्ट्रॉपी की गणना नियमित विधि की तुलना में एलज़ेड-एलसीएस का प्रयोग करके 2% से भी अधिक सटीकता से कर सकते हैं तथा किसी भी आम सांख्या की प्रणाली से जैसे कि- रेग्रेस्सिओन, एसटीडी, एव्रेज इत्यादि काफी अधिक सटीक परिणाम देती है।

नतीजन, एलजेड और एलसीएस का संयोजन अधिक सटीकता से पूर्वानुमान दर की गणना करता है तथा एलसीएस की टाइम कॉम्प्लैक्सिटी को ध्यान में रखते हुए, किसी भीं इंटेलीजेंट सिस्टम जैसे कि-आईटीएस (इंटेलीजेन्ट ट्रान्सपोर्ट सिस्टम) इत्यादि में आसानी से उपयोग किया जा सकता है, जो दिए गए यात्रा समय अनुक्रम के आधार पर अगली यात्रा समय के सम्भावित मान की गणना

करने के लिए पूर्वानुमान दर के परिकलित मान का प्रयोग कर सकता है।

संदर्भ

1. Huiping Lia, Fang He, Xi Lin, Yinhai Wang & Meng Li ,2019. Travel time reliability measure based on predictability using the Lempel-Ziv algorithm, *Transportation Research Part C: Emerging Technologies* **101** (2019) 161-180.
2. Song C, Qu Z, Blumm N & Barabasi A, *Limits of predictability in human mobility, Science,* **327** (2010) 1018-1021.
3. Thomas H Cormen, Charles E Leiserson, Ronald L Rivest & Clifford Stein, *Introduction to Algorithms,* The MIT Press, Cambridge, Massachusetts London,England, McGraw-Hill Book.
4. Ziv J & Lempel A, *A Universal Algorithm for Sequential Data Compression.* IEEE Press (1977).