레벤슈타인 거리 및 문자열 유사도 기반 로그 파서 성능의 정량적 평가 방법에 관한 연구

김희재, 송현석, 윤창범, 박택근*

한전KDN 전력ICT연구원

{hola.halo2, hyunseok.song.17, changbe0m.yun, *reply_1997}@kdn.com

A Study on the Quantitative Performance Evaluation of Log Parser Using Levenshtein Distance and String Similarity Metrics

Hee-Jae Kim, Hyun-Seok Song, Chang-Beom Yun, *Taek-Keun Park

KDN Electric Power ICT Research Institute

요 약

본 논문은 로그 파서(Log Parser)의 성능을 정량적으로 평가하기 위한 방법을 제안한다. 기존의 로그 파서 성능 평가는 특정 로그 형식이나 환경에 종속된 인식률 중심의 방식으로, 로그 전체 문자열 수준에서의 정보 손실이나 재현 정확도를 충분히 반영할 수 없는 한계를 가진다. 이를 개선하기 위해 본 연구에서는 로그 파서를 통해 생성된 로그 패턴과 로그 데이터를 이용하여 재현 로그를 생성하고, 레벤슈타인 거리 기반 문자열 유사도를 활용하여 원본 로그와의 비교를 통해 성능을 평가하는 방법을 제안하고자 한다.

Ⅰ. 서론

클라우드 환경과 분산 컴퓨팅 인프라의 확산으로 시스템 운영 과정에서 생성되는 로그 데이터의 양과 종류가 급격히 증가하고 있다. 로그 데이터는 시스템 모니터링, 장애 분석, 보안 이벤트 탐지 등에서 핵심 데이터로 활용되며, 이를 효율적으로 분석하기 위해 비정형의 로그를 구조화하는 로그 파서(Log Parser)가 필수적이다. 기존의 규칙 기반 로그 파서는 고정된 로그 형식에는 효과적이지만 새 형식이 추가되거나 구조가 변화하는 환경에서 적용 범위가 제한되는 한계가 있다. 최근 범용 및 기계학습 기반 파서가 등장했으나 파서의 범용성이 확대될수록 파싱 정확도와 정보 손실을 객관적으로 평가할 방법의 필요성이 커지고 있다. 기존 평가는 주로 인식률 중심으로 이루어져 파싱 과정의 정보 손실이나 재현 정확도의 정량적인 측정이 어려웠고, 다양한 로그 형식 환경에서 일관된 평가 기준이 부재하여 파서 간 성능 비교가 곤란한 문제가 있다.[1] 이러한 문제를 해결하기 위해 본 연구에서는 레벤슈타인 거리와 문자열 유사도 지표를 활용한 로그 파서 성능 평가 방법을 제안한다. 파싱 결과로 재현 로그를 생성하고 원본과의 유사도를 측정하여 손실률과 재현율을 결합한 정량적 성능 지표를 제공함으로써, 다양한 로그 형식을 지원하는 로그 파서의 성능을 비교·검증할 수 있는 객관적이고 일관된 평가 기준을 확보하고자 한다.

Ⅱ. 본론

2.1. 재현 로그(Reconstructed Log) 기반 평가 개요

로그 파서 성능을 정량적으로 평가하기 위해서는 파싱 결과가 원본 로그 정보를 정확하게 유지하고 있는 정도를 판단할 수 있어야 한다. 이를 위해 본 연구에서는 로그 파서를 통해 생성된 로그 패턴과 로그 데이터를 이용하여 원본 로그를 재구성한 재현 로그를 정의하고, 재현 로그와 원본 로그 간의 문자열 비교를 수행하고자 한다. 재현 로그는 로그 파서가 인식한 고정 패턴과 가변 파라미터를 결합하여 생성된 문자열로, 로그 파서의 처리 결과를 원본 로그 문자열 수준에서 직접적으로 비교할 수 있는 기준 데이터를 제공한다. 이러한 비교 방식은 로그 항목 단위 평가가 아닌 문자열 전체 수준의 평가를 가능하게 한다.

2.2 레벤슈타인 거리(Levenshtein Distance) 활용 문자열 유사도 판단

레벤슈타인 거리는 두 개의 문자열 간 차이를 편집 연산의 최소 횟수로 표현하는 문자열 거리 지표이다. 편집 연산은 삽입(Insert), 삭제(Delete), 치환(Substitution)의 세 가지 연산으로 구성된다.

원본 로그 문자열을 S_o, 재현 로그 문자열을 S_r라고 할 때, 레벤슈타인 거리는 S_o를 S_r로 변환할 때 필요한 최소 편집 연산 횟수로 정의된다. 이 값이 작을수록 두 문자열은 유사하며, 재현 로그가 원본 로그를 재현하고 있다는 것을 의미한다.[2] 본 연구에서는 레벤슈타인 거리를 직접적인 거리 값으로 사용하기보다는 문자열 간 유사도를 판단하기 위한 기초 지표로 활용한다. 이를 통해 로그 파서가 원본 로그의 정보를 정확하게 유지하고 있는 정도에 대해 정량적으로 평가할 수 있다.

2.3 문자열 유사도 지표

재현 로그와 원본 로그 간의 유사도를 보다 일반화된 형태로 표현하기 위해, 문자열 유사도 지표를 활용한다. 문자열 유사도는 두 개의 문자열 간의 유사성을 0과 1 사이의 값으로 정규화하여 표현할 수 있는 지표로 값이 1에 가까울수록 두 문자열이 유사함을 의미한다.[3]

No.	유사도 산출방법
1	레벤슈타인 거리(Levenshtein Distance) 기반 유사도
2	코사인 유사도(Cosine Similarity)
3	자카드 유사도(Jaccard Similarity)

〔표 1〕 본 연구에서 고려하는 문자열 유사도 산출 방법

[표 1]에 제시한 방법은 모두 재현 로그와 원본 로그 간의 문자열 비교를 통해 유사도를 산출할 수 있으며, 로그 파서 성능 평가를 위한 재현율 계산의 기반으로 활용될 수 있다.

2.4 인식률(Recognition Rate) 및 손실률(Loss Rate) 산출

로그 파서의 성능을 평가하기 위해 먼저 인식률과 손실률을 정의한다. 인식률은 원본 로그 문자열 대비 재현 로그 문자열이 차지하는 비율로 정의되며, 로그 파서가 원본 로그를 어느 정도까지 인식하고 재구성했는지를 나타낸다.

인식률 R은 다음과 같이 정의된다. 여기서 L_r는 재현 로그 문자열의 길이, L_o는 원본 로그 문자열의 길이를 나타낸다.

$$R = \frac{L_r}{L_o}$$

손실률 L은 인식률을 기반으로 정의되며, 원본 로그 중 로그 파서가 인식하지 못한 문자열의 비율을 나타낸다. 손실률은 다음과 같이 산출된다.

$$L = 1 - R$$

손실률은 로그 파서가 로그 파싱 과정에서 발생시킨 정보 손실의 정도를 정량적으로 나타내는 지표이다.

2.5 재현율(Reproduction Rate) 산출

재현율은 재현 로그가 원본 로그를 얼마나 유사하게 재현하였는지를 나타내는 지표이다. 재현율은 재현 로그와 원본 로그 간의 문자열 유사도를 측정하여 산출한다. 문자열 유사도 측정 결과를 S라고 할 때, 재현율은 해당 유사도 측정 결과값으로 정의되며, 그 값이 클수록 로그 파서가 원본 로그의 구조와 내용을 정확히 재현했음을 의미한다. 재현율은 인식률보다 같거나 작은 값을 가지며, 로그 재현 품질을 직접적으로 반영하는 핵심 지표로 활용될 수 있다.

2.6 손실 무시 정확도 및 손실 보정 정확도 산출

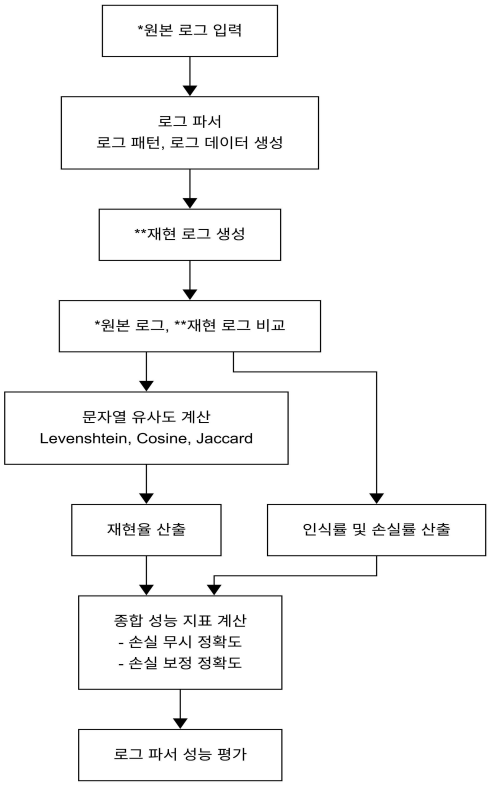
재현율과 손실률을 종합적으로 고려하기 위해 본 연구에서는 손실 무시 정확도(Loss-Ignored Accuracy)와 손실 보정 정확도(Loss-Corrected Accuracy)를 정의한다. 손실 무시 정확도 A_i는 재현율 S과 손실률 L 값의 합산으로 정의되며, 로그 재현 정확도와 손실 정도를 동시에 반영한다.

$$A_i = S + L$$

손실 보정 정확도 A_c는 재현율 S과 손실률 L에 각각 가중치를 적용하여 산출되는 지표로, 손실된 정보에 대한 패널티를 보다 명확히 반영할 수 있다.

$$A_c = S \times L$$

이 두 지표를 통해 로그 파서의 성능을 단일 지표가 아닌 복합적인 관점에서 평가할 수 있으며, 다양한 로그 파서 간 성능 비교를 보다 객관적으로 수행할 수 있다. 다음 [그림 1]에서는 순서도를 통해 제안 로그 파서 성능 평가 방법의 전체적인 성능 평가 과정을 표현하였다.



[그림 1] 제안 로그 파서 성능 평가 개념도

Ⅲ. 결론

본 논문에서는 로그 파서의 성능을 정량적으로 평가하기 위한 방법을 제안하였다. 로그 파서를 통해 생성된 로그 패턴과 데이터를 기반으로 재현 로그를 생성하고, 레벤슈타인 거리 및 문자열 유사도 지표를 활용하여 원본 로그와의 비교를 통해 재현율을 산출하였다. 또한 인식률과 손실률을 결합한 성능 지표를 통해 로그 파싱 과정의 정보 손실과 재현 정확도를 동시에 평가할 수 있음을 보였다. 본 연구의 의의는 개별 로그 항목이 아닌 문자열 전체 수준에서 로그 파서 성능을 평가함으로써, 특정 로그 형식이나 시스템 환경에 종속되지 않는 객관적인 평가 기준을 제시한 데 있다. 제안된 방법은 다양한 로그 파서를 동일한 기준에서 비교·분석하고, 성능 개선을 위한 기준 지표로 활용될 수 있다는 점에서 실용적 의미를 가진다. 향후 연구에서는 제안된 평가 방법을 다양한 로그 유형과 파서에 적용하여 활용 가능성을 확장하고, 문자열 유사도 지표 간 특성 차이에 따른 평가 기준의 세분화에 대한 추가 논의가 가능할 것으로 기대된다.

참 고 문 헌

[1] J. Zhu, S. He, J. Liu, P. He, Q. Xie, Z. Zheng, and M. R. Lyu, “Tools and benchmarks for automated log parsing,” ICSE-SEIP '19: Proceedings of the 41st International Conference on Software Engineering: Software Engineering in Practice, pp. 121 - 130, 2019.

[2] V. I. Levenshtein, "Binary codes capable of correcting deletions, insertions, and reversals," Doklady Akademii Nauk SSSR, vol. 163, no. 4, pp. 845 - 848, 1966.

[3] G. Navarro, “A guided tour to approximate string matching,” ACM Computing Surveys(CSUR), Vol. 33, Issue 1, pp. 31 - 88, 2001.