Web アクセスログのステータスコードと並列可能数の分割による検索の応答時間の短縮

Reducing search response time in the web access log by dividing the status code and the number of possible parallelisms

大野 有樹
指導教員　串田 高幸

東京工科大学 大学院 バイオ・情報メディア研究科 コンピュータサイエンス専攻

キーワード：アクセスログ，ログ検索

1． はじめに

システム管理者は EC サイトにシステム障害が発生した際にログを検索することで原因を調査する．EC サイトを迅速に復旧させなければユーザーは購入処理ができなくなる．システム管理者はシステム障害の全容を調査するためにアクセスログを調べる[1]． アクセスログは Web サーバーに送られるリクエストを記録したログである．

課題は並列で検索できるようにログを分けて保存する場合に，分けた先のログの数が異なると，ログの件数が多い保存先のログ検索の応答時間が長くなる．

2．提案

本提案手法の目的は検索の応答時間が早くなることでシステム障害が起きた場合に，システム障害の全容を迅速に知ることができるようにすることである．本提案手法は正常系であるステータスコード 200 番台から 300 番台と異常系である 400 番台から 500 番台でアクセスログを分割する．さらに正常系と異常系で分けたログをログサーバーで並列処理できる数で分割する．並列処理できる数はログサーバーの CPU のスレッド数とする．分割数に合わせて 1 件ずつログをブロックに追記して保存する．並列処理で検索するそれぞれのログの件数の差が減り，検索それぞれの応答時間の差が減る．

図 1 は提案の概要図を示している．サーバーは大きく分けて 2 つで，ログサーバーと Web サーバーがある．ユーザーから Web サーバーにアクセスした場合に生成されたログはログサーバーに転送する．ログサーバー側の”ログ配置”は Web サーバーから受け取ったログをブロックとしてまとめて保存する．保存したログは検索の索引として利

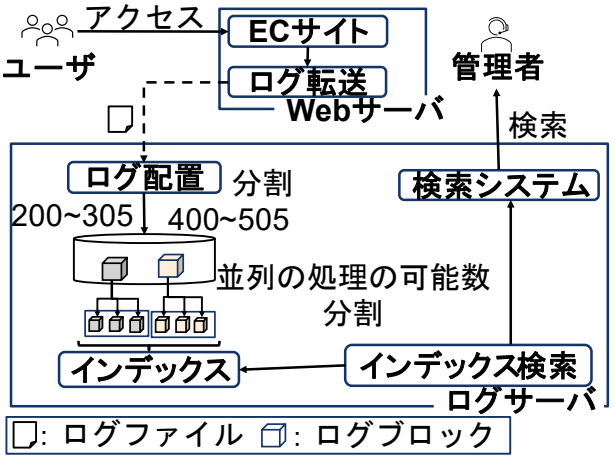

図 1：提案の概要図

用するためのインデックスにログの保存場所をログブロックごとに記録する．

3．デモンストレーション

デモンストレーションの流れを説明する．デモンストレーションは Web ブラウザを通してアクセスログを検索する．使用するログはポーランドのオンラインストアのログである EClog を使用する[2]．システム管理者は現状を把握するために，ロ

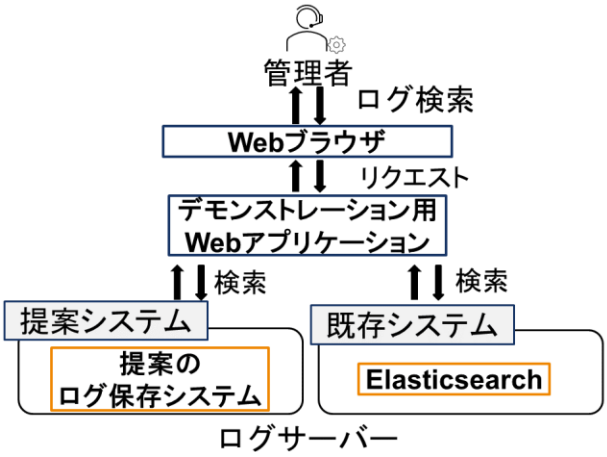

図 3：デモンストレーションの構成

グを検索してシステム障害の全容を把握する．検索クエリは，ステータスコード 500 番台のアクセスログとする．理由は現在起きている状態がステータスコードから分かり，エラーが発生した時刻が分かるためである．この 2 つが分かることで，どのシステムログを見るかと発生時刻からシステムログの時間の条件を絞り込みできる．提案のログ保存システムと既存の Elasticsearch の検索の応答時間を比較する.検索の結果と応答時間を Web ブラウザに載せる．

デモンストレーションの構成図を図 2 に示す．提案のログ保存システムと Elasticsearch の違いはログの保存の仕方である．保存方法の違いによって検索の応答時間にどれだけの違いができるかを比較する.そのため,検索のクエリを発行するシステムを統一し，ログを保存して検索するシステムのみ異なる状況を設定した．検索は Web ブラウザを通してデモンストレーション用 Web アプリケーションで検索クエリを発行して行なう．

あらかじめアクセスログを提案のログ保存システムと Elasticsearch に保存しておく．デモンストレーションでは，システム障害が発生した場合を想定して，検索結果が表示されるまでの時間を測定する.計測方法はデモンストレーション用 Web アプリケーションで測定し,Web ブラウザに表示する．図 3 はデモンストレーションで見せるブラウザ画面の形式を示している．一番上は検索ボックスであり，検索の条件を検索クエリとして入力することができる．検索クエリを入力して検索ボタ

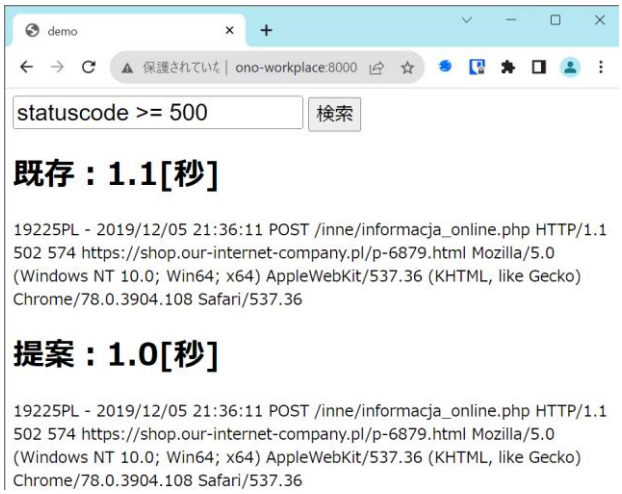

図 2 ブラウザ画面

ンを押すと，検索結果が出力される．検索クエリは同時に実行され，デモンストレーション用 Web アプリケーションが提案のログ保存システムと既存システムの Elasticsearch に検索クエリを発行する．検索結果が出力されると検索の応答時間が表示され，提案手法の検索の応答時間が早くなることが分かる．

4．おわりに

課題は並列で検索できるようにログを分けて保存する場合に，分けた先のログの数が異なると，ログの件数が多い保存先のログ検索の応答時間が長くなる．課題に対して，ステータスコードごとにアクセスログを分割し，更に並列処理できる数で分割し保存することで，一つあたりのログのエントリー数を減らす手法を提案する．既存手法 Elasticsearch と提案手法で同じ検索条件で検索速度を比較し,Web アプリケーションを通じてブラウザから検索の応答時間の違いを視覚的に示す．

参考文献

[1] Grace, L. K., V. Maheswari, and Dhinaharan Nagamalai. ”Analysis of web logs and web user in web mining.” arXiv preprint arXiv:1101.5668 (2011).

[2] Chodak, G., Suchacka, G. and Chawla, Y.: EClog: HTTP-level e-commerce data based on server access logs for an online store (2020).