稀にアクセスされるWebページを検出するシステムに関する研究

立花 一樹 (0810037)

北陸先端科学技術大学院大学 情報科学研究科

2012年2月6日

キーワード: 稀なWebアクセス、長期観測、べき乗則.

1991年にWorld Wide Webが発明されてから、ホームページの閲覧だけでなく電子決済やニュースの閲覧、Webメールの利用などHTTPを利用した様々なサービスが実現されるようになった。Webを利用するユーザ層の拡大によってアクセス目的が多様化した結果、そのアクセス先となるインターネット上のWebコンテンツの情報量は莫大なものとなっている。しかし、Webページのアクセス頻度とそれに基づくアクセス順位との関係は、一般的にべき乗則に従うことが知られている。このことから、大半のアクセスは一部の人気の高いWebページへ集中していることが分かっている。これまでは、Webページのランキング調査のように人気のある一部のWebページの存在やそれへのアクセス分布は詳しく調べられてきたが、その他の膨大な数となるアクセス頻度が低いWebページへのアクセスやその存在はほとんど知られていない。それらへの稀なアクセスの中には、単に人気のないWebページへのアクセスだけでなく、例えば現在社会問題となっているWebを介して感染するマルウェアによる通信のような特異な目的の通信が含まれている可能性がある。

そこで、本研究はこれまで本格的に調査されることがなかった稀にアクセスされるWebページ群へのアクセスを検出するために長期間に渡って観測可能なシステムを提案する。そして、これまで知られていないWebページの存在やアクセスパターンを発見することで、これまで見過ごされてきた危険なアクセスを迅速に検出したり、Webの生態を解明する上で社会的に有用な知見を得ることができると考えている。研究を進めるにあたり、まずアクセス頻度とそれに基づく順位の関係から本研究が対象とする稀にアクセスされ

る Web ページ群の分布とそのデータ量を把握した。

次に、「稀な Web ページへのアクセス」を定義し、長期間に渡って実時間で検出可能なシステムを設計、試作した。検出システムには、様々な目的でアクセスされたあらゆる Web ページの URL を記録できる容量効率の高いデータ構造と、アクセス履歴に含まれる種々のパラメータの中から稀にアクセスされた Web ページを実時間で峻別するアルゴリズムが求められる。本研究では、観測期間に依存しない公正な判別方法を有限の計算機リソースの中で実現するためにアクセス間隔に着目し、長期間に渡ってアクセス履歴を効率的に記録可能なデータ構造を考案した。そして、試作した稀な Web アクセスを検出するシステム (Web-Prospector) を本学学内ネットワークに設置し、提案システムの動作検証を行った。

16 日間の検出実験の結果、観測された URL の総数は約 2,900 万個となった。稀な URL の検出実験では判定条件を固定し、学習期間を変化させて稀にアクセスされた URL の総数を計測した。

提案システムは実時間検出が可能であるが、本方式では膨大な URL 文字列をメインメモリ内に記録することが困難であることが判明した。この問題に対し、ハッシュテーブルを動的に拡張可能にすることで静的に確保される固定領域を削減することや、パトリシアトライの導入によって URL 文字列を圧縮するなど、本問題を解決する手法について考察した。