大規模データ処理法 課題 1

Shu Sakamoto
学籍番号: 71775236
ログイン名: t17523ss

解析の目的

2019 年度春学期「大規模データ処理」の授業では, 201[]年 4-5 月における Wikipedia のアクセスログを解析した. このとき我々は 201[]年 4-5 月において人気のあるページを確認するため, 4-5 月における総アクセス数の順位に着目した. しかし, この手法を用いると "トップページ" や "[]" のように Wikipedia の性質上アクセスされやすいページが上位に表示されてしまい, これはノイズとなってしまう. また, "[]" のように常に検索されやすいページも表示されてしまうが, これは "4-5 月において人気のあるコンテンツ" としての側面が薄いページであるため, これもなるべく取り除きたい. つまり, 我々が着目したいのは任意の記事の 4-5 月におけるニュース性である. そこで本課題では, 慢性的に人気のあるページではなく, コンテンツのニュース性が高いページが上位に表示されるような解析を行なった.

解析の手順

ニュース性をはかる指標として, $\frac{\Delta A}{\Delta t}$ を用いる (Aは任意の時間窓におけるアクセス数, tは時間). つまり, ある時間においてアクセス数がどれだけ増えたかに着目する. この指標を用いれば "トップページ" や "[]" といったページは上位に表示されず, 例えば "令和" のようなニュース性の高いページのみが上位に表示されるはずである.

わかったこと

やってみた感想