4P-1

周波数分析を利用した周期的にブックマークされるwebページの特定

綱本 圭祐† 亀井 清華†† 藤田 聡††
†広島大学 工学研究科 ††広島大学 工学研究院

1 はじめに

web上には, 年賀状のイラストに関するページや梅雨に関するページなど一年周期で利用されるページが存在する [1]. 一方で, 一年以外の周期で利用されるページも存在する可能性がある. 例えば, 半年に一度開催される音楽イベントに関するページは半年周期で利用されるであろう. これらの周期は人間の生活に関係のある周期であるが, 生活に関係のない周期でもこのようなページが存在する可能性がある.

本研究ではソーシャルブックマークを利用して, 周期的に利用されるページに関して調査を行う. 具体的にはブックマークされるタイミングに周期性があるページがどの程度の割合で存在するのか確認する. さらに, ある周期と結びつきが強い特定のジャンルが存在するのかどうかについても確認する. 周期性の判定には, 離散フーリエ変換を利用した周波数分析を利用する.

2 関連研究

山家らはブックマークされるタイミングが一年周期であるページに着目し, 一年のうち特定の時期に需要が増えるページを特定した [1]. さらにそのようなページを利用して時期連動型のweb検索結果のランキング手法を提案した. 彼らの手法はブックマークを集計する粒度が**月**のようなある程度荒い粒度であれば有効であるが, **日**のような細かい粒度になるとブックマークされるタイミングに周期性があるかどうか適切に判定することはできない. 本研究では離散フーリエ変換を利用した周波数分析を行うため, ブックマークを集計する粒度に関係なくブックマークされるタイミングに周期性があるかどうか判定できる. このため, ブックマークされる周期が年単位よりも短いページも特定できる.

3 周波数分析

本研究ではページのブックマーク数を**日**で集計する. データセット上でデータが格納してある最後の日を D_{last} とし, 各ページがデータセット上で初めてブックマークされた日を B_{first} とする. B_{first} から D_{last} までの経過日数に対応するブックマーク数の系列をブックマーク時系列と呼ぶことにする. ブックマーク時系列に対して離散フーリエ変換を利用した周波数分析を行い, 周波数を求める. 離散フーリエ変換は以下の式で行われる.

$$X(k) = \frac{1}{\sqrt{N}} \sum_{n=0}^{N-1} x(n) e^{-i2\pi kn/N}, \quad k = 0, 1 \ldots N-1$$

$x(n)$ は経過日数 n におけるブックマーク数である. N はブックマーク時系列のデータ数であり, B_{first} から D_{last} までの経過日数を表している. k/N が周波数 f_k であり, その逆数が周期 $T(f_k)$ となる. また周波数の影響力の大きさを表すパワースペクトルは以下の式で計算される.

$$P(f_k) = |X(k)|^2, \quad k = 0, 1 \ldots \lceil \frac{N-1}{2} \rceil$$

計算された各パワースペクトルから, ブックマーク時系列の形成に特に大きな影響を与えている主要な周波数を求める. Vlachos らは, 各パワースペクトルは指数分布に従うと仮定し, 閾値以上のパワースペクトルを保持している周波数を主要な周波数としている [2]. 本研究ではこの手法で求めた閾値以上の周波数の逆数を主要な周期とする. ここで, ページによっては複数の主要な周期が特定される可能性があることに注意する.

4 分析結果

分析にははてなブックマークのデータセットを使用した. データセットには196385ページ格納してあるが, 10回以上ブックマークされている149206ページを対象とした. また閾値の設定に使用する確率 P は 10^{-4} とした. 周波数分析の結果, ブックマークされるタイミングに周期性があると判定されたページは32917ページとなり, データセット全体の約17%にあたる.

4.1 ページの除去

32917ページから**周波数0のページ**と**一過性の情報のみ保持しているページ**を除去する. **周波数0のページ**は周波数0のみを主要な周波数として特定されているページであり, 周期は存在しないとみなせる. **一過性の情報のみ保持しているページ**は短期間に集中してブックマークされた後, ほとんどブックマークされなくなるページであり, データ数 N のブックマーク時系列に対して $\lceil N/2 \rceil$ 以下の周期を一つも保持していないページである. これらのページを除去した結果, 残っ

Detection of web page periodically bookmarked by using frequency analysis
† Keisuke Tsunamoto
†† Sayaka Kamei
†† Satoshi Fujita
† Graduate School of Engineering, Hiroshima University
†† Graduate School of Engineering, Hiroshima University

Copyright ©2012 Information Processing Society of Japan.
All Rights Reserved.

たページは2496ページとなり, データセット全体の約1.3%にあたる.

4.2 周期の分類

ブックマークされる周期は, 長さによって**生活に関係がある周期** (**以下** L **周期**) と**生活に関係のない周期** (**以下** NL **周期**) に分類することができる. 分類対象の周期は周波数0を除いて最大のパワースペクトルを保持している周期とした. 従って, ページと周期は一対一で対応する. 複数の周期を持つページで, 分類対象の周期が一過性の情報の周期であった場合, そのページは調査の対象から外した. 最終的に1385ページに対して, L周期とNL周期を持つページがどの程度の割合で存在するのか調査を行った. 表1に分類結果を示す. Tは分類対象の周期である. この結果から, 1385ページの内L周期のページが約23%存在し, NL周期のページが約77%存在することが分かった.

表1: 周期の分類結果

L 周期		NL 周期	
6<T≤8(一週間)	49	0<T≤6	139
12<T≤18(半月)	34	8<T≤12	33
25<T≤35(一ヶ月)	41	18<T≤25	28
55<T≤65(二ヶ月)	28	35<T≤55	52
85<T≤95(三ヶ月)	28	65<T≤85	56
115<T≤125(四ヶ月)	20	95<T≤115	35
175<T≤190(半年)	45	125<T≤175	100
350<T≤380(一年)	63	190<T≤350	264
		380<T	366
合計	312	合計	1073

4.3 ジャンルの分類

周期と結びつきが強いジャンルが存在するのかどうか確認するために, 各周期にどのようなジャンルのページが多く含まれているか調査を行った. ジャンルは10種類に分類し, 結果を表2に示す. 各周期列の左側の数値はジャンルに属するページ数を表しており, 右側は各ジャンルの合計ページ数に対する割合である. ジャンル名の一番下に示してあるベースラインは1385ページに対するL周期とNL周期のページ数の割合を表している. この割合より大幅に高い割合を持つジャンルは周期と結びつきが強いジャンルであると考えることができる. L周期ではイベントジャンルが, NL周期では社会ジャンルのページが他のジャンルと比較してベースラインより大幅に高い割合で存在していることが分かった. 詳細に調査を行っていくとイベントジャンルは175<T≤190(半年) と 350<T≤380(一年) に, 社会ジャンルでは190<T≤350にそれぞれ高い割合で存在していた.

表2: L周期とNL周期のジャンル分け

ジャンル名	合計	L 周期		NL 周期	
IT	399	82	21%	317	79%
web	248	54	22%	194	78%
レクリエーション	198	40	20%	158	80%
趣味	98	30	31%	68	69%
生活	65	18	28%	47	72%
社会	39	4	10%	35	90%
教育	25	5	20%	20	80%
ビジネス	66	16	24%	50	76%
イベント	44	21	48%	23	52%
その他	203	42	21%	161	79%
ベースライン	1385	312	23%	1073	77%

5 考察

L周期を保持しておりイベントジャンルに属するページには台風に関するページ等があった. このようなページがブックマークされる周期に近づいている場合, 同程度の評価を得ているページと比較してより高く評価することができる. NL周期を保持しており社会ジャンルに属するページには雇用問題に関するページ等があった. なぜこのような周期が発生するのかは不明であるが, 普段意識しにくい周期で社会に関する何らかの事象が起こっている可能性がある. ページ間の比較に利用できることに加え, 人間が意識しにくい周期で適宜な情報の提供に利用できる.

6 まとめ

本研究では, ブックマークされるタイミングに周期性があるページがどの程度存在するのか, またある周期と結びつきが強い特定のジャンルが存在するのかどうか調査した. 調査の結果, 周期性があるページは2496ページ存在することを確認し, L周期とNL周期のそれぞれに結びつきが強いジャンルの存在を確認した. 今後の課題は実際にこれらの周期を利用したシステムを実装し, 評価を行うことである.

参考文献

[1] 山家雄介, アダムヤトフト, 中村聡史, 田中克己. ソーシャルブックマーキングの周期性発見と時期連動型検索ランキングへの適用. **情報処理学会論文誌** *Vol.2, No.3, pp.130-140*, 2009.

[2] M. Vlachos, C. Meek, Z. Vagena and D. Gunopulos. Identifying similarities, periodicities and bursts for online search queries. *In Proceedings of the 2004 ACM SIGMOD international conference on Management of data pp.131-142*, 2004.

Copyright ©2012 Information Processing Society of Japan.
All Rights Reserved.