

関心の視覚化

問題

WEBログからサイト利用者の関心の変化を抽出できないか？

もしできれば、

宣伝が、どの程度にサイト利用者の関心の変化を引き起こすか
イベントが、どの程度にサイト利用者の関心の変化を引き起こすか

といったことをツリーモデルなどを使うことにより調べることができる

実際のWEBログ

```
210.233.115.11 -- [13/May/2004:20:51:37 +0900] "GET /images/common_image03.png
HTTP/1.1" 200 154
210.233.115.11 -- [13/May/2004:20:51:37 +0900] "GET /research/images/title.png
HTTP/1.1" 200 45961
210.233.115.11 -- [13/May/2004:20:51:38 +0900] "GET /images/common_image01.png
HTTP/1.1" 200 7742
210.233.115.11 -- [13/May/2004:20:51:38 +0900] "GET /images/common_image02.png
HTTP/1.1" 200 6486
210.233.115.11 -- [13/May/2004:20:51:38 +0900] "GET /research/images/title_rese
arch.png HTTP/1.1" 200 1087
210.233.115.11 -- [13/May/2004:20:51:48 +0900] "GET /research/download/keitaike
ssai.pdf HTTP/1.1" 200 210012
210.233.115.11 -- [13/May/2004:20:51:49 +0900] "GET /research/download/keitaike
ssai.pdf HTTP/1.1" 304 -
210.233.115.11 -- [13/May/2004:20:52:13 +0900] "GET /licenses/index.shtml HTTP/
1.1" 200 13550
210.233.115.11 -- [13/May/2004:20:52:13 +0900] "GET /licenses/images/title.png
HTTP/1.1" 200 41457
210.233.115.11 -- [13/May/2004:20:52:13 +0900] "GET /images/menu_02_sel.png HTT
P/1.1" 304 -
210.233.115.11 -- [13/May/2004:20:52:13 +0900] "GET /licenses/images/title_lice
nses.png HTTP/1.1" 200 988
64.242.88.10 -- [13/May/2004:21:21:33 +0900] "G
```

要請

- 膨大なデータの中から、関心の変化として考えられる部分を見つけたい
- 3Dで視覚化し、関心の変化を見たい

概念の説明1

- カテゴリーが4つあれば、サイト利用者の行動は、4次元空間内の折線で表現できる

	車	車 (累積)	花	花 (累積)	食事	食事 (累積)	住宅	住宅 (累積)
セッション1	3	3	5	5	1	1	0	0
セッション2	4	7	5	10	2	3	5	5
セッション3	2	9	6	16	1	4	1	6
セッション4	6	15	2	18	4	8	0	6

概念の説明2

サイト利用者 u の関心の空間は、カテゴリ c_1 に属するファイルの数、カテゴリ c_2 に属するファイルの数 c_2 ...カテゴリ c_n に属するファイルの数によって張られる n 次元の空間である。

サイト利用者の関心の変化はこの空間内のベクトルとして定義される。
関心移動ベクトルは次のように計算される。まず

$$V(u,s,c) \dots\dots\dots (1)$$

によって、サイト利用者 u がセッション s で閲覧したカテゴリ c に属するファイル数を表す。

$$V(u,s_1,s_2,c) = V(u,s_1,c) - V(u,s_2,c) \dots\dots\dots (2)$$

によってセッション s_1 とセッション s_2 との間に起こった関心の変化。 s_1 は例えばプログラムが起動された時刻の直前のセッション、 s_2 は1週前の日に一番近いセッションである。次に

$$S(u,s_1,s_2) = \sum_c |V(u,s_1,s_2,c)| \dots\dots\dots (3)$$

によって総変動を定義する。

$$R(u,s_1,s_2,c) = \frac{V(u,s_1,s_2,c)}{S(u,s_1,s_2)} \dots\dots\dots (4)$$

はカテゴリ c に関するセッション s_1 とセッション s_2 との間の関心の変化である。

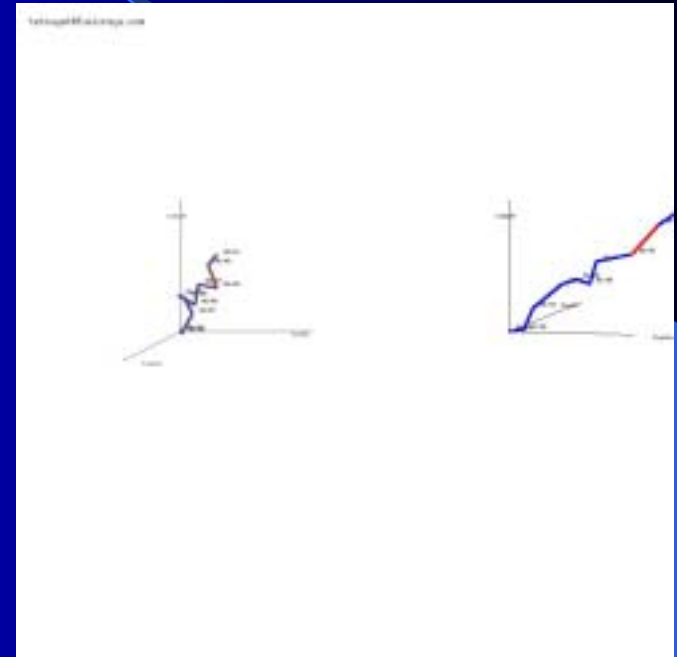
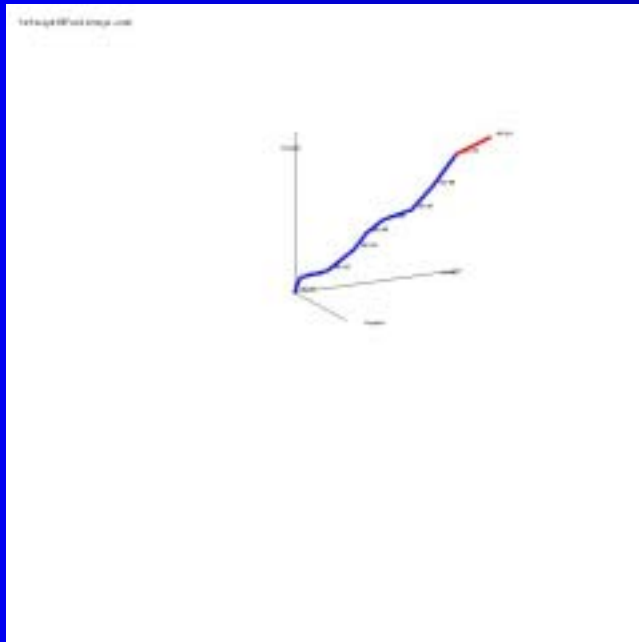
$$\{c_1, c_2, c_n\} \quad \vec{I}(u,s_1,s_2) = (R(u,s_1,s_2,c_1), R(u,s_1,s_2,c_2), \dots, R(u,s_1,s_2,c_n)) \dots\dots (5)$$

を“厳密な意味での関心の移動ベクトル”と定義する。正の値をとる $R(u,s_1,s_2,c)$ の中で大きいものから順に3つとる。その3つを成分とするベクトルを

$$\vec{I}(u,s_1,s_2) \dots\dots\dots (6)$$

で表し、“関心の増加ベクトル”とする。(6)で表されるベクトルを“関心の増加”と呼ぶことにする。

3つのカテゴリーの変化



システムの概要

- OSはLINUX
- C++ (gcc) による高速なプログラム
- 100万行程度のWEBサーバ (apache) のログを解析できる
- 画像やレポートの自動生成機能

関心の変化を説明変数とした解析の例

目的変数

リコメンドをしたあとに、あらかじめ決められた日数以内にサイト利用者がサイトを訪れなかった --> 成功度 1

リコメンドをしたあとに、あらかじめ決められた日数以内にサイト利用者がリコメンドの対象ではない商品のページを閲覧した --> 成功度 2

リコメンドをしたあとに、あらかじめ決められた日数以内にサイト利用者がリコメンドの対象である商品のページを閲覧した --> 成功度 3

リコメンドをしたあとに、あらかじめ決められた日数以内にサイト利用者がリコメンドの対象ではない商品の購入をした --> 成功度 4

リコメンドをしたあとに、あらかじめ決められた日数以内にサイト利用者がリコメンドの対象である商品の購入をした --> 成功度 5

説明変数

- ・1 回のリコメンドをする時点からあらかじめ決められた日数以内に関心の対象が変化すると判定された回数
- ・サイト利用者の累積セッション数
- ・サイト利用者の累積閲覧ファイル数
- ・サイト利用者の関心変化ベクトルの長さの総和
- ・リコメンドする製品の価格
- ・リコメンドする製品の知名度
- ・リコメンドするメールを送った曜日と時間帯
- ・それまでにそのサイト利用者がリコメンドメールを受け取ったことがあるか否か
- ・そのサイト利用者が会員登録をしているか否か