国立情報学研究所殿

ユーザサイド検索ログからの情報出力機能
サーチエンジン定義仕様書

FS-103

1.0 版

KYA group Corporation

版	発行日	発行者	変更内容概要
1.0 版	2013 年 5 月 17 日	KYA group	

目次

1 本資料の概要

本資料は、ユーザサイド検索ログからの情報出力機能(qthlog2tsv)のサーチエンジン定義の仕様とカスタマイズ手順について解説する。

イベントを記録したログファイルを元に検出したイベントを記録したログファイルからアクションを検出する際に使用するサーチエンジンの定義に関して説明した後、具体的に変更する手順をチュートリアル形式で示す。

2 サーチエンジン定義

サーチエンジンの定義、設定は perlmod/serps.pm で行う。List 1 は設定の例である。

List 1 serps.pm:「Google 日本」に関する定義

```
(省略)
 { ↵
    'search_label' => 'google-jp', ↵
    'base_url' => '^http://www\.google\.co\.jp/search', # manually modified
↵
    'keyword_type' => 'parameter', ↵
    'index_key' => 'start', ↵
    'keyword_key' => 'q', ↵
   observe_keys => [qw/q num/], # manually modified ↵
   preference_url => '^http://www.google.co.jp/setprefs$', # manually
modified ↵
  }, ↵
(省略)
```

2.1 パラメータの説明

サーチエンジンの設定は以下の 7 個のパラメータから構成されている。

- `search_label`
  - サーチエンジンの名称を記述する
  - 設定値は文字列を想定している
  - この値が「検索エンジンラベル」(識別子 selabl)、または「検索エンジンラベル（遷移後）」(識別子 postse)として出力される
- `base_url`
  - SERP (Search Engine Result Page; サーチエンジン結果ページ)の URL マッチングパターンを記述する
  - 設定値は文字列を想定している
  - Perl 正規表現で記述する
  - イベントログの「requestURI」の値が、ここに設定した正規表現にマッチした場合、SERP として判定される
- `keyword_type`
  - キーワードの抽出方式を記述する
  - 抽出されたキーワードが「キーワード」(識別子 kw)、または「キーワード（遷移後）」(postkw)として出力される
  - 以下のいずれかを記述する
    - `parameter`
      - CGI パラメータ(URL クエリ部)にキーワードがある場合に使用する
      - 後述のパラメータ `keyword_key` でキーワードの CGI パラメータ名を指定する

    - in_url
        - URL のパス部にキーワードがある場合に使用する
        - キーワードの範囲は base_url の該当部分を括弧で括って指定する
- keyword_key
    - パラメータ keyword_type に parameter を指定した場合に、キーワードの CGI パラメータ名を記述する
    - 設定値は文字列を想定している
    - 例えば key1=value1&key2=value2&...の value1 がキーワードの場合、keyword_key には”key1”を指定する
- index_key
    - ページ番号のマッチングパターンを記述する
    - 設定値は文字列を想定している
    - 正規表現にマッチした部分が「検索結果のページ番号」(識別子 serpno)、または「検索結果のページ番号（遷移後）」(postno)として出力される
- observe_keys
    - SERP から SERP への遷移が search アクション（検索条件を変更した）か、browse アクション（変更しない）かを判定するために用いる
    - 設定値は文字列の配列を想定している
    - / (スラッシュ)で囲まれた部分に CGI パラメータ名を記述する
    - 記述した名前の CGI パラメータのいずれかが遷移の前後で異なった場合、search アクションとみなされる
    - それ以外は browse アクションとみなされる
    - 複数の CGI パラメータ名を指定する場合、各パラメータの間に半角スペースを挿入する
- preference_url
    - SERP の表示設定などを変更するページ(以後「設定ページ」)からの遷移先 URL を正しく出力するために用いる
    - このパラメータの記述を省略した場合、form の submit URL が「対象 URL」 (識別子 o-url)として出力される。
      この URL に対応するページが画面に表示されるわけではないため、アクションの対象 URL とは異なる
    - 設定ページの form の submit URL を記述する
    - 設定値は文字列を想定している
    - Perl 正規表現で記述する
    - 設定値とイベントログの「requestURI」がマッチした場合、次の遷移先 URL が「対象 URL」 (識別子 o-url)として出力される

2.2 サーチエンジンと判別する場合の解説

実際に検索サイトで検索ボタンをクリック(送信先が SERP の URL パターンに適合する form を submit)したイベント情報を記録したイベントログファイルを読み込み、TSV にアクションとして出力された場合の例を以下に示す。

2.2.1 サーチエンジンの判別

ここでは、Google での検索を記録したイベントログ(List 1)を入力とした場合の例を示す。

List 2 イベントログ(入力)

```
(省略)
{"event_label":"click","target":"[object XPCNativeWrapper [object
HTMLSpanElement]]","target_id":"","button":0,"outerHTML":"<span
class=\"csb\" style=\"background-position: -74px 0pt; width:
20px;\"></span>","modifiers":
{"alt":false,"ctrl":false,"shift":false,"meta":false},"anchor_outerHTML":"<
a href=\"/search?q=%E6%83%85%E5%A0%B1%E5%87%A6%E7%90%86%E6%8A
%80%E8%A1%93%E8%80%85&amp;num=100&amp;hl=ja&amp;client=firefox-
a&amp;rls=org.mozilla:ja:official&amp;prmd=ivns&amp;ei=BbkdUcf5K6OKmwX9v4GI
AQ&amp;start=100&amp;sa=N\" class=\"fl\"><span class=\"csb\"
style=\"background-position: -74px 0pt; width:
20px;\"></span>2</a>","anchor_href":"/search?q=
%E6%83%85%E5%A0%B1%E5%87%A6%E7%90%86%E6%8A
%80%E8%A1%93%E8%80%85&num=100&hl=ja&client=firefox-
a&rls=org.mozilla:ja:official&prmd=ivns&ei=BbkdUcf5K6OKmwX9v4GIAQ&start=100
&sa=N","tab_id":"panel13608292197282","timestamp":1360902417702,"title":"情
報処理技術者 - Google 検索","url":"http://www.google.co.jp/search?q=①
%E6%83%85%E5%A0%B1%E5%87%A6%E7%90%86%E6%8A%80%E8%A1%93%E8%80%85 ⓑ
&oe=utf-8&rls=org.mozilla%3Aja%3Aofficial&hl=ja&client=firefox-a&oq=
%E6%83%85%E5%A0%B1%E5%87%A6%E7%90%86%E6%8A
%80%E8%A1%93%E8%80%85&gs_l=heirloom-
serp.12...0.0.0.73189238.0.0.0.0.0.0.0..0.0...0.0...1c.Df2kI4nrb44"}
(省略)
```

- requestURI の値が、パラメータ base_url に記述した条件「**^http://www\.google\.co\.jp/search**」と一致している(List 2中の 1)

  serp.pm で定義した条件と一致したため、このイベントを SERP に関連するアクションであると判別され、出力情報に SERP に関連情報が出力、追加される

2.2.2 TSV ファイルに出力される付加情報

a) パラメータ search_label に記述した値「google-jp」が「検索エンジンラベル(遷移後)」(識別子 postse)として出力される

b) パラメータ keyword_key に記述した値「q」と名前が一致する CGI パラメータから抽出された値が「キーワード(遷移後)」(識別子 postkw)として出力される(List 2中の b)

2.2.3 TSV ファイルへの出力結果

TSV ファイルへの出力結果は表 1 となる

表 1　List 2 からの出力結果

識別子	出力された値
#time-y	20130215042657.702
action	search
tab-id	1
loadid	1
url	http://www.google.co.jp/search?q=%E6%83%85%E5%A0%B1%E5%87%A6%E7%90%86%E6%8A%80%E8%A1%93%E8%80%85&oe=utf-8&rls=org.mozilla%3Aja%3Aofficial&hl=ja&client=firefox-a&oq=%E6%83%85%E5%A0%B1%E5%87%A6%E7%90%86%E6%8A%80%E8%A1%93%E8%80%85&gs_l=heirloom-serp.12...0.0.0.73189238.0.0.0.0.0.0.0.0..0.0...0.0...1c.Df2kI4nrb44
title	情報処理技術者 - Google 検索
type	serp
postse	**google-jp**
postkw	**情報処理技術者**
postno	100
selabl	google-jp
kw	情報処理技術者
serpno	
anchort	
o-url	http://www.google.co.jp/search?q=%E6%83%85%E5%A0%B1%E5%87%A6%E7%90%86%E6%8A%80%E8%A1%93%E8%80%85&num=100&hl=ja&client=firefox-a&rls=org.mozilla:ja:official&prmd=ivns&ei=BbkdUcf5K6OKmwX9v4GIAQ&start=100&sa=N
bookmk	
object	
form_params	
postti	1
postli	2
time-e	1360902418
dwell	

※TSV 形式で出力されたデータを表形式にまとめている

3 カスタマイズ方法

perlmod/serps.pm の具体的なカスタマイズ手順について実例を挙げながら説明する。

3.1 サーチエンジン「BIGLOBE」を追加する

3.1.1 設定に必要な URL と検索パターンを調べる

- 実際に「http://www.biglobe.ne.jp/」で検索を行い、検索結果の URL を取得する。
  取得した URL は URL 1～3 である

URL 1　トップページから検索文字列候補「キューティーハニー」を選択した場合に表示されたページの URL

```
http://cgi.search.biglobe.ne.jp/cgi-bin/search-sgst-btop?search=%E6%A4%9C
%E7%B4%A2&q=%E3%82%AD%E3%83%A5%E3%83%BC%E3%83%86%E3%82%A3%E3%83%BC%E3%83%8F
%E3%83%8B%E3%83%BC&web_s.x=1&sug_qc=4&sug_is=%E3%81%8D
%E3%82%85%E3%83%BC&st_sug=7
```

URL 2　検索結果画面で再検索した場合に表示されたページの URL

```
http://cgi.search.biglobe.ne.jp/cgi-bin/search_bl_top?q=%E3%82%AD
%E3%83%A5%E3%83%BC%E3%83%86%E3%82%A3%E3%83%BC%E3%83%8F%E3%83%8B
%E3%83%BC&sub=%E6%A4%9C%E7%B4%A2&ie=utf8&num=10&start=0
```

URL 3　トップページから検索文字列入力欄に「キューティーハニー」と入力し Enter キー
または検索ボタンを選択した場合に表示されたページの URL

```
http://cgi.search.biglobe.ne.jp/cgi-bin/search2-b?search=%E6%A4%9C
%E7%B4%A2&web_s.x=1&q=%E3%82%AD%E3%83%A5%E3%83%BC
%E3%83%86%E3%82%A3%E3%83%BC%E3%83%8F%E3%83%8B%E3%83%BC&bt01=%E6%A4%9C
%E7%B4%A2
```

- 検索結果を表示したページにあるリンク「検索オプション」をクリックし、URL を取得する。
  取得した URL は URL 4 である

URL 4　検索オプションページの URL

```
http://search.biglobe.ne.jp/option.html?q=%E3%82%AD%E3%83%A5%E3%83%BC
%E3%83%86%E3%82%A3%E3%83%BC%E3%83%8F%E3%83%8B%E3%83%BC
```

3.1.2 設定に必要な URL と検索パターンを抽出する

- パラメータ `base_url` に使用する URL を抽出する
  - http://cgi.search.biglobe.ne.jp/cgi-bin/search-sgst-btop
  - http://cgi.search.biglobe.ne.jp/cgi-bin/search_bl_top
  - http://cgi.search.biglobe.ne.jp/cgi-bin/search2-b

  ➔ Perl 正規表現で記述すると
  ^http://cgi\.search\.biglobe\.ne\.jp/cgi-bin/(search-sgst-btop|search_bl_top|search2-b)
- キーワードの抽出方式を判定する
  - キーワードの抽出方式は parameter である

  ➔ パラメータ `keyword_type` は「**`parameter`**」
- ページ番号の CGI パラメータを抽出する
  - 「次の 10 件」などでページ遷移し、変動するパラメータを確認する

  ➔ パラメータ `index_key` は「**`start`**」
- キーワードの CGI パラメータの判別
  - 検索キーワード「キューティーハニー」が URL エンコードで変換されたものが含まれる CGI パラメータを確認する

  ➔ パラメータ `keyword_key` は「**`q`**」
- search アクションの CGI パラメータの判別
  - 検索結果表示ページ上で検索キーワードを「きゅーていはにー」に変更して検索を行う。パラメータ `keyword_key` に設定した以外に CGI パラメータの値が変わっているものがあればパラメータとして追加する

  ➔ パラメータ `observe_keys` は「**`q`**」
- パラメータ `preference_url` に使用する URL を抽出する
  - http://search.biglobe.ne.jp/option.html

  ➔ Perl 正規表現で記述すると
  ^http://search\.biglobe\.ne\.jp/option\.html

3.1.3 perlmod/serps.pm を編集する

List 3 serps.pm: サーチエンジン「BIGLOBE」を追加する場合の記述例

a) `our $serps =` に続く[]内に記述する(List 3中の a)

a) 一つの定義は{}内に記述し、複数の定義を記述する場合は
{定義 1},{定義 2},{定義 3}の様に各定義間に,(カンマ)を記述する(List 3中の b)

1. パラメータ `search_label` に「biglobe」を記述する(List 3中の 1)
2. パラメータ `base_url` に「^http://cgi\.search\.biglobe\.ne\.jp/cgi-bin/(search-sgst-btop|search_bl_top|search2-b」を記述する(List 3中の 2)
3. パラメータ `keyword_type` に「parameter」を記述する(List 3中の 3)
4. パラメータ `index_key` に「start」を記述する(List 3中の 4)
5. パラメータ `keyword_key` に「q」を記述する(List 3中の 5)
6. パラメータ `observe_keys` のスラッシュ間に「q」を記述する(List 3中の 6)
7. パラメータ `preference_url` に「^http://search\.biglobe\.ne\.jp/option\.html$」を記述する(List 3中の 7)

3.2 定義を追加後の入出力結果

実際に BIGLOBE で「キューティーハニー」の検索を行い、そのイベント情報を記録したイベントログファイルを読み込み、TSV にアクションとして出力するまでの例を以下に示す。

3.2.1 サーチエンジンの判別

List 4 イベントログ(入力)

```
(省略)
{"event_label":"click","target":"[object XPCNativeWrapper [object
HTMLInputElement]]","target_id":"search-btn","button":0,"outerHTML":"<input
name=\"bt01\" value=\"検索\" id=\"search-btn\"
type=\"submit\">","modifiers":
{"alt":false,"ctrl":false,"shift":false,"meta":false},"type":"submit","tab_
id":"panel1366885592981","timestamp":1366885655261,"title":"BIGLOBE","url":
"http://www.biglobe.ne.jp/"}
{"form_id":"srch","eventType":"submit","form_action":"http://cgi.search.big
lobe.ne.jp/cgi-bin/search2-
b","form_name":"srch","tab_id":"panel1366885592981","submit_data":"search=
%E6%A4%9C%E7%B4%A2&web_s.x=1&q=%E3%82%AD%E3%83%A5%E3%83%BC
%E3%83%86%E3%82%A3%E3%83%BC%E3%83%8F%E3%83%8B%E3%83%BC&bt01=%E6%A4%9C
%E7%B4%A2","timestamp":1366885655263,"url":"http://www.biglobe.ne.jp/","tit
le":"BIGLOBE"}
{"eventType":"http_req","requestURI":"http://cgi.search.biglobe.ne.jp/cgi-
bin/search2-b?search=%E6%A4%9C%E7%B4%A2&web_s.x=1&q=%E3%82%AD
%E3%83%A5%E3%83%BC%E3%83%86%E3%82%A3%E3%83%BC%E3%83%8F%E3%83%8B
%E3%83%BC&bt01=%E6%A4%9C
%E7%B4%A2","method":"GET","http_req":"Host:cgi.search.biglobe.ne.jp\u000aUs
er-Agent:Mozilla/5.0 (Windows; U; Windows NT 5.1; ja; rv:1.9.2.28)
Gecko/20120306 Firefox/3.6.28 ( .NET CLR
3.5.30729)\u000aAccept:text/html,application/xhtml+xml,application/xml;q=0.
9,*/*;q=0.8\u000aAccept-Language:ja,en-us;q=0.7,en;q=0.3\u000aAccept-
Encoding:gzip,deflate\u000aAccept-Charset:Shift_JIS,utf-
8;q=0.7,*;q=0.7\u000aKeep-Alive:115\u000aConnection:keep-
alive\u000aReferer:http://www.biglobe.ne.jp/\u000aCookie:BLS_SGF=1;
web_ABtest_flg35=100;
__utma=103156754.84371860.1366885640.1366885640.1366885640.1;
__utmb=103156754.2.10.1366885640; __utmc=103156754;
__utmz=103156754.1366885640.1.1.utmcsr=biglobe.ne.jp|utmccn=(referral)|
utmcmd=referral|
utmcct=/","tab_id":"panel1366885592981","timestamp":1366885655277,"title":"
BIGLOBE","url":"http://www.biglobe.ne.jp/"}
(省略)
```

- requestURI の値が、パラメータ base_url に記述した条件「**^http://cgi\.search\.biglobe\.ne\.jp/cgi-bin/(search-sgst-btop|search_bl_top|search2-b)**」と一致している(List 4中の 1)。

  serp.pm で定義した条件と一致したため、このイベントを SERP に関連するアクションであると判別され、出力情報に SERP の関連情報が出力、追加される

3.2.2 TSV ファイルに出力される付加情報

a) パラメータ `search_label` に記述した値「biglobe」が「検索エンジンラベル（遷移後）」(識別子 postse)として出力される

b) パラメータ `keyword_key` に記述した値「q」と一致するパラメータから出力した値が「キーワード（遷移後）」(識別子 postkw)として出力される(List 4 中の b)

3.2.3 TSV ファイルへの出力結果

TSV ファイルへの出力結果は表 2 となる

表 2 List 4 からの出力結果

識別子	出力された値	
	アクション1	アクション2
#time-y	20130425102735.2	20130425102737.9
action	search	end
tab-id	1	1
loadid	1	2
url	http://www.biglobe.ne.jp/	http://cgi.search.biglobe.ne.jp/cgi-bin/search2-b?search=%E6%A4%9C%E7%B4%A2&web_s.x=1&q=%E3%82%AD%E3%83%A5%E3%83%BC%E3%83%86%E3%82%A3%E3%83%BC%E3%83%8F%E3%83%8B%E3%83%BC&bt01=%E6%A4%9C%E7%B4%A2
title	BIGLOBE	“キューティーハニー”で検索した結果
type	non_serp	serp
postse	**biglobe**	**biglobe**
postkw	キューティーハニー	キューティーハニー
postno		
selabl		biglobe
kw		キューティーハニー
serpno		
anchort		
o-url	http://cgi.search.biglobe.ne.jp/cgi-bin/search2-b	
bookmk		
object	srch	
form_params	{'web_s.x' => '1','search' => '検索','q' => 'キューティーハニー','bt01' => '検索'}	
postti	1	1
postli	2	2
time-e	1366885655	1366885658
dwell		2.726

※TSV 形式で出力されたデータを表形式でまとめている

4 付録

4.1 凡例

- ↵は改行(Enter/Return)キーを表す