De Webxref aux traits

SF 02.06.2004

(ce texte au format PDF)

1 Rappels : sorties webxref

1.1 Fichiers de Stat

La version de webxref utilisée pour le projet SENSNET reprend la version ultime de ce programme utilisée dans le projet Typweb (Webxref-038) mais se distingue de celle-ci par l'insertion de fonctionnalités concernant la création d'indicateurs de qualité (pour les pages web scrutées), en s'inspirant des indicateurs utilisés dans WebTango (http://webtango.ischool.washington.edu/).

Dans cette version, 4 "fichiers de stat" sont créés à l'issue du traitement réalisé:

StatWordByFile.txt
StatWordFull.txt
StatQualityByFile.txt
StatQualityMeanByFile.txt

1.1.1 StatWord

StatWordByFile.txt
StatWordFull.txt

Ces 2 fichiers contiennent un état statistique du nombre de `tags HTML` et de mots dans les pages scrutées par webxref (le premier fichier construit cet état pour chaque page HTML, le second le construit pour le site complet). Le deux fichiers ont l'allure suivante :

```
<TAGS>
<SITE>sitename</SITE>                                   ; nom du site
<PAGE>pagename</PAGE>                                   ; nom de la page
<ELEMENTS>
<ITEM>P</ITEM><FRQ>17</FRQ>
...                                                     ; liste et fréquence des éléments HTML
</ELEMENTS>
<ELEMENTS_ATTR>
<ITEM>FONT(COLOR)</ITEM><FRQ>10</FRQ>
<ITEM>FONT(FACE)</ITEM><FRQ>9</FRQ>
...                                                     ; liste et fréquence des élements+attributs
...
</ELEMENTS_ATTR>
<ELEMENTS_ATTRVALUE>
<ITEM>FONT(FACE=ARIAL)</ITEM><FRQ>7</FRQ>
<ITEM>IMG(ALIGN=BOTTOM)</ITEM><FRQ>4</FRQ>
```

```
...                                                    ; liste et fréquence des élements+attributs+valeurs
</ELEMENTS_ATTRVALUE>
<WORDS>
<SITE>res1$jura$speleo1</SITE>
<PAGE>menu_acc</PAGE>
...                                                    ;;; liste et fréquence des words
</WORDS>
```

1.1.2 StatQuality

StatQualityByFile.txt
StatQualityMeanByFile.txt

Ces 2 fichiers contiennent un descriptif de certains ***traits calculés*** par webxref, le contenu de ces fichiers a l'allure suivante (première ligne de la matrice contenue dans ce fichier) :

```
Filename       wordsinTAG-A   wordsinTAG-B   imageinTAG-A   wordsinTAG-h1  wordsinTAG-h2
       wordsinTAG-h3  wordsinTAG-h4  wordsinTAG-h5  wordsinTAG-h6  wordsinTAG-h7  wordsinTAG-TD
       wordsinTAG-P   wordsinTAG-TABLE      wordsinTAG-FONT
```

La matrice contient donc les valeurs des indicateurs créés pour chaque site/page scruté par webxref (décompte brut pour le premier fichier et moyenne pour le second), ces indicateurs ont pour valeur :

WordsinTAG-A	nombre de mots dans un tag A
WordsinTAG-B	nombre de mots dans un tag B
ImageinTAG-A	nombre d'image dans un tag A
WordsinTAG-h1	nombre de mots dans un tag H1
WordsinTAG-h2	nombre de mots dans un tag H2
WordsinTAG-h3	nombre de mots dans un tag H3
WordsinTAG-h4	nombre de mots dans un tag H4
WordsinTAG-h5	nombre de mots dans un tag H5
WordsinTAG-h6	nombre de mots dans un tag H6
WordsinTAG-h7	nombre de mots dans un tag H7
WordsinTAG-TD	nombre de mots dans un tag TD
WordsinTAG-P	nombre de mots dans un tag P
WordsinTAG-TABLE	nombre de mots dans un tag TABLE
WordsinTAG-FONT	nombre de mots dans un tag FONT

1.2 Traits calculés et intégrés dans la base Sensnet

Webxref calcule aussi des décomptes pour un certain nombre de traits, ces décomptes sont intégrés directement dans la base Sensnet. Ces traits donne des indications sur le nombre de liens (interne, externe), d'images (interne, externe) etc.

1.2.1 Typage des liens et décomptes

Rappel : "Désossage" des pages par webxref[1]

L'élément <*html*> de chaque document et ses descendants subissent un traitement sélectif, selon qu'ils constituent des « déclarations » (habituellement la déclaration du type de document), des « tag objects » (balises ouvrantes ou fermantes), des « text objects » (séquences textuelles) ou des « comments » (des balises de commentaire). La sous-fonction *elements* appelée à ce point fait appel à une version adaptée du parser distribué avec *Perl5* pour *Windows* (*HTML::Parser.pm*) pour typer ces éléments et les imprimer après encodage-décodage des entités HTML. Les attributs *SGML* des « tag objects » et les séquences

[1] http://www.cavi.univ-paris3.fr/ilpga/ilpga/sfleury/typweb/manual/WEBXREF-Typweb.htm

textuelles seront affichés dans le format demandé (*.html* ou *.txt*) selon la grammaire du script *dissectsite.hts* (cf. Mode d'emploi webxref en ligne).

Les attributs des liens sont soumis à un test pour le typage. La typologie est reprise en partie à l'auteur de *Webxref035.pl.*

Procédure ***links*** de *webxref* : Applique la routine de typage selon que les attributs des liens sont reconnus. Les attributs des liens sont soumis à un test pour le typage. La typologie est reprise en partie à l'auteur de *Webxref035.pl*.

Algorithmes:

```
* mailto: <mailto:>
* cgi-bin: <cgi-bin>
* gopher: <gopher:>
* news: <news:>
* lien ftp : <ftp:>
* image : <img src>
* lien http : <http:>
* lien hypertextuel: <link rel><link rev>
* lien interne vers fichier : <file:>
* lien vers ancre : <#>
```

Les liens typés sont imprimés dans le fichier de rapport selon la grammaire :

numéro courant
classe
type
attributs

```
LINK_INTERNALIMAGE            $imgNb      $imgIntNb
LINK_EXTERNALIMAGE            $imgNb      $imgExtNb
LINK_INTERNALFILE             $linkNb     $intLinkNb
LINK_EXTERNALHYPERTEXTUAL     $linkNb     $extHypertextLinkNb
LINK_INTERNALHYPERTEXTUAL     $linkNb     $intHypertextLinkNb
LINK_INTERNALHTMLFILE         $linkNb     $intHtmlFileLinkNb
LINK_INTERNALDOCFILE          $linkNb     $intDocFileLinkNb
LINK_INTERNALANCHOR           $linkNb     $intAncNb
LINK_EXTERNALANCHOR           $linkNb     $extAncNb
LINK_EXTERNALMAILTO           $linkNb     $extMailNb
LINK_EXTERNALHTTP             $linkNb     $extLinkNb
LINK_EXTERNALHTTP_CGI         $linkNb     $extCgiLinkNb
LINK_EXTERNALHTTP_NEWS        $linkNb     $extNewsLinkNb
LINK_EXTERNALHTTP_FTP         $linkNb     $extFtpLinkNb
LINK_EXTERNALHTTP_GOPHER      $linkNb     $extGopherLinkNb
```

1.2.2 Descriptif des traits calculés

Les traits calculés par webxref associés au processus décrit dans la partie précédente sont donc :

Lien vers image

`ImgIntNb` *: compteur image interne*
`imgExtNb` *: compteur image externe*

Liens internes
`intLinkNb` *: compteur lien interne (type* **`<file >`** *obsolète a priori)*
`intHypertextLinkNb` *: compteur lien interne (type* **`<link href>`***)*
`intHtmlFileLinkNb` *: compteur lien interne(lien interne classique vers fichier HTML)*
`intDocFileLinkNb`*: compteur lien interne (lien interne classique vers fichier non HTML)*

liens externes
`extHypertextLinkNb`*: compteur lien externe (type* **`<link href=http>`***)*
`extLinkNb` *: compteur lien externe (lien externe classique type* **`<a href=http>`***)*

Liens "ancre"
`intAncNb`*: compteur ancre interne*
`extAncNb`*: compteur ancre externe*

Liens "mail"
`ExtMailNb`*: compteur lien mail*

Lien "cgi"
`extCgiLinkNb`*: compteur lien cgi*

Liens "news"
`extNewsLinkNb`*: compteur lien news*

Liens "Ftp"
`extFtpLinkNb`*: compteur lien ftp*

Liens "gopher"
`extGopherLinkNb`*: compteur lien gopher*

Compteurs globaux
`linkNb`*: nb global de lien (somme de 13 types de lien précédents)*
`imgNb`*: nb global d'image (somme de 2 types d'image précédents)*

Le comptage de liens internes est donc la somme des 4 types de liens internes définis ci-dessus, pour les liens externes, idem avec les 2 types de liens externes définis.

Remarque : dans le projet Typweb, le comptage des liens (interne, externe, mail etc.) était réalisés par un script Perl qui parcourait le fichier `StatWordByFile.txt` pour constituer ces indicateurs (cf infra).

2 Exemple : sorties webxref sur échantillon de pages de JFV

(un peu plus de 1700 pages)

2.1 Extraction des traits calculés par webxref

Vous trouverez ici une matrice reprenant l'ensemble des traits présentés supra. La première colonne correspond à un nom de chemin des rapports associés aux pages de ce échantillon de page, la page est référencée dans la seconde colonne.

Version Excel
Version HTML

Remarque :

- dans ce fichier les 3 colonnes "Liens externes", "Liens internes", "Liens ancre" ne correspondent pas à des indicateurs directement calculés par webxref (contrairement aux autres colonnes), ce sont des colonnes construites ici sur la base de sommes de colonnes idoines.
- Il y a 2 traits calculant le nombre global d'image sur la page : dans le tableau présenté ci-dessus, la colonne "image Desc" est la somme des colonnes précédentes (EXT et INT), la colonne "Nb image" est associée à un indicateur particulier.

2.2 Extraction des traits à partir du fichier StatWordByFile

A l'image de ce qui était fait dans le projet TYPWEB notamment pour le calcul des traits de type LIEN, un programme parcourt le fichier StatWordByFile pour compter les occurrences d'un certain nombre de traits (nb de lien, images, mots…). L'objectif est ici en partie de vérifier les comptages de traits précédents et de calculer des traits complémentaires (comptage des mots en particulier).

Vous trouverez ici une matrice contenant l'ensemble des traits visés ici :

Version Excel
Version HTML

Remarques :

- dans ce fichier la colonne "Liens externes" est la somme des 2 colonnes précédentes,
- la colonne"lienjavascript" correspond a priori à un lien qui serait classé en lien interne en 2.1,
- la colonne "lienlink(ext/int)" correspond à des définitions de lien vers des feuilles de style par exemple,
- la colonne "lienautre (interne)" contient tous les liens qui ne sont pas reconnus dans les autres colonnes, a priori les liens internes se retrouvent donc ici
- la colonne "Liens internes" est la somme de 3 colonnes la précédant (i.e. avec le lien javascript)
- la colonne "Liens ancre" est la somme de 2 colonnes la précédant
- la colonne "nb images" est la somme de 2 colonnes la précédant
- les 2 dernières colonnes concernant le nombre de mots par page : nb de formes, nb d'occurrences

2.3 Extraction des traits à partir du corpus XML

Enfin, en reprenant le programme utilisé dans TYPWEB pour le comptage des liens qui essaie d'identifier un certain nombre de traits pour comptage (liens, images), on obtient la matrice suivante :

Version HTML
Version Excel

Dans ce fichier, on distingue le type de lien et sa répétition éventuelle (les occurrences du même lien), c'est ce dernier indicateur qui est à rapprocher des précédents indicateurs similaires.