Oct 2000

WalkAll
(Suite de Broad HTML Analyzer)

1# Problématiques diverses: cas réels

1.1# Le contenu est en plusieurs morceaux

```
<tr>
  <td></td>
  <td></td>
  <td></td>
</tr>
```

Impossible visuellement de récupérer le contenu de <tr>. Seul <td>, le premier, le second, le troisième, sera obtenu.

1.2# Les titres des news sont composés d'un gros titre et de n-1 titres

Le gros titre est à entendre par un titre avec l'utilisation d'un style ou d'une fonte de taille particulière. Les n-1 autres titres ont une autre taille, identique entre elles, mais différente de celle du gros titre. Que faire ?

- Garder la présentation, quitte à ce que le code agrégé ne soit qu'une recopie
- Reformuler la présentation des titres, ce qui suppose l'utilisation d'un parser HTML complet ainsi qu'un outil de modification relativement intelligent et complexe.

1.3# Lorsqu'une news comporte un lien hypertexte, avec un lien relatif

Se contenter de récupérer le lien ne permet pas d'accéder à la news à partir d'un site d'aggrégation de contenu.

Exemple : <a href="/news/technology/0,1282,39235,00.html" target="_top">

Il faut reformuler **href**, de façon à ce que tout se passe comme si nous avions <a href="http://www.wired.com/news/technology/0,1282,39235,00.html" target="_top">, lorsque ce lien provient de http://www.wired.com.

Deux solutions à cela :

- exploiter un parser HTML complet, et complexe à mettre en œuvre.
- exploiter le fait que le DOM en mémoire n'est pas nécessairement équivalent au code HTML :

Explications de la seconde solution : imaginons que nous aggrégeons le site Wired dans une page web finale, peu importe son nom, dans laquelle on trouvera un bloc de news spécifique à Wired. Nous savons que dans ce bloc, tous les liens sortants vont vers le site Wired, sauf ceux qui sont exprimés en absolu avec un http://www…

Oct 2000

Le D.O.M. en mémoire a, une fois la page web chargée, tous ces liens en mémoire. Il ne tient qu'à nous de mettre en œuvre une petite fonction en Javascript qui, une fois la page chargée parcourt tous les liens d'un bloc donné, et concatène à tous les liens relatifs du bloc la particule principale http://www.wired.com, connue à l'avance. Lorsqu'un lien est exprimé en absolu, il n'y a pas de traitement à faire.

1.4# Lorsqu'une news comporte un lien vers une procédure javascript

Exemple : sur le site de l'INA, une offre d'emploi pointe vers **JavaScript:go_submit(20278)**. Pour que ce lien fonctionne, il faut capturer la procédure go_submit. La procédure go_submit peut elle même référencer des tags HTML non capturés. Ce qui est le cas en pratique puisque go_submit valide un formulaire en ayant préalablement intégré l'identifiant 20278 en tant que paramètre.

Impossible de s'en sortir sans tout capturer : la procédure Javascript et le formulaire HTML.

Ce processus est très complexe et empêche en pratique de manipuler une base de données très simple comportant des chemins d'accès de tags pour chaque site capturé.

2# Capture des données

Si le but est d'aggréger du contenu provenant de sites de news alimentés quotidiennement, il faut procéder quotidiennement à une capture. La capture est un processus offline qui extirpe pour un site de news donné (site de news au sens large du terme) un ensemble de titres.
Pour accéder à chaque titre, le processus accède au D.O.M., à travers C++ et dans un contexte hors navigateur web.
Chaque site de news comporte un certain nombre de titres sur sa page d'accueil, mais il n'est pas nécessaire de spécifier le nombre exact de titres à capturer car une modélisation mathématique permet d'établir la relation entre deux titres qui se suivent, et par itération obtenir tous les titres, quel qu'en soit le nombre.

2.1# Paramétrage

Le processus est lancé à l'aide d'un programme exécutable et d'un fichier de description. Ce fichier de description est susceptible d'évoluer au gré de l'ajout de nouveaux sites de news, ou de la refonte de certains déjà capturés. C'est un fichier texte aisément compréhensible et manipulable.

Il comprend la modélisation de l'ensemble des sites de news à capturer, c'est-à-dire *n* blocs descriptifs élémentaires.

2.2# Modélisation d'un site de news

Les news sont affichées sur une page, dite page principale, d'un site de news. L'hypothèse forte que nous faisons afin de pouvoir automatiser un processus de capture, c'est qu'au fil des jours la structure de la page du site ne change pas fondamentalement. Seule la forme varie, c'est-à-dire le contenu.
Un site portail est refondu dans le pire des cas tous les 2 mois, et en moyenne tous les 6 mois ou tous les ans. Notre fichier de description a donc toutes les chances d'être valable un certain temps.
Les news sont généralement disposées le long d'une bande verticale. Les news sont éventuellement (et même souvent) cliquables et mènent alors à une page spéciale.

2.2.1# Les chemins d'accès

Parce qu'il y a une bande verticale, cela veut dire que chaque news partage un tag père commun. Ce tag père commun est un tag que nous indexons.

Exemple :

```
<table>
 <tr>
   <td> news 1 : oitueiortuoiert oery </td>
   <td> news 2 : ieyiuzytiur y uiyiu   </td>
   <td> news 3 : trpotureioturoet   </td>
   <td> news 4 : ksjdhfiuyturt  </td>
   <td> news 5 : poypuirtopyrty  </td>

 </tr>
</table>
```

Dans notre exemple, le tag père commun est le tag <tr>. Son chemin d'accès dans notre exemple est 0;0;0 (et non 0;0 du fait de l'insertion d'un tag TBODY que l'on ne voit jamais).

On ne capture qu'une série de news par site.

Chaque news a une relation privilégiée avec la précédente. Et la première news est une "headline" et joue elle aussi une relation privilégiée avec les autres.

Notre modélisation revient à indiquer les chemins d'accès des 3 premières news.

Le chemin d'accès de la première news est indiqué, car c'est une news, et que la première news joue fréquemment un rôle particulier par rapport aux autres, notamment sa présentation est souvent différente.

Les chemins d'accès des deux news suivantes sont indiquées. Le but est à la fois d'avoir l'accès direct à celles-ci, mais aussi, par "soustraction" entre ces deux chemins d'accès, nous avons le moyen d'accéder à la quatrième news, à la cinquième, etc. jusqu'à plus soif.

Pourquoi la troisième news est-elle nécessaire, alors que manifestement on pourrait très bien soustraire les chemins d'accès des deux premières ? Tout simplement, la première news joue un rôle particulier. Il est donc fort à parier que la soustraction entre la première et la deuxième soit différente de la soustraction entre la deuxième et la troisième. Mais comme ce sont les news autres que la première qui constituent l'ossature de la bande verticale, il est assez logique de faire ce choix.

Exemple :

```
        news 1   0;0;0;1
        news 2   0;0;0;2
        news 3   0;0;0;3

        soustraction 1-2   0;0;0;1
        soustraction 1-3   0;0;0;1
```

Cet exemple est simple et ne particularise pas la première news.

Exemple :

```
<table>
 <tr>
  <td>
   <font size=4>news 1</font><br><br>
   <font size=2>news 2 </font><br>
   <font size=2>news 3 </font><br>
```

```
    <font size =2>news 4 </font><br>
    <font size=2>news 5 </font><br>
   </td>
  </tr>
</table>
```

Cet exemple particularise la première news. La modélisation est la suivante :

news 1 0;0;0;0;0
news 2 0;0;0;0;3
news 3 0;0;0;0;5

soustraction 1-2 0;0;0;0;3
soustraction 2-3 0;0;0;0;2

Et l'on constate bien qu'il faut tenir compte de la soustraction entre la deuxième et la troisième news puisque les news 4 et 5 se trouvent respectivement aux indices 0;0;0;0;7 et 0;0;0;0;9.

Autre exemple, qui pose problème :

Imaginons une bande verticale de news non structurées, c'est-à-dire pour lesquelles il n'y a pas de tag englobant les unes par rapport aux autres :

```
<table>
 <tr>
  <td>
   news 1<br><br>
   news 2<br>
   news 3<br>
   news 4<br>
   news 5<br>
  </td>
 </tr>
</table>
```

Le problème est qu'il est tentant d'indiquer le chemin d'accès 0;0;0;0 pour la news 1, ainsi que 0;0;0;0;1 pour la news 2, etc. Mais cela ne marche pas du tout. En particulier, pour la news 1, le tag fermant associé à 0;0;0;0 englobe toutes les news !!
Et les autres news n'ont pas de tag fermant puisque un tag </br> est facultatif, voire vide de sens.

La meilleure solution ici est de ne capturer que la news 1, et capturer le bloc entier.

Il faut tenir compte de ce cas de figure dans notre modélisation.

Par analogie, une bande verticale peut comporter moins de 3 news. A modéliser aussi.

Ces deux dernières remarques nous conduisent à adjoindre aux chemins d'accès un paramètre. Ce paramètre vaut n lorsque nous sommes en présence d'une bande verticale très générale, où l'on sait que l'on trouvera un certain nombre de news. Ce paramètre vaut 1, 2, 3… si nous devons nous limiter à une, deux, trois…. news à capturer.

Pour le premier exemple, la modélisation complète est :

Bandeaux de news

http://www.wired.com
n
0;0;0;1
0;0;0;2
0;0;0;3

Pour le second exemple, la modélisation complète est :

News 1 particulière

http://www.wired.com
n
0;0;0;0;0
0;0;0;0;3
0;0;0;0;5

Pour le troisième exemple :

News sans séparation

http://www.wired.com
1
0;0;0;0

D'autres cas particuliers peuvent se présenter, et le présent document se veut la référence de tous les cas possibles du "monde réel". Dans le monde réel, les news renvoient via des liens vers des pages de contenu. Cela veut dire qu'il y a toutes les chances de trouver pour chaque news des tags englobants du type news 1. Par conséquent, le troisième exemple, qui pose un vrai problème de fond, a peu de chances de se trouver sur des sites réels.

De plus, indépendemment de cette dernière remarque, le troisième exemple est symptomatique du cas général de présentation de news au sein de code HTML. Il ne peut être traité sereinement que dans un cadre très général, qui appelle une autre modélisation, et une nouvelle approche. La version actuelle du logiciel WalkAll ne l'envisage pas.

2.2.2# L'accès à la page de news

La problématique est la suivante : un site est "taggé" par son DNS principal, exemple www.wired.com

Mais la page de news peut être située, et c'est même quasiment exclusivement le cas, à un autre point d'entrée que http://www.wired.com. Comme par exemple http://www.wired.com/news4/index.html
Comme la capture suppose d'accéder à la page de news, et que cet accès doit être facilité le plus possible, il faut d'une manière ou d'une autre en réalité indiquer à la fois le DNS principal, et le point d'entrée aux news.

De plus, un problème supplémentaire est que la page de news peut être dynamique y compris du point de vue de son nom. On pourrait très bien trouver une page de news de la forme http://www.wired.com/news4/24_octobre_2000.html. Comme le nom de cette page change tous les jours, c'est un non sens que de tagger un site avec un nom de page de news complètement statique. L'intérêt d'un nom dynamique est de garantir l'accès aux news de la veille, le lendemain. Certains sites de news écrasent en effet purement et simplement les news tous les jours !

Il y a toutefois une logique. Un site de news comporte les news sur sa page principale, ce qui veut dire qu'en tapant http://www.wired.com, fondamentalement la logique de redirection se fait toute seule et conduit à http://www.wired.com/news4/24_octobre_2000.html. Pour accéder à la page de news, il n'est donc pas nécessaire de connaître effectivement cette page de news. Tout rendre en ordre, car du point de vue de notre outil de capture, tout se passe comme si la page de news était http://www.wired.com.

Un cas de figure qui ne fonctionne pas, mais qui est très rare est celui qui consiste à ce que la homepage du site soit un frameset, c'est-à-dire la déclaration d'un ensemble de cadres. Et que l'un de ces cadres soit effectivement la page de news. Si nous indiquons à la capture la homepage du site, nous n'aurons l'accès qu'au frameset, et pas à la page de news.
Pour se sortir du problème, il faut, à la main, étudier préalablement le site, afficher la source HTML correspondant au frameset et détecter le cadre correspondant à la page de news. Mais il faut prier pour que cette page de news comporte un nom statique. Si le nom est statique, comme dans news.html, alors il suffit d'indiquer

Oct 2000

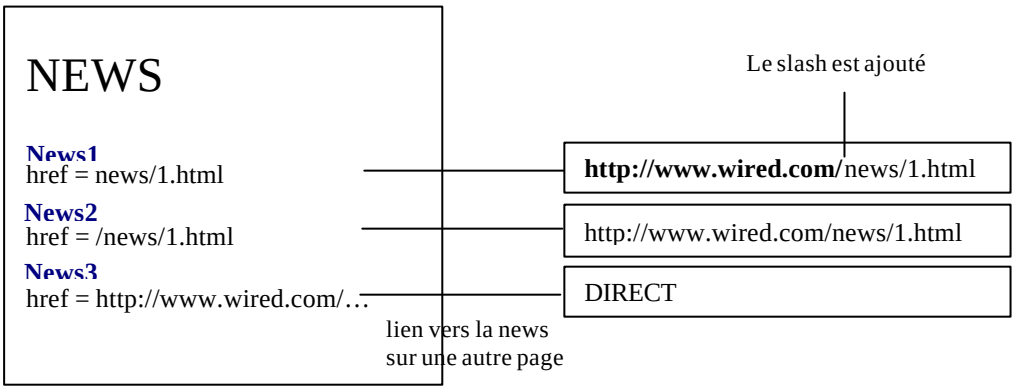
finalement que la page de news est http://www.wired.com/news.html, alors que si le nom de page est dynamique, il faut bien qu'il y ait un processus capable de comprendre un frameset, de fouiner l'ensemble des cadres et chercher parmi ceux-ci quel est celui qui héberge le site de news, pour enfin accéder aux news elles-mêmes. On l'a compris, cela revient à simuler du code HTML. Compliqué, mais c'est un cas réel.
Ce dernier cas, problématique, appelle une évolution du logiciel WalkAll. A prévoir.

En résumé, nous avons 3 cas de figure.

Cas 1 : la page de news est en accès direct à partir de la homepage

HOMEPAGE : http://www.wired.com
remappé en http://www.wired.com/index.html (page par défaut du serveur web)

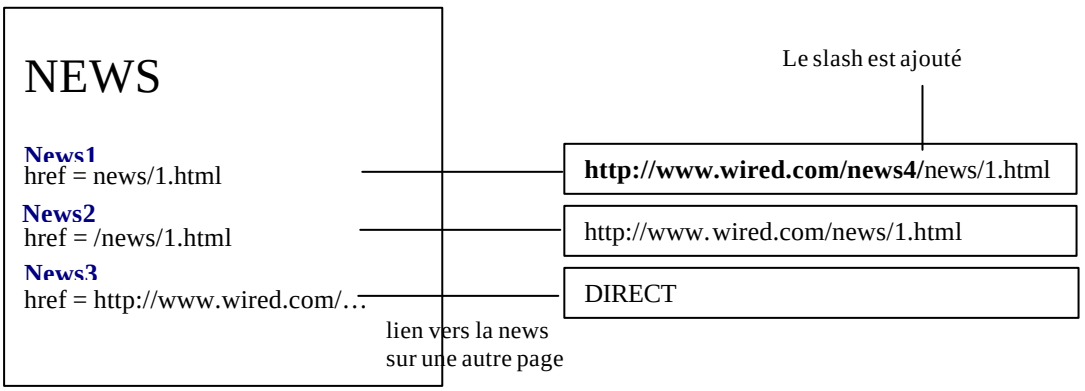
PAGE DE NEWS : http://www.wired.com/index.html



Cas 2 : la page de news n'est pas la homepage

HOMEPAGE : http://www.wired.com
remappé en http://www.wired.com/index.html (page par défaut du serveur web)

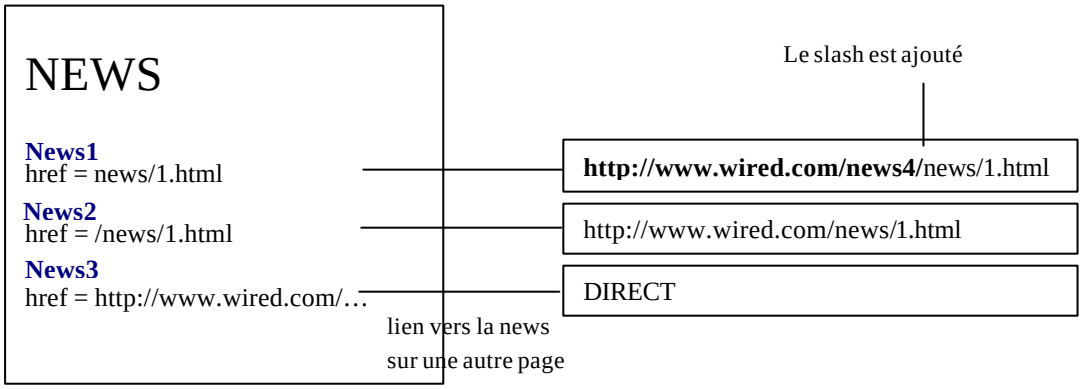
PAGE DE NEWS : http://www.wired.com/news4/index.html



Cas 3 : la page de news n'est pas la homepage et a un nom dynamique

HOMEPAGE : http://www.wired.com
remappé en http://www.wired.com/index.html (page par défaut du serveur web)

PAGE DE NEWS : http://www.wired.com/news4/24_oct_2000.html



Ce que les 3 cas de figure montrent, c'est qu'il peut y avoir deux distinctions :

- la première entre la homepage du site et la page de news. La première doit être indiquée car elle est utilisée notamment pour construire un lien absolu à partir d'un lien relatif commençant par un / (cf. News 2). La seconde doit être indiquée car la capture se base sur celle-ci.
- Pour obtenir un lien absolu à partir d'un lien relatif, il faut choisir en fonction du type de relativité. S'il y a un slash, c'est que la ressource HTML est accessible à partir de la racine du site. S'il n'y en a pas, il faut partir de la page courante, c'est-à-dire la page de news.

En conclusion, nous devons adjoindre les deux informations homepage et page de news.

Mais ce qui vient à notre secours, c'est que la page de news est une URL qui elle-même référence le DNS principal.

Ce qui pour nos 3 exemples nous donne :

Bandeaux de news

http://www.wired.com/index.html
n
0;0;0;1
0;0;0;2
0;0;0;3

Pour le second exemple, la modélisation complète est :

News 1 particulière

http://www.wired.com/news4/index.html
n
0;0;0;0;0
0;0;0;0;3
0;0;0;0;5

Pour le troisième exemple :

News sans séparation

http://www.wired.com/news4/24_oct_2000.html
1
0;0;0;0

2.2.3# Nom final pour le site

Chaque site capturé possède un nom. Les news sont capturées et agrégées à leur tour dans de nouvelles bandes verticales. Ce sont en fait des tableaux. Et pour distinguer les tableaux il convient de décorer un "site", résumé par ses news du jour, par un nom, voire un template HTML.

Le template HTML appelle une nouvelle version de WalkAll. Une version dans laquelle la génération de la page web finale est paramétrable et ouverte, notamment par un code source en plugin.

Mais rien n'empêche dans un premier temps d'indiquer un nom. De plus, ce nom peut être exprimé sous format HTML. On peut aussi bien mettre Wired, comme **Wired news**. Ce qui ne gâche rien.

Pour nos 3 exemples, cela donne une entrée supplémentaire :

Bandeaux de news

Wired news
http://www.wired.com/index.html
n
0;0;0;1
0;0;0;2
0;0;0;3

Pour le second exemple, la modélisation complète est :

News 1 particulière

Wired news
http://www.wired.com/news4/index.html
n
0;0;0;0;0
0;0;0;0;3
0;0;0;0;5

Pour le troisième exemple :

News sans séparation

Wired news

Oct 2000

http://www.wired.com/news4/24_oct_2000.html
1
0;0;0;0

2.2.4# Code ouvert

Les descriptifs sont "propriétaires", c'est-à-dire qu'il n'est pas du tout évident de l'extérieur de comprendre comment ils sont agencés. De façon à préparer une évolution des propriétés de la capture, des pré-traitements et des post-traitements, il est très souhaitable de choisir en fait de déclaration des règles de production.

De cette manière, chaque règle de production est cataloguée de manière homogène en interne dans le logiciel de capture, et si le code doit évoluer ou s'ouvrir, notamment via des plugins, l'approche structurée aura grandement facilité son exploitation maximale.

Pour nos 3 exemples, cela donne :

Bandeaux de news

site=Wired news
newspage=http://www.wired.com/index.html
amount=n
news=0;0;0;1
news=0;0;0;2
news=0;0;0;3

Pour le second exemple, la modélisation complète est :

News 1 particulière

site=Wired news
newspage=http://www.wired.com/news4/index.html
amount=n
news=0;0;0;0;0
news=0;0;0;0;3
news=0;0;0;0;5

Pour le troisième exemple :

News sans séparation

site=Wired news
newspage=http://www.wired.com/news4/24_oct_2000.html
amount=1
news=0;0;0;0

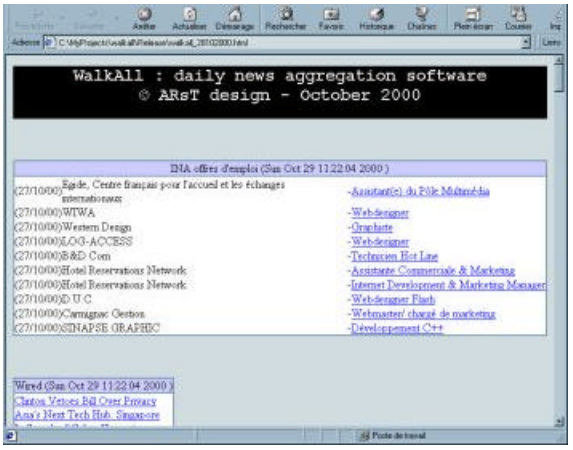
2.3# Fichier de description

Le fichier de description n'est autre que la concaténation de tous les descripteurs de sites.

Exemple :

Oct 2000

2.4# Walkall, écrit en C++



Ce que produit WalkAll, une aggrégation de contenu.
Un point d'entrée unique pour une infinité de sites

2.4.1# Accès à IE et au DOM

Capturer un site de news revient à capturer puis exploiter une ou plusieurs pages web. Il faut se donner les moyens de charger en mémoire une page web. C'est IE qui fait ce travail, à l'aide d'un moniker URL. Nous nous contentons de lui donner l'ordre de charger la page, et nous enregistrons pour être notifiés des évènements : erreur, chargement en cours, chargement terminé.

L'accès à InternetExplorer sans interface graphique se fait à l'aide d'un composant COM : CLSID_HTMLDocument. L'interface principale est IHTML_DOCUMENT2. Cette interface n'est autre que le D.O.M.

L'environnement de développement InternetClient SDK, qui fait aujourd'hui parti intégrante du Platform SDK de Microsoft fourni toutes les APIs, points d'entrée, fichiers headers et librairies nécessaires.

2.4.2# Processus

<Rien de spécial à expliquer>

2.4.3# Pré-traitements

<Rien de spécial à expliquer>

2.4.4# Post-traitements

Les post-traitements sont importants. Un post-traitement, ou simplement traitement, est une opération appliquée sur chaque news entre l'instant où elle est extirpée de la page web et l'instant où elle est archivée, sur disque dans un fichier, dans une base de données, voire dans une nouvelle page web.

On peut appliquer un ou plusieurs traitements, et l'ordre des traitement peut être important.

2.4.4.1# Suppression des effets de police de caractères

Un exemple de traitement ? imaginons que nous décidons que les news sont toujours affichées avec un style de police de caractères normal. Pas de gras ni d'italique ni de souligné. Il faut d'une manière ou d'une autre analyser

chaque news et enlever tous les tags <B> </B> <I> </I> <U> </U> quel que soit l'ordre dans lequel ils se trouvent. Ce faisant, les news obtenues sont parfaitement homogènes visuellement.

Commandes :
REMOVE <B> </B>
REMOVE <I> </I>
REMOVE <U> </U>

2.4.4.2# Suppression des tags IMAGE

Seule la news textuelle est capturée, certainement pas l'image (les images) éventuelle qui serait accolée. Car u tag image est une référence vers un fichier binaire, et rien d'autre. Il convient donc de supprimer ces tags.

Exemple : <img src="logo.gif> News1

Commande :
REMOVE <IMG> </IMG>

2.4.4.3# Mise à niveau des tailles de police de caractères

But de la manipulation : homogénéiser la présentation des news entre elles d'une part, et entre les news de deux sites distincts d'autre part. Ceci est particulièrement dû au fait que la première news d'une série de news est souvent proposée avec une police plus grande, pour attirer l'attention. Des sites de news sont en taille 3, et d'autres en taille 2 : pour mettre tout le monde d'accord la taille est mise à 2 systématiquement.

Exemple : avant traitement <FONT FACE="Arial,…." SIZE="3">….</FONT>
après traitement <FONT FACE="Arial,…." SIZE="2">…..</FONT>

Commande :
FORCE FONTSIZE 2

2.4.4.4# Suppression des attributs CLASS

Un attribut CLASS d'un tag HTML permet de moduler sa présentation visuelle, mais l'attribut CLASS référence un style déclaré soit dans la page HTML, ailleurs, soit dans un fichier .css externe. Dans les deux cas de figure, ces informations sont capturées mais elles sont en trop, il faut donc les supprimer. Il faut les supprimer car il est souhaitable de stocker des news brut de fonderie, sans artifice visuel. Une des applications de la capture de news est justement de fournir en sortie des news ayant un aspect visuel homogène et uniforme, quel que soit le site capturé.

Commande:
REMOVE ATTRIB CLASS

2.4.4.5# Ajout des guillemets sur les attributs des news

Pour une raison que j'ignore, le D.O.M. de IE supprime la plupart des guillemets des attributs des tags HTML. C'est assez mauvais car cela peut engendrer des confusions entre deux attributs qui se suivent, et notamment du fait de la présence d'un caractère espace. Sans les guillemets, seul le caractère espace permet de séparer deux attributs de tags. Si la valeur d'un attribut de tag comporte un espace, il y aura erreur d'interprétation.
Pour prendre les devants, et éviter les surprises, mieux vaut mettre les guillemets que IE a enlevé.

Exemple : avant traitement …….

après-traitement …….

Commande:
ATTRIBUTES QUOTES

A noter, on doit ajouter intelligemment les quotes et choisir en fait entre **apostrophes** et **guillemets** selon si le contenu contient déjà respectivement des guillemets ou des quotes. C'est le principe de l'alternance.

2.4.4.6# Reformulation des liens

Il y a deux types de news sur les portails. Les news purement informatives, non cliquables. Et les news cliquables qui renvoient vers l'article, en remplaçant la homepage du portail par une page dédiée, ou en faisant apparaître une fenêtre séparée prévue à cet effet.

Les news non cliquables n'ont pas besoin de post-traitement.

Mais lorsqu'une news est cliquable, chaque lien est problématique car il renvoie soit :
- à une URL absolue, de la forme http://……. sur le site de news, ou sur un autre site
- à une URL relative partant de la racine, de la forme /news/1.html,
- à une URL relative partant du répertoire courant (celui où se trouve la page des news), de la forme news/1.html

Dans le premier cas de figure, rien n'est à faire puisque l'URL est en quelque sorte opaque et fonctionne que le lien soit cliqué à partir du site de news ou à partir d'ailleurs.

Mais les deux autres cas de figure sont problématiques. Lorsqu'on clique sur les news une fois reformulées dans une nouvelle page web, il ne faut pas oublier qu'elles s'affichent dans un nouveau site web, et donc dans un autre domaine (exemple dans http://www.wired.com/news/1.html, le domaine est wired.com).
Puisque le domaine est différent, les liens exprimés en relatifs doivent nécessairement être transformés en liens absolus. Pour ce faire, il suffit de construire une nouvelle URL par concaténation de **http://**, du domaine, et de l'URL exprimée en relatif.
A un détail près, c'est que les cas de figure 2 et 3 se distinguent par un caractère / initial très important. Il faut le prendre en compte.

Après cette reformulation les liens sont parfaitement cliquables et opérationnelles.

Commande:
REMAP A HREF

Lorsqu'on clique sur un lien, la page des news est remplacée. Pour conserver la page des news, pour une raison de confort ergonomique, il faut faire une reformulation spéciale :

Plutôt que de reformuler /news.1.html en http://www.wired.com/news/1.html
Il faut reformuler /news.1.html en javascript:window.open(http://www.wired.com/news/1.html ,"News").

Ou mieux, indiquer l'attribut **target="_blank"** pour chaque lien, ce qui permet que le lien fonctionne si l'on clique sur le bouton gauche ou sur le bouton droit (option : ouvrir dans nouvelle fenêtre).

Pas très compliqué, mais visuellement cela change tout. Se référer à la documentation MSDN pour connaître les options de la fonction Javascript Window.open, permettant de décider des caractéristiques de la fenêtre qui s'ouvre.

Un vrai problème : on capture une news pointant non pas vers un lien absolu ni relatif, mais un lien de type code Javascript, comme par exemple : **Click here**.

Que faire ? rien dans l'état. C'est une grosse difficulté, car il faut en fait capturer le code de la fonction **go()**. Qui elle-même peut faire appel à d'autres fonctions, ainsi qu'à des éléments HTML. Problème quasi inextricable dans le cas général, sauf en construisant au moins un interpréteur javascript complet. Pas trivial.

Fort heureusement, ce type de lien est assez rare. L'essentiel des sites de news sont "bruts" de fonderie.

2.4.5# API d'accès au tags HTML et à leurs attributs

Les post-traitements sont d'autant plus facile à implémenter que l'on dispose d'une interface de programmation efficace, permettant littéralement de naviguer dans le contenu HTML comme si de rien n'était. Les concepts en jeu sont les tags, les attributs, et les valeurs associées aux attributs.
S'il y a besoin pour une raison ou pour une autre de gérer les relations père-fils entre des tags HTML, il est plus simple d'utiliser le D.O.M., mais l'API du D.O.M. ne fournit pas d'interface pour les tags eux-mêmes.

[IN] : szString : contenu HTML brut de fonderie
[OUT] : szBeginIndex
[OUT] : szEndIndex

BOOL getNextTag(szString, szBeginIndex, szEndIndex);

[IN] : szString : contenu HTML brut de fonderie
[IN] : szBeginIndex, début du tag courant
[IN] : szEndIndex, fin du tag courant
[OUT] : szBeginAttribute, début du nom de l'attribut courant
[OUT] : szEndAttribute, fin du nom de l'attribut courant
[OUT] : szBeginValue, début de la valeur de l'attribut courant
[OUT] : szEndValue, fin de la valeur de l'attribut courant

BOOL getNextAttribute(szString,szBeginIndex,szEndIndex,szBeginAttribute, szEndAttribute, szBeginValue, szEndValue)

A partir de ces deux fonctions il est possible d'extraire et de modifier un tag quelconque et les attributs associés. La valeur de retour BOOL permet de savoir à l'avance si les paramètre de retour [OUT] sont du sens.

2.5 Quelques cas réels

2.5.1# Temps de réponse

Pour capturer 30 sites, il faut 5 minutes environ. Soit un site toutes les dix secondes. Dans un contexte où la bande passante est non pas de 3kb/seconde mais plutôt 20kb/ seconde, ceci serait très prometteur. Quelques chiffres :
à 3kb/ seconde (bande passante d'un particulier, connection Internet par modem analogique)
Pour capturer 100 sites, il faut 16 minutes environ.
Pour capturer 1000 sources d'info, il faut donc 2 heures 40 minutes environ.

à 20kb/ seconde (bande passante professionnelle)
Pour capturer 100 sites, il faut 3 minutes environ.
Pour capturer 1000 sources d'info, il faut 30 minutes environ.

à 60kb/ seconde (ligne spécialisée)
Pour capturer 100 sites, il faut 1 minute environ.

Pour capturer 1000 sources d'info, il faut 10 minutes environ.

2.5.2# Le cas de la news scindée en deux parties

Sur le site de Netslaves, la news est un titre non cliquable. Immédiatement suivi d'un résumé de taille variable et d'un lien cliquable, la news elle est particulière puisque scindée en une partie non cliquable et une partie cliquable. Une idée très simple consiste à ne pas récupérer les liens.
Le plus simple, pour conserver le code actuel est de capturer le bloc news complet, avec comme effet de bord de récupérer un véritable paquet de données plutôt qu'un simple titre cliquable.
La troisième solution consiste à être capable d'indiquer non pas un mais deux chemins d'accès dans le fichier de description. Le premier pour le titre, et le second pour le lien. Ceci double la taille du fichier de description, et le rend nettement plus opaque. Surtout que ce cas de figure est très rare. Ce serait dommage de généraliser la méthode à cause d'un cas particulier. On peut donc imaginer au pire une syntaxe spéciale. La syntaxe est normale, c'est-à-dire simple, pour les titres cliquables, dits à lecture directe, et la syntaxe est complétée par un appendice pour le cas de figure. Reste à définir une syntaxe : on peut tout simplement séparer les deux chemins d'accès par un signe -. Ce qui donne par exemple **news=1;0;0;0;1-1;0;0;0;1;0;0**
Quatrième solution, la meilleure car automatique, consiste à ce que WalkAll soit capable tout seul de récupérer le lien cliquable. On entre de plein pieds dans le domaine du pattern matching, car le cas de figure de Netslaves peut se retrouver ailleurs mais par exemple dans une configuration où le lien cliquable est disposé très différemment sur le plan hiérarchique.

Dans les troisième et quatrième solutions, il faut dire qu'il reste, une fois les données capturées, à les mixer pour en faire une news classique, un titre cliquable. C'est loin d'être trivial, sans modélisation préalable.

Une modélisation consiste à définir un titre **T** et un lien **L**. Le bloc de données est de la forme T….L, où ….. est du code HTML très général. On modélise le bloc à obtenir par **T** .

On voit bien ici qu'une paramétrisation des données capturées et des transformations est une évolution naturelle du produit tel qu'il est aujourd'hui.

D'un simple logiciel de capture de news formatées, WalkAll deviendrait un outil capable d'extirper dynamiquement des masses de données moyennement structurées. Une valeur technique insoupçonnable.

2.5.3# Le problème de la session courante

Sur le site américain *The Economist*, les news pointent vers des articles dont l'URL fait passer des identifiants de session. Dans l'URL relative **displayStory.cfm?Story_id=386104&CFID=258705&CFTOKEN=89513687**, les paramètres **CFID** et **CFTOKEN** sont créés par le serveur ColdFusion lorsque le visiteur arrive sur le site. Le serveur garde en mémoire ces identifiants pour suivre le visiteur, jusqu'à expiration.
Cela implique qu'**éventuellement** ces paramètres sont stockés par des cookies sur le disque dur du visiteur, ou d'une manière ou d'une autre obligent l'utilisateur à passer par la case départ, la homepage, pour obtenir sans le voir ces identifiants et ensuite seulement pouvoir accéder aux articles.
C'est problématique car cela veut dire que WalkAll doit d'abord simuler une connection sur la homepage, récupérer des identifiants, modélisés pour ce site, puis les intégrer dans les URLs des news au moment de la capture.
Le travail de transformation est important et WalkAll ferait beaucoup plus que ce qu'il fait aujourd'hui puisqu'avec un tel mécanisme nous nous donnons le moyen de simuler des clics souris et donc de traverser des sites.

De plus, une fois les données capturées, le problème se pose de nouveau lorsque l'utilisateur clique sur les données capturées pour accéder aux articles. Une fois de plus, la valeur des paramètres CFID et CFTOKEN ne correspond pas aux valeurs obtenues par un utilisateur allant effectivement sur le site, via la homepage. Selon si un mécanisme vérifie la cohérence des données, la porte d'entrée de The Economist peut bel et bien rester fermée.

Dans ce cas aussi, il faut être capable de simuler une connection préalable sur la homepage. Mais ce qui change, c'est que cette action est à faire lorsque l'utilisateur clique sur un titre de la page Web listant toutes les captures. Nous ne sommes plus là dans le contexte de WalkAll.
Pour simuler une connection, il faut par exemple rediriger l'URL vers un serveur d'application, un processus capable de faire un certain travail pour nous, sans qu'il n'y ait rien de spécial à faire ni à installer sur la machine. Ce serveur d'application est un composant complexe qui joue notamment le rôle de proxy web : il surfe sur le site de The Economist pour votre compte tout en gérant pour vous les paramètres de session, cookies, etc.

WalkAll aurait alors une composante "offline", celle d'aujourd'hui, et une composante online, celle à venir.

Le serveur d'application est composant basé sur une librairie HTTP par dessus un serveur web classique. Cette librairie HTTP est un ensemble de fonctions permettant à un code assez simple de charger une page web en transmettant des paramètres type GET/POST/Cookie, traiter la page web si besoin, et renvoyer les données à l'utilisateur comme si de rien n'était, comme si c'était juste un serveur web.

2.5.4# Tags mystérieux

Sur un site, webfaster.net, pour une raison inconnue, les pages web contiennent des tags propriétaires, des tags non HTML. Ce site va à l'encontre des conditions d'utilisation du web. Comme par exemple **wmlinfo**. Ces tags apparaissent dans le code HTML et peuvent causer de graves problèmes aux navigateurs. Ceux-ci peuvent soit refuser d'afficher la page web, générer une ou plusieurs boites de message à l'écran, mal restituer la présentation graphique du reste du code HTML, etc.
Le plus génant c'est que quand bien même le browser HTML passerait à travers sans trop souffrir, le D.O.M. est susceptible d'accéder à ce genre d'éléments, ce qui posera problème : au moment de parcourir la hiérarchie, il est possible qu'une vérification des méthodes de programmation supportées par cet élément soit mise en œuvre, donnant lieu à un "plantage" du D.O.M.. Impossible d'accéder au code HTML fils de ce genre de tags. Impossible donc pour WalkAll de capturer les données.
Il faut soit que le D.O.M. soit excessivement conciliant et supporte tout de même les méthodes, ou alors il faut avoir les moyens de détecter chaque élément, savoir les méthodes qu'il supporte et agir en conséquence.
A expérimenter.
Dans le cas du D.O.M. d'Internet Explorer 4, clairement il y a bug lorsqu'on essaye d'accéder à des méthodes comme **.children**.

Autre cas de figure : vu sur le site JournalDuMail

```
<p>
  <font face="verdana,helvetica,arial" size="-1">
    <b>Vous consultez vos emails....</b><br>
    plusieurs fois par :<br>
     <li>heure 56,4%
     <li>jour 38,3%
     <li>mois 2,5%
     <li>siècle 2,5%<br>
    (plus de 400 votants !)<br>
    <i>(JournalDuMail - 30/10/2000)</i><br>
  </font>
</p>
```

Mal compris par le D.O.M. IE car il manque les tags <ul> et devant entourer les éléments de liste <li>.

2.5.5# Affinage automatique des chemins d'accès réels

Ce paragraphe est déjà une vision pour la prochaine version de WalkAll, une version plus intelligente, et toujours automatique.

Capturer un site de news, c'est capturer une bande verticale de liens structurés habillés dans une présentation HTML.

Oct 2000

Capturer un site de news, c'est donc supposer un déterminisme dans la relation qui fait passer de la news courante à la news suivante. Ce déterminisme est vérifié dans la majorité des sites de news.

Malgré tout, il suffit que, pour prende comme exemple Wired news, certaines news (cela semble aléatoire) soient habillées par un titre supplémentaire les chapeautant, ou une image adjacente, cela détruit la relation triviale de la news courante à la news suivante.

Mais fondamentalement chaque news a une structure simple avec un habillement commun.

Il est donc raisonnable de penser qu'un système tirant partie de la cohérence de la structure d'une news, au fil des news, est susceptible de prendre en main ces décalages de positionnement. Et fournir ainsi une solution de capture indépendante de certaines fioritures.
On peut d'ailleurs se poser la question de savoir jusqu'à quel point sur les sites web ces fioritures sont aléatoires, plutôt que parfaitement prévues par les journalistes pour précisément casser les logiciels de capture. Il faut comprendre qu'un logiciel de capture est susceptible de diminuer substantiellement l'audience d'un site, au moins l'audience sur la homepage. Or il se trouve que cette audience est une des sources majeures de revenu, sur la base de PLUS D'AUDIENCE = PLUS DE VENTE D'ESPACES PUBLICITAIRES.
On peut aisément comprendre que les webmasters voient d'un mauvais œil tout logiciel de "navigation automatique".

Il y a deux cas de figure fréquents :
- le cas d'un surtitre
- le cas d'une image accolée à la news

La bonne nouvelle, c'est qu'ils sont simples à traiter !

2.5.5.1# Le cas du surtitre

Capture provenant du site WiredNews

Ce cas de figure montre clairement qu'une news, aléatoirement, est chapeautée par un titre la catégorisant : WirelessNotebook. Cela se traduit pour la news précédente et cette news par le schéma HTML suivant :

```
<p>
  <font face="Arial,Geneva,sans-serif" size="3" color="#000000">
  <b><a href="/news/business/0,1367,39531,00.html" target="_top">Got Content? Think Syndication</a></b>
  </font>
  <br>
  <font face="Verdana,Arial,Geneva,sans-serif" size="2" color="#000000"> If you think content creation is dead just
because DEN, Pseudo, and Pop.com crashed and burned, think differently. New content creation is on the way -- through
syndication. By Brad King.
  </font>
  <br>
  <font face="Verdana,Geneva,sans-serif" size="1" color="#000000"> <i><a href="/news/business/0,1367,,00.html"
target="_top">in Business</a></i>
  </font>
</p>


<p>
  <font face="Verdana,Geneva,sans-serif" size="1" color="#FF0000"> Wireless Notebook </font>
  <br>
  <font face="Arial,Geneva,sans-serif" size="3" color="#000000">
  <b><a href="/news/technology/0,1282,39572,00.html" target="_top">What's Next: Jewelry Waiting?</a></b>
  </font>
```

```
  <br>
  <font face="Verdana,Arial,Geneva,sans-serif" size="2" color="#000000"> IBM's wacky interpretation of the future phone is
jewelry with a converged cell phone and PDA. Also in Elisa Batista's wireless notebook: Two-way instant messaging ... the
wireless Web is no Internet ... and the Los Angeles Lakers have wireless scouts.
  </font>
  <br>
  <font face="Verdana,Geneva,sans-serif" size="1" color="#000000"> <i><a href="/news/technology/0,1282,,00.html"
target="_top">in Technology</a></i>
  </font>
</p>
```

La différence de structure est indiquée en **rouge**. Chaque news est composée d'un lien, suivi d'un résumé, et d'un lien vers la catégorie à laquelle appartient la news.
Une solution doit tout de suite être écartée. C'est celle qui consiste à tagger non pas le lien vers la news, mais le tag <p>, c'est-à-dire le bloc entier. On ne veut que le lien vers la news.

Ce qui nous intéresse est du contenu habillé par du code HTML selon la structure FONT-B-A, où B est fils de FONT et A fils de B. (Modèle : si deux éléments se suivaient et étaient frères, on utiliserait le signe + plutôt que le signe - qui dénote la relation père-fils).

Dans la news suivante, on trouve FONT+BR+FONT-B-A.

D'autre part, le chemin d'accès calculé pointe sur le premier tag FONT.

Puisque le premier tag FONT n'a pas de fils, et encore moins B puis A, la news ne peut pas être récupérée avec la version actuelle de WalkAll. La capture sur le site s'arrête à ce niveau.

Si on va plus loin, on se rend compte qu'il suffirait de parcourir les frères du premier tag FONT pour tomber sur le bon tag FONT. Pas très compliqué techniquement. Mais effectivement, c'est une méthode qui procède par tatonnement autour d'une position centrale plus qu'un *pattern matching* évolué.

Et cette méthode n'est pas sure, puisqu'on peut très bien tomber sur une mauvaise suite de tags FONT-B-A, même si ce n'est pas le cas ici.

En l'absence d'un pattern matching évolué, cette solution peut être mise en œuvre. Pour la valider, il suffit de traiter un certain nombre d'exemples. En particulier il est intéressant de traiter le cas de l'image accolée.

Deuxième exemple : France.internet.com

```
<font face="Arial, Helvetica, Sans Serif" size="2" color="#CC0000"><b>-- <a href="chaine.asp?chaine_id=11"><font
color="#CC0000">Etudes / Statistiques</font></a> --</b> </font>
</font>
<br>
<font face="Arial, Helvetica, sans-serif" size="2">[28/10]</font>
<font face="Arial, Helvetica, sans-serif" size="2"><a href="news.asp?news_ID=1808"><b>
          Un tiers des entreprises européennes croient à la nouveauté du m-business </a></b><br>
          Selon Arthur Andersen, le m-business, qui peine encore à convaincre les entreprises, devrait  entraîner une
nouvelle croissance dès 2002.
</font>


<font face="Arial, Helvetica, Sans Serif" size="2" color="#CC0000"><b>-- <a href="chaine.asp?chaine_id=9"><font
color="#CC0000">E-commerce</font></a> --</b> </font>
<font face="Arial, Helvetica, Sans Serif" size="2" color="#CC0000"><b>-- <a href="chaine.asp?chaine_id=15"><font
color="#CC0000">Produits / Solutions</font></a> --</b> </font>
</font>
<br>
<font face="Arial, Helvetica, sans-serif" size="2">[28/10]</font>
<font face="Arial, Helvetica, sans-serif" size="2"><a href="news.asp?news_ID=1807"><b>
          Sun et Commerce One s'allient sur le marché de l'e-marketplace </a></b><br>
          Le constructeur Sun et Commerce One, l'éditeur de solutions de places de marchés virtuelles et d'e-procurement,
s'allient dans le but développer des outils d'e-marketplace pour la plate-forme Internet de Sun. Une alliance stratégique
de poids sur un marché hautement concurrentiel.</font>
```

Dans ce cas, il y a un tag supplémentaire dans la news suivante. Un déplacement par tatonnement permet de s'en sortir.
Il faut également noter que ce site comporte plusieurs erreurs HTML d'intensité variable :

- tag tout seul, ici en vert
- tags <a> et entrelacés : <a>…., incompatibles sur le fond avec HTML.

2.5.5.2# Le cas de l'image accolée

Il y a principalement deux cas de figure :

Oct 2000

- l'image est incluse dans le bloc lien de la news
- l'image n'est pas incluse dans le bloc lien de la news

Premier cas, on se retrouve avec un code HTML de la forme :

news courante: (pas d'image)
<font size="2Breaking news4

news suivante: (une vignette sur le côté)
<**img src="vignette1.jpg" align="left"**>Breaking news

et l'image n'est pas cliquable ou même,

<**img src="vignette1.jpg" align="left"**>Breaking news

et l'image est cliquable.

Pas de problème car la capture va embarquer tout ce qu'il y a dans le tag , qu'il y ait un tag <img> ou non. Puisque les post-traitements sont prévus pour supprimer tout tag <img>, la capture de la news fonctionne bien. Au final, la présentation de la news est standardisée et sans image.

Techniquement, le chemin d'accès est valide. On tombe sur un élément, on extraie le contenu, et comme ce contenu n'est pas vide on passe ce dernier dans la moulinette du post-traitement.

Second cas, l'image n'est pas incluse dans le bloc lien de la news :

On peut très bien être confronté à :

<img src="vignette1.jpg" align="left">Breaking news

Traitement : il faut bien comprendre qu'il n'y aurait pas de problème si le tag <font> correspondait au chemin d'accès pour accéder à la news. On se trouve dans le cas de figure où, parce qu'il y a une image alors qu'il n'y en a pas sur la news précédente, l'accès à cette news ne se passe pas bien. On tombe sur le tag <img> et on ne ramène aucune news puisque le tag <font> est frère du tag <img> et non fils.
La solution ? On l'a évoqué précédemment : il faut être capable de parcourir les frères à la recherche d'une suite de tags connue, ici FONT-A

Techniquement, le chemin d'accès est valide. On extrait le contenu, mais on ne récupère rien. Le fait que le contenu extrait soit vide attire l'attention. Alors on essaye par tatonnement de trouver le bon frère. Si le chemin d'accès est 4;5;1;0;0, il faut rendre variable la dernière composante, et essayer, 4;5;1;0;1 puis 4;5;1;0;2, etc. à la recherche de la suite de tags FONT-A.

Puisque la méthode n'est pas sure, mieux vaut ne pas boucler sur 4;5;1;0;X pour X prenant des valeurs de 1 à 10. On risquerait de tomber sur tout à fait autre chose que la news, comme un lien dit d'enrichissement, ou pire un lien de renvoi.

2.5.6# Le fichier d'index pour les 30 sites

site=INA offres d'emploi
newspage=http://www.ina.fr/cgi-bin/INA/Media/New/search.pl?num_type=1&sous_date=6&sortby1=0&sous_lieu1=0&sous_lieu2=0&sous_typo1=0&sous_typo2=0&sous_typo3=0&sous_typo4=0&sous_typo5=0&sous_typo6=0&sous_typo7=0&sortby2=0&sortby=0&langue=fr&mode=emplois
amount=10

news=1;1;0;0;0;0;1;0;1;1;0;0;0
news=1;1;0;0;0;0;1;0;1;1;0;0;1
news=1;1;0;0;0;0;1;0;1;1;0;0;2
site=Wired
newspage=http://www.wired.com/news/nc_index.html
amount=0
news=1;1;9;0;0;4;16;2;0;0
news=1;1;9;0;0;4;17;0;0;0
news=1;1;9;0;0;4;18;0;0;0
site=01net
newspage=http://www.01net.com
amount=0
news=2;1;2;0;0;2;0;0;0;0;0;0;0;2;2;2;0
news=2;1;2;0;0;2;0;0; 0;0;0;0;0;2;2;4;0
news=2;1;2;0;0;2;0;0;0;0;0;0;0;2;2;6;0
site=Transfert
newspage=http://www.transfert.net/fr/index.cfm
amount=8
news=0;1;0;0;1;0;0;0;0;1;1;0;0;1;0;0;0;0;0;0;1;0;1;0;0;0
news=0;1;0;0;1;0;0;0;0;1;1;0;0;1;0;0;0;0;0;0;1;0;2;0;0;0
news=0;1;0;0;1;0;0;0;0;1;1;0;0;1;0;0;0;0;0;0;1;0;3;0;0;0
site=Netslaves
newspage=http://www.disobey.com/netslaves/
amount=8
news=1;1;3;0;0;4;2;0;0;0;0;0
news=1;1;3;0;0;4;5;0;0;0;0;0
news=1;1;3;0;0;4;8;0;0;0;0;0
site=Economist
newspage=http://www.economist.com
amount=8
news=0;1;1;0;0;8;0;0;1;1;1;0
news=0;1;1;0;0;8;0;0;1;1;4;0
news=0;1;1;0;0;8;0;0;1;1;7;0
site=GRC Privacy
newspage=http://grc.com/x/talk.exe?cmd=xover&group=news&utag=
amount=8
news=1;1;0;9;2;3;8;0;0;0;0;0;0
news=1;1;0;9;2;3;8;0;0;0;0;0;1
news=1;1;0;9;2;3;8;0; 0;0;0;0;2
site=Zdnet
newspage=http://www.zdnet.com
amount=8
news=3;1;5;0;0;0;1;7;0;1;2;3;0;1;1;0;0
news=3;1;5;0;0;0;1;7;0;1;2;3;0;2;1;0;0
news=3;1;5;0;0;0;1;7;0;1;2;3;0;3;1;0;0
site=Cyperus
newspage=http://www.cyperus.fr
amount=8
news=1;1;0;0;0;1;0;1;0;1;1;5;0;0;1;0;0;0;3;0;2
news=1;1;0;0;0;1;0;1;0;1;1;5;0;0;1;0;0;0;5;0;2
news=1;1;0;0;0;1;0;1;0;1;1;5;0;0;1;0;0;0;7;0;2
site=Journal du net
newspage=http://www.journaldunet.com
amount=8
news=1;1;0;0;2;1;0;0;0;1;0;0;0;0;27;0;1;0;0
news=1;1;0;0;2;1;0;0; 0;1;0;0;0;0;0;27;0;5;0;0
news=1;1;0;0;2;1;0;0;0;1;0;0;0;0;0;27;0;9;0;0
site=Multimedium
newspage=http://www.mmedium.com/
amount=8
news=0;1;1;0;1;0;2;0;1;2;0;0;0
news=0;1;1;0;1;0;2;0;2;2;0;0;0
news=0;1;1;0;1;0;2;0;3;2;0;0;0
site=Business 2.0
newspage=http://www.business2.com/
amount=8
news=0;1;0;0;1;0;0;0;0;1;1;0;0;2;0;0

news=0;1;0;0;1;0;0;0;0;1;3;0;3;2;0;0;3;0;0;0
news=0;1;0;0;1;0;0;0;0;1;3;0;3;2;0;0;10;0;0;0
site=Clickz
newspage=http://www.clickz.com
amount=8
news=0;1;0;1;0;0;1;4;0;10;1;3;2
news=0;1;0;1;0;0;1;4;0;13;1;3;2
news=0;1;0;1;0;0;1;4;0;16;1;3;2
site=france.internet.com
newspage=http://france.internet.com
amount=8
news=0;1;0;0;0;5;1;1;5;0;0
news=0;1;0;0;0;5;1;6;0;0
news=0;1;0;0;0;5;1;12;0;0
site=Journal du mail
newspage=http://www.journaldumail.com
amount=8
news=1;1;5;0;0;1;3;0;0;4;1;1;0;0
news=1;1;5;0;0;1;3;0;0;4;1;2;0;0
news=1;1;5;0;0;1;3;0;0;4;1;3;0;0
site=Le monde du Web
newspage=http://www.mondeduweb.com
amount=8
news=1;1;0;0;0;1;1;1;2;0
news=1;1;0;0;0;1;1;1;6;0
news=1;1;0;0;0;1;1;1;10;0
site=Maximum PC
newspage=http://www.maximumpc.co.uk
amount=3
news=0;1;3;0;0;0;1;0;0;0;0;3;0;1;0;0;0;0;0;0;0;0;0;0;0;0;1;0;0;0;0;0
news=0;1;3;0;0;0;1;0;0;0;0;3;2;0;0;0;0;0;0;1;0;0;0;0;0;0;1;0;0;0;0;0;0;0
news=0;1;3;0;0;0;1;0;0;0;0;3;4;0;0;0;0;0;1;0;0;0;0;0;0
site=Newsbytes
newspage=http://www.newsbytes.com
amount=8
news=0;1;5;0;1;0;0;0;4;0;1;2;0;0
news=0;1;5;0;1;0;0;0;4;0;1;6;1;0;0
news=0;1;5;0;1;0;0;0;4;0;1;7;1;0;0
site=Nomade Infos multimedia
newspage=http://actu.nomade.fr/html/med/
amount=8
news=1;1;0;13;0;0;1;1;0;0;0;1;0;0;0;0;1;0;0;0
news=1;1;0;13;0;0;1;1;0;0;0;4;0;0;0;0;1;0;0;0
news=1;1;0;13;0;0;1;1;0;0;0;7;0;0;0;0;1;0;0;0
site=RedHerring
newspage=http://www.redherring.com
amount=8
news=0;1;4;0;1;0;2;0;2;2;1;0;0
news=0;1;4;0;1;0;3;1;0;4;0;1;0;0;0
news=0;1;4;0;1;0;3;1;0;4;0;3;0;0;0
site=Streaming Media
newspage=http://www.streamingmedia.net
amount=8
news=0;1;2;0;4;1;0;2;0;1;0;1;0
news=0;1;2;0;4;1;0;4;0;0;0;5;0;0;2;0;0
news=0;1;2;0;4;1;0;4;0;0;0;5;0;2;2;0;0
site=WashingtonPost Technologies
newspage=http://www.washtech.com
amount=8
news=0;1;10;0;0;3;1;0;0;0;4;0;0;6;7;0;0
news=0;1;10;0;0;3;1;0;0;0;8;0;0;6;2;1;0;1;0;0;0
news=0;1;10;0;0;3;1;0;0;0;8;0;0;6;2;1;1;1;0;0;0
site=Winmag
newspage=http://www.winmag.com
amount=8
news=0;1;1;0;4;4;27;0
news=0;1;1;0;4;4;35;0

news=0;1;1;0;4;4;43;0
site=Forbes
newspage=http://www.forbes.com
amount=8
news=1;1;2;0;1;3;0;0;0;0;0
news=1;1;2;0;1;3;6
news=1;1;2;0;1;3;10
site=Business Week
newspage=http://www.businessweek.com
amount=8
news=1;1;8;0;0;0;0;0;0;0;0;0
news=1;1;8;0;0;0;0;0;0;0;0;3
news=1;1;8;0;0;0;0;0;0;0;0;6
site=Dr dobb's journal
newspage=http://www.ddj.com
amount=8
news=0;1;0;0;1;4;0;0;1;0;0;0;0;0;2;0;0;0
news=0;1;0;0;1;4;0;0;1;0;0;0;0;0;2;0;1;0
news=0;1;0;0;1;4; 0;0;1;0;0;0;0;0;2;0;2;0
site=Net imperative
newspage=http://www.netimperative.com
amount=8
news=0;1;2;0;0;0;1;0;0;0;1;0;0;0;0;0;0;0;0;0;1;0;0;0;0;0;0;2;0;0;0;0
news=0;1;2;0;0;0;1;0;0;0;1;0;0;0;0;0;0;0;0;0;1;0;0;0;0;0;0;4;0;0;0;0
news=0;1;2;0;0;0;1;0;0;0;1;0;0;0;0;0;0;0;0;0;1;0;0;0;0;0;0;6;0;0;0;0
site=Security advisories
newspage=http://www.cert.org/advisories
amount=8
news=0;1;6;0;0;0;5;0;0;1;0
news=0;1;6;0;0;0;5;0;0;4;0
news=0;1;6;0;0;0;5;0;0;7;0
site=Journal du Wap
newspage=http://www.journalduwap.com
amount=8
news=0;1;0;0;0;0;6;1;1;0;2
news=0;1;0;0;0;0;6;1;1;0;5
news=0;1;0;0;0;0;6;1;1;0;8
site=MSDN Online
newspage=http://msdn.microsoft.com/default.asp
amount=8
news=1;1;0;15;0;0;1;0;0;0;0;0;2;0;0;0;0
news=1;1;0;15;0;0;1;10;0;1;2;0;0;0;0
news=1;1;0;15;0;0;1;10;0;2;2;0;0;0;0
site=Nomade Info
newspage=http://actu.nomade.fr/html/une/index.asp
amount=8
news=1;1;0;13;0;0;1;1;0;0;0;1;0;0;0;0;1;0;0;0
news=1;1;0;13;0;0;1;1;0;0;0;4;0;0;0;0;1;0;0;0
news=1;1;0;13;0;0;1;1;0;0;0;7;0;0;0;0;1;0;0;0
site=Vakooler
newspage=http://www.vakooler.com
amount=8
news=0;1;3;0;0;0;0;0;0;1;0;0;0;0;0;0;0;0;0;0
news=0;1;3;0;0;0;0;0;0;1;2;0;0;0;0;0;0;0;0;0
news=0;1;3;0;0;0;0;0;0;1;4;0;0;0;0;0;0;0;0;0
site=Vnunet
newspage=http://www.vnunet.fr
amount=8
news=0;1;6;0;0;1;1;0;0;0;0;0;0;0;1;0
news=0;1;6;0;0;1;1;0;0;0;0;0;0;0;5;0;0
news=0;1;6;0;0;1;1;0;0;0;0;0;0;0;5;1;0

Oct 2000

3# WalkAll version 2.0

WalkAll est un outil parfaitement opérationnel. Il sait capturer des news provenant d'un certain nombre de sites, les aggréger et en faire un point d'entrée unique, une page de news ! (une page web).

Du fait de l'instabilité notable des portails de news, il est nécessaire d'effectuer une vérification régulière du fichier de description. Pour chaque site, le fait de ne récupérer aucune news alors qu'il n'y a pas d'erreur de connection, met la puce à l'oreille. D'une certaine manière, WalkAll ne pourra jamais se passer d'un fichier de description. WalkAll ne sera jamais réellement automatique. Si tel était le souhait il faudrait mettre une intelligence cérébrale dans ce logiciel. Irréaliste tant que les portails sont faits de code HTML fourre-tout et non d'un assemblage de composants identifiables.
De plus, la versatilité d'HTML et la créativité des designers font qu'il y autant de blocs de news différents en structure qu'il y a de sites de news sur le web !

Certaines améliorations sont à une meilleure portée. La première, la plus visuelle porte sur l'ajout d'une catégorie aux news aggrégées ainsi que sur l'accès direct à un bloc de news particulier, éventuellement noyé dans la masse.

Le second type d'améliorations porte sur les aléas de la connection réseau. Ne pourrais-t-on pas envisager une stratégie de capture par reprise ou essais retardés ?

Le troisième type d'améliorations porte sur l'évolution globale du système de capture de news. Il reprend les concepts détaillés dans les paragraphes précédents.

Enfin, WalkAll reste ouvert. Une présence sur le web pourrait susciter le besoin d'un public de capturer les news d'un site donné. Une sorte de **My WalkAll**.

Bref, WalkAll version 2.0 s'enrichit, devient plus intelligent et plus exploitable. Et ce n'est qu'un premier pas, le paragraphe qui vient juste après, décrit une déclinaison server-side de WalkAll. Une déclinaison qui à la fois permet la mise en jour des news en temps réel, autorise l'organisation des blocs selon désiderata et enfin le partage et l'envoi de ces news à des tiers. WalkAll serait alors un outil de travail et un portail de portails.

3.1# Améliorations sur la présentation finale des news

Les news sont en effet de plusieurs types (sans prétendre à l'exhaustivité) :
- news de la net économie (française, européenne, américaine)
- news financières
- actualités françaises
- actualités mondiales
- news de développeurs
- news du monde de la santé
- offres d'emploi

Pourquoi ne pas regrouper les blocs de news appartenant manifestement à une même catégorie ? D'où l'idée qui consiste dans un premier temps, sans rien changer au code de WalkAll, à simplement organiser le fichier de description naturellement par catégorie. Sans rien changer au code, cela veut dire que la catégorie n'apparaît autrement que par une succession de blocs de news qui, apparemment, traitent d'un même thème.
Ceci se fait en quelques minutes.

Dans un second temps, rien n'empêche d'adjoindre un nouveau champ descriptif à l'ensemble des champs d'un bloc news. Exemple : **category=offres d'emploi**.

Ce nouveau champ contribue à faire évoluer la génération de la page web finale afin de bien montrer la déclinaison des catégories.

Plusieurs présentations naturelles peuvent être mises en jeu :
- une présentation par bloc d'une certaine catégorie : une décoration HTML spécifique entoure l'ensemble des blocs de news de la catégorie courante.
- une présentation dépliable : grâce à du code HTML dynamique, il est assez aisé de rendre dépliable et rétractable un bloc de news. Ceci permet de mettre en œuvre une hiérarchie à un niveau.
- une présentation en plusieurs pages : la homepage listerait alors les catégories et servirait de point d'entrée
- une présentation listant les catégories en haut de la page web, de façon à rendre plus efficace et plus directe la navigation dans les blocs de news.

De plus, chaque bloc de news lui-même peut être modélisé. Il suffit pour cela de considérer un code HTML HEADER, *n* codes HTML CONTENT ainsi qu'un code HTML FOOTER. Les codes HEADER et FOOTER fournissent toute la logique de présentation en amorce et en fin de bloc de news, alors que le code HTML CONTENT contient en particulier un mot-clef CONTENT remplacé à la volée par chaque news capturée. En somme, la présentation des news d'un bloc est paramétrable et personnalisable.

3.2# Evolution de la modélisation des blocs de news

En terme d'ouverture, pour assurer l'évolutivité de la modélisation d'un bloc de news quelconque, il est possible soit :
- le plus simple (et le moins évolutif) est de rajouter un champ de description pour chaque bloc de news. Exemple : **category = offres d'emploi**
- de fournir un champ de description, unique dans le fichier de description, décrivant la constitution élémentaire de chaque bloc de news
- de fournir un champ de description par catégorie de bloc de news : ce champ revient à décrire complètement un bloc de news. Exemple : **newsbloc = category newspage amount news***, et où les champs category, newspage, amount et news sont décrits à leur tour. On voit bien que cette description n'est autre qu'un modèle de langage type XML.

Par ailleurs, la méthode de capture peut elle-même être modélisée à l'aide d'une description type XML. C'est-à-dire qu'au lieu de considérer 3 chemins d'accès de news, dont la soustraction logique des deux dernières fournit une méthode incrémentale pour accéder aux autres news du bloc en cours de capture, on peut modéliser à la fois l'accès à une news et l'accès aux autres news du bloc par une loi logique.
Dans WalkAll 1.0, tout se passe comme si la modélisation suivante était mise en œuvre :
FIRSTNEWS_ACCESS = PATH
OTHERNEWS_ACCESS = PATH_ESTIMATION

où PATH et PATH_ESTIMATION sont deux entités connues nativement par WalkAll.

3.3# Améliorations sur la prise en charge des erreurs réseau

Par *erreurs réseau* on entend tous les types d'erreurs possibles ou imaginables empêchant de recevoir la page web comportant des news à capturer.

L'erreur réseau la plus naturelle est l'absence de réseau, due à une coupure de la bande passante entre la machine où WalkAll s'exécute et Internet. L'absence de réseau, tant qu'elle ne dépasse pas un délai fixé à l'avance, est un simple retard. Une fois le délai écoulé, de l'ordre de 3 minutes, la couche applicative rend la main en affichant un message d'erreur. Mais ce message d'erreur est modal, ce qui est génant car il requière l'action physique de l'utilisateur pour passer à la suite, en l'occurrence aborter la capture sur le site courant, et passer au site suivant.

Une erreur réseau fréquente également est l'incapacité d'accéder à un site web donné, soit parce que l'identifiant indiqué dans le fichier de description est mauvais, soit parce que le fournisseur d'accès à Internet à quelques

problèmes avec la résolution DNS, soit parce que le serveur web est "tombé", soit encore parce qu'une page web indiquée n'existe plus. Certaines de ces erreurs sont remontées par l'intermédiaire d'un message modal. Là encore, l'utilisateur est mis à contribution pour passer à la suite.

Une erreur réseau au sens large est constituée par une entête spécifique qui empêcherait d'accéder à la page web. Le problème peut se poser en amont ou en aval. En amont, cela veut dire que la page web peut forcer l'envoi d'u cookie ou d'un identifiant de session quelconque, une preuve qu'un utilisateur physique accède au site. C'est une limitation actuelle de WalkAll, cf paragraphe suivant. En aval, le serveur web peut de lui-même commander le navigateur web à exécuter certaines actions. Ou, pire, cette action peut être indiquée dans le code HTML de la page web, une fois téléchargée, par l'intermédiaire de tags META http-equiv. En particulier la commande <META http-equiv="Pragma" content="no-cache"> indique au navigateur web de vider son cache, ce qui du coup vide le cache de la page web qui vient d'être téléchargée. De ce fait, la page web n'est exploitable qu'une fois sur deux accès, lorsqu'on considère qu'elle est déjà en cache, ou pas en cache.
Pour contrer cette limitation, il est nécessaire de pouvoir indiquer à InternetExplorer, l'outil donnant l'accès au D.O.M., de ne pas utiliser son cache.

Dans tous les cas, lorsque la page web d'un site n'est pas accessible, on doit l'indiquer dans un fichier journal, et de façon complémentaire soit réitérer l'essai, soit passer en revue tous les sites du fichier de description, puis reprendre tous les sites ayant posé un problème, jusqu'à 5 essais, seuil à partir duquel un site est jugé purement et simplement inaccessible. Ceci met en œuvre deux notions complémentaires : le système de reprise, et les essais retardés. Une simple comptabilisation des erreurs en pratique.

3.4# Améliorations sur l'accès intelligent aux news

Qu'entend-on par accès intelligent à l'information ? Par abus de langage, on entend par accès intelligent un accès plus souple qu'un accès parfaitement rigide et prévu à l'avance. Par exemple, un accès intelligent aux news peut simplement se constituer d'un système d'accès par tatonnements. Ces tatonnements sont une mise en œuvre d'un apprentissage. Bien sûr, cet apprentissage est réel ou non, c'est-à-dire est persistent ou non suivant les sites et au jour le jour. Ce qui induit de nombreuses possibilités.

Dans les paragraphes précédents 2.5#, l'intérêt immédiat de l'accès par tatonnements a été évoqué. Il permet d'accéder aux news pour lesquelles des titres intersticiels ont été rajouté, cassant le cas échéant la logique de chemin d'accès.

C'est une solution qui a aussi ses limites, en particulier lorsqu'un bloc de news est entrecoupé par des sections publicitaires par exemple. Auquel cas, une modélisation plus explicite doit être mise en œuvre, afin de séparer le bon grain de l'ivraie.
La modélisation par *pattern matching* consiste à apprendre (ou comprendre) au fil des news capturées la constitution d'une news, de constituer un motif en chainant les types de tags HTML par exemple, puis de chercher ce motif dans le bloc de news. Ce motif peut être sauvegardé pour être réutilisé le lendemain. Un exemple de motif : FONT-B-A dans le cas d'une news pour laquelle le lien cliquable est en gras et spécifie une police de caractères particulière.
Chercher ce motif ailleurs dans le bloc de news ne garantie aucunement que l'on tombe systématiquement sur une news, et non sur un lien purement et simplement avec un FONT et un B. En probabilités pourtant, le risque d'erreur est faible car on ne cherche pas n'importe où dans le bloc de news, on utilise le modèle incrémental de chemin d'accès. Dans le pire des cas, il faut pouvoir spécifier un motif plus complet. Tout est question de modélisation. Si la modélisation est externalisée dans un fichier ou une base de données, il devient possible d'améliorer la recherche à l'aide d'un outil intéractif.

3.5# Des blocs de news aux comptes privés

Oct 2000

De très nombreux sites Internet offrent des services privés, comme une messagerie web-based sur Hotmail. On constate qu'une fois que l'utilisateur a entré son nom d'utilisateur et son mot de passe, il accède à une page web privée constituée de zones tabulaires qui ressemblent fortement sur le principe à un bloc de news.
Bref, les comptes emails, les comptes bancaires, et de nombreuses autres choses, peuvent être capturés sur le même principe. Ceci étend très largement l'interêt de WalkAll, qui devient dès lors non seulement un outil d'aggrégation automatique, voie par laquelle il fait gagner du temps aux internautes, et leur évite de se "logguer" individuellement sur chaque site à contenu, et chercher soi-même l'information.
Mais on a bien remarqué que l'accès a un compte email n'est pas trivial car il passe par un formulaire, une bannière de login, qui elle-même peut être issue de l'utilisation de cookies ou d'autres identifiants de session. Pour accéder automatiquement au compte email sans plus jamais avoir à se logguer, il faut être capable de traverser le formulaire de login, c'est-à-dire de simuler parfaitement l'envoi du login et du mot de passe.
Et comme des cookies ou identifiants de session peuvent être mis en jeu, il faut aussi être capable de transmettre et gérer en amont et en aval ces objets.
Ceci est traité dans un autre outil, Webh4ck. Webh4ck repose lui-même sur une modélisation de cookies, et de paramètres get et post. Pour trouver ces identifiants, il faut utiliser un outil capable de remonter les données invisibles produites par le navigateur web et le serveur web lors de la navigation. Ceci se fait avec un outil déterministe comme Webtracker, ou tout analyseur de paquet TCPIP.

En combinant Webh4ck et WalkAll, il devient possible d'aggréger dans la page web à la fois des données publiques, les news, et des données privées.
L'intérêt ici est de ne plus avoir de quelque manière que ce soit à naviguer soi-même sur les sites, et y passer un temps fou, proportionnel au nombre de sites à visiter, et proportionnel aux problèmes de bande passante et d'erreurs impromptues.

3.5# Enrichissement externe des sites capturés

Supposons que WalkAll soit présent sur le web d'une manière ou d'une autre. Un simple formulaire ADD THIS SITE renverrait l'adresse d'un nouveau site web, particulièrement pertinent pour quelqu'un, à l'administrateur WalkAll (!) et lui permettrait de le "tagger", et de l'insérer dans la liste des news à capturer.

Modèle de business : un client ouvre un compte WalkAll et dispose gratuitement d'un certain nombre de sources capturées par défaut, ainsi que des sources personnalisées, payées par lot de 5.

Le compte est administré via un serveur web.

4# WalkAll sous forme de serveur d'application

Une capture d'écran vaut mieux qu'un long discours :

Oct 2000

Lorsqu'un composant est alimenté par un contenu mis à jour automatiquement quotidiennement, et que ce composant a la bonne idée de pouvoir s'agglomérer à d'autres composants, être archivé sur disque ou partagé sur le net à la façon des post-it, on obtient quelque chose d'unique, ludique et fonctionnel.