UNIVERSITE DE YAOUNDE I
Faculté des Sciences
Département d'Informatique
BP 812 Yaounde-Cameroun

UNIVERSITY OF YAOUNDE I
Faculty of Sciences
Department of Computer Science
P.O Box 812 Yaoundé-Cameroon

Essai d'adaptation au Web Scraping de l'algorithme de Zhang et Shasha's d'appariement des arbres de recherche

Mémoire présente et soutenu en vue d'obtention du diplôme de Master en Informatique par :

NOUTEGNE TCHEUP Lionel Sedar
Matricule : **15U2887**

Sous la direction du **Professeur FOUDA NDJODO Marcel**

Yaoundé, Juillet 2018

Sommaire

Context

- Les sites web ne cessent de croitre de nos jours ;
- Comprendre et exploiter les données du web ;

Remarque

Données disponibles sous forme de **page web** sans aucune option de **récupération.**

Définition

Le **Web Scraping** est l'ensemble des techniques **automatiques** de **récupération** et de **structuration** d'informations des pages Web.

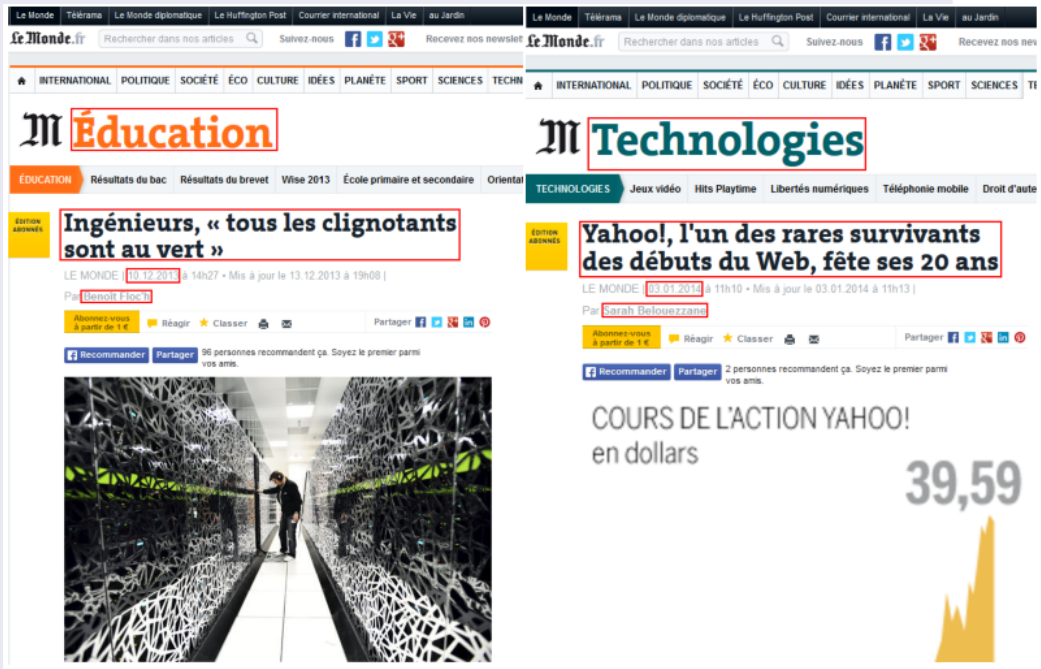


Figure i Page Web du site ***Le Monde.fr***

De nombreux champs d'application

Figure : 1) Veille concurrentielle

Figure : 3) Recrutement digitales

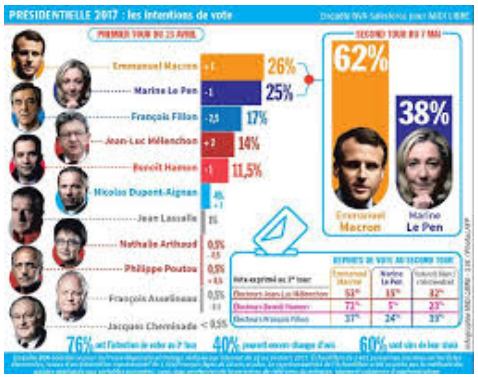


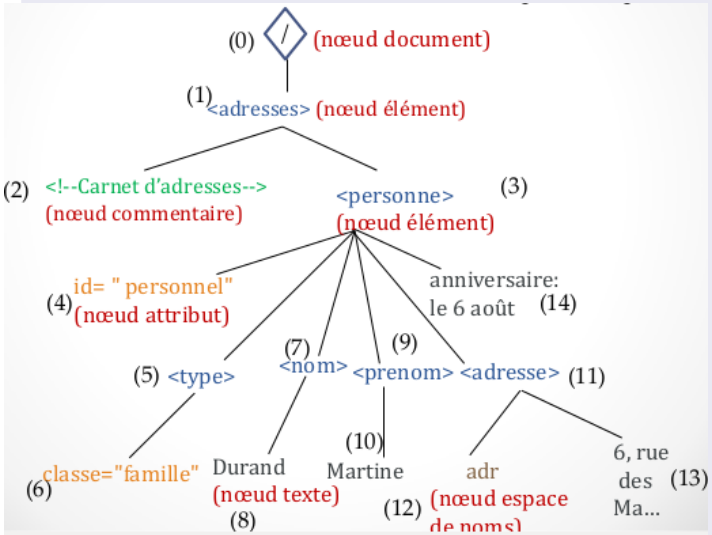
Figure : 2) Sondage électorales

Figure : 4) Retargeting

Problématique

Les méthodes axées XPATH

```
<adresses>
    <!--Carnet d'adresses-->
    <personne id= " personnel " >
        <type classe= " famille " />
        <nom>Durand</nom>
        <prenom>Martine</prenom>
        <adr:adresse>6, rue des Magnolias
        </adr:adresse>
        anniversaire: le 6août
    </personne>
</adresses>
```



Caractérisation de l'approche Xpath

- Localise des composants dans un document XML ou HTML ;
- S'applique sur **XDM** ou **DOM** ;
- Intègre les bibliothèques **lxml** ou **urllib** .
- Utiisé dans **RapidMineur**, **WebScraper**, **Selenium,Scrapy,BeautifulSoup.**

Difficultés

Écriture très complexe et s'applique sur la structure globale de la page ;
Des changements même mineurs demande de nouvelles expressions.

Problématique

- HTML,permet de structurer les données sur les pages Web ;
- La mise en forme des pages web constitue une des difficultés du web scraping ;

Question

Comment renseigner les changements structurelle au sein d'une page web sans toutefois vouloir réécrire la règle d'extraction dans un contexte du web scraping ?

Objectif

1. Comprendre et expérimenter un processus d'extraction pour comprendre une situation de tir au but en football ;
2. Expliquer l'apport de l'appariement des modèles arbres sur des données disponibles sous forme de page web dans un contexte du web Scraping.

Architecture fonctionnelle d'un Scraper

Elle repose sur deux composants

- **Response Manager :** englobe les points (1),(2),(3) et (4) ;
- L'analyseur **Response Parser :** intègre les points (5) et (6).

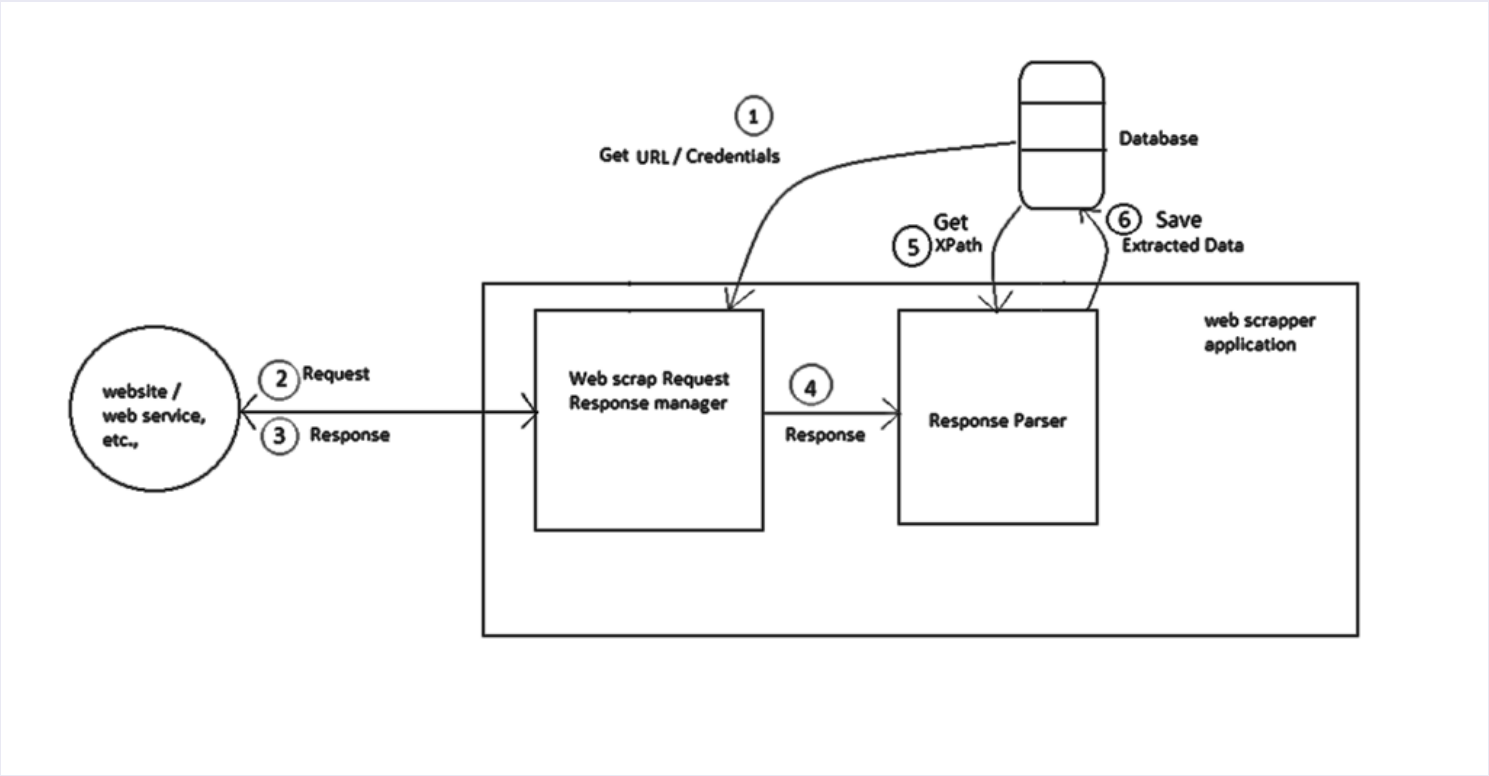


Figure : 5) Architecture logique d'un Scraper [1]

Difficultés du Web Scraping

Lutter contre le Scraping indesirable

- Elles sont **techniques** et même **déontologiques** en raison de nombreuses **affaires juridiques** ;
- De nombreux litiges juridiques comme ;
  - L'affaire opposant American Airlines à Farechase ;
  - Facebook et d'un **Scraper** anonyme ;
  - L'affaire Cambridge Analytica de mars 2018 ;
- Des procéder techniques du site cible du "**Scraper**" ;
  - Se sont des barrières **CAPTCHA** ;
  - La technique de **bannissement d'adresse IP** ;
- Des procédés déontologiques sont généralement des règles ;
  - Récemment la **RGPD** (Union européenne) du 25 mai 2018 ;
  - La loi Camerounaise n° 2010/012 du 21 décembre 2010.

Web Scraping au service des tirs au but

Description du cas d'étude

Une façon de départager deux équipes dans un match de football

- Est d'appliquer les tirs au but ;
- L'on reproche aux tirs au but de s'apparenter à une loterie ;
- Apparement les équipes débutantent remportent la séance ;
- Un échantillon pour vérifier cette probabilité été étudié ;

Outil de Scraping

Le Web Scraper de Chrome est :

- L'oeuvre de **Martins Balodis** ;
- On accède à via l'outil de développement de Chrome ;
- On y trouve des **menus** tels que le menu **Create new Sitemap**, le **Selector graph,** le menu **Scrape,** le **Browser**, menu **Export** et bien d'autres.

Web Scraping au service des tirs au but

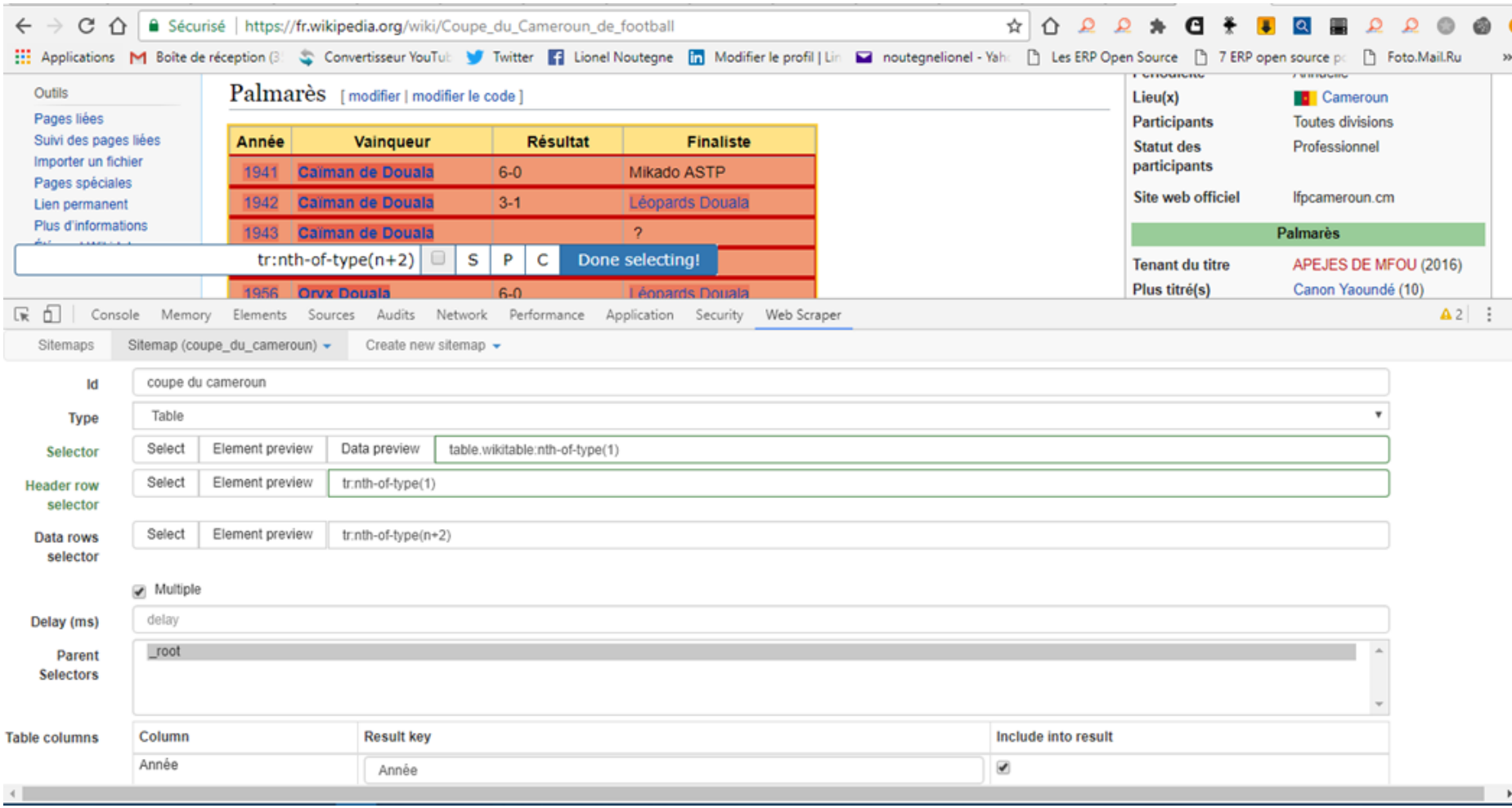
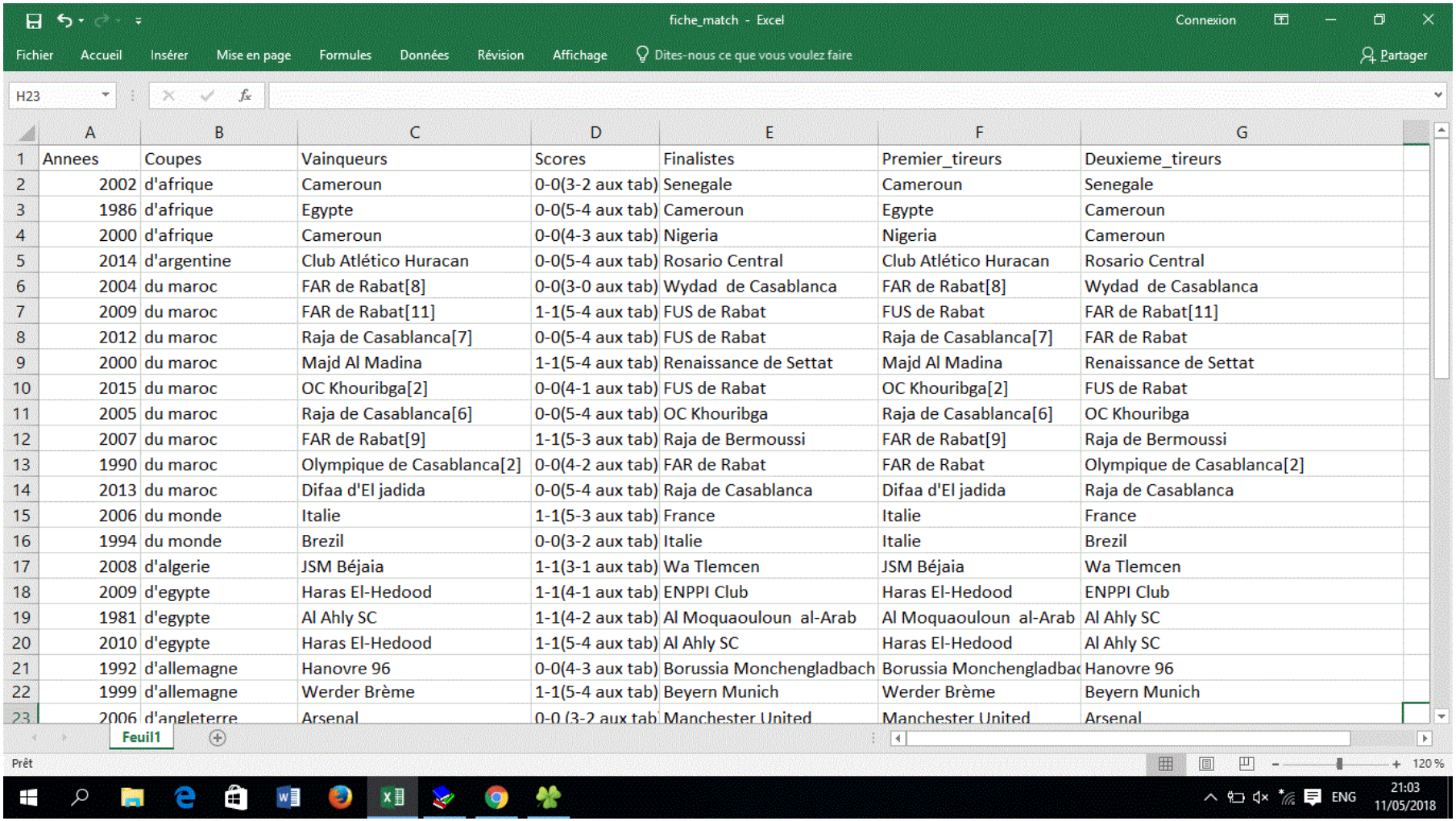


Figure : 6) Capture d'écran montrant l'interface du web scraper en action

Web Scraping au service des tirs au but



fiche_match - Excel

	A	B	C	D	E	F	G
1	Annees	Coupes	Vainqueurs	Scores	Finalistes	Premier_tireurs	Deuxieme_tireurs
2	2002	d'afrique	Cameroun	0-0(3-2 aux tab)	Senegale	Cameroun	Senegale
3	1986	d'afrique	Egypte	0-0(5-4 aux tab)	Cameroun	Egypte	Cameroun
4	2000	d'afrique	Cameroun	0-0(4-3 aux tab)	Nigeria	Nigeria	Cameroun
5	2014	d'argentine	Club Atlético Huracan	0-0(5-4 aux tab)	Rosario Central	Club Atlético Huracan	Rosario Central
6	2004	du maroc	FAR de Rabat[8]	0-0(3-0 aux tab)	Wydad de Casablanca	FAR de Rabat[8]	Wydad de Casablanca
7	2009	du maroc	FAR de Rabat[11]	1-1(5-4 aux tab)	FUS de Rabat	FUS de Rabat	FAR de Rabat[11]
8	2012	du maroc	Raja de Casablanca[7]	0-0(5-4 aux tab)	FUS de Rabat	Raja de Casablanca[7]	FAR de Rabat
9	2000	du maroc	Majd Al Madina	1-1(5-4 aux tab)	Renaissance de Settat	Majd Al Madina	Renaissance de Settat
10	2015	du maroc	OC Khouribga[2]	0-0(4-1 aux tab)	FUS de Rabat	OC Khouribga[2]	FUS de Rabat
11	2005	du maroc	Raja de Casablanca[6]	0-0(5-4 aux tab)	OC Khouribga	Raja de Casablanca[6]	OC Khouribga
12	2007	du maroc	FAR de Rabat[9]	1-1(5-3 aux tab)	Raja de Bermoussi	FAR de Rabat[9]	Raja de Bermoussi
13	1990	du maroc	Olympique de Casablanca[2]	0-0(4-2 aux tab)	FAR de Rabat	FAR de Rabat	Olympique de Casablanca[2]
14	2013	du maroc	Difaa d'El jadida	0-0(5-4 aux tab)	Raja de Casablanca	Difaa d'El jadida	Raja de Casablanca
15	2006	du monde	Italie	1-1(5-3 aux tab)	France	Italie	France
16	1994	du monde	Brezil	0-0(3-2 aux tab)	Italie	Italie	Brezil
17	2008	d'algerie	JSM Béjaia	1-1(3-1 aux tab)	Wa Tlemcen	JSM Béjaia	Wa Tlemcen
18	2009	d'egypte	Haras El-Hedood	1-1(4-1 aux tab)	ENPPI Club	Haras El-Hedood	ENPPI Club
19	1981	d'egypte	Al Ahly SC	1-1(4-2 aux tab)	Al Moquaouloun al-Arab	Al Moquaouloun al-Arab	Al Ahly SC
20	2010	d'egypte	Haras El-Hedood	1-1(5-4 aux tab)	Al Ahly SC	Haras El-Hedood	Al Ahly SC
21	1992	d'allemagne	Hanovre 96	0-0(4-3 aux tab)	Borussia Monchengladbach	Borussia Monchengladbac	Hanovre 96
22	1999	d'allemagne	Werder Brème	1-1(5-4 aux tab)	Beyern Munich	Werder Brème	Beyern Munich
23	2006	d'angleterre	Arsenal	0-0 (3-2 aux tab	Manchester United	Manchester United	Arsenal

Figure : 7) Jeu de données finale destiné à l'analyse

Web Scraping au service des tirs au but

Analyse des données du fichier finale dans Qlikview

- Comprendre le phénomène des tirs au but dans **Qlikview** ;
- **Qlikview** est un outils d'élaboration de tableaux de bord ;
- Le premier tableau de bord montre les pourcentages des finalistes ;
- Le deuxième tableau de bord donne les pourcentages des vainqueurs ;

Intéressons nous au pourcentage des vainqueurs.

Les fonctions associées au deuxième tableau de bord sont :

=round((sum(IF(premier_tireurs=vainqueurs,1,0))/count(TOTAL vainqueurs))*100,0.01)&'%'
=round((sum(IF(deuxieme_tireurs=vainqueurs,1,0))/count(TOTAL vainqueurs))*100,0.01)&' %'

Web Scraping au service des tirs au but

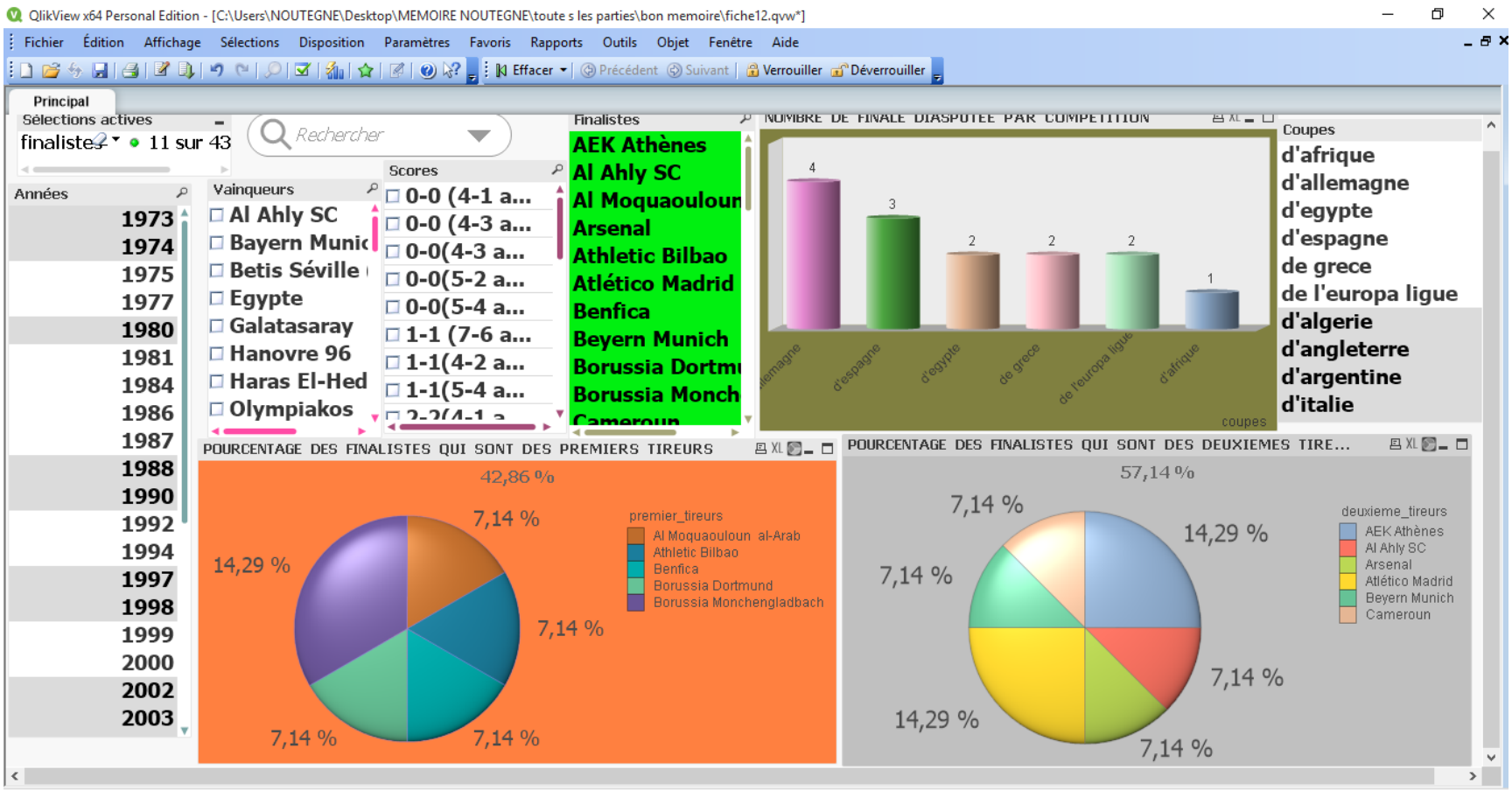


Figure : 8) Tableau de bord sur la classification du nombre de finales disputées par type de compétitions pour un type de finalistes et leurs pourcentages respectifs qu'ils soient premiers tireurs ou deuxièmes tireurs

Web Scraping au service des tirs au but

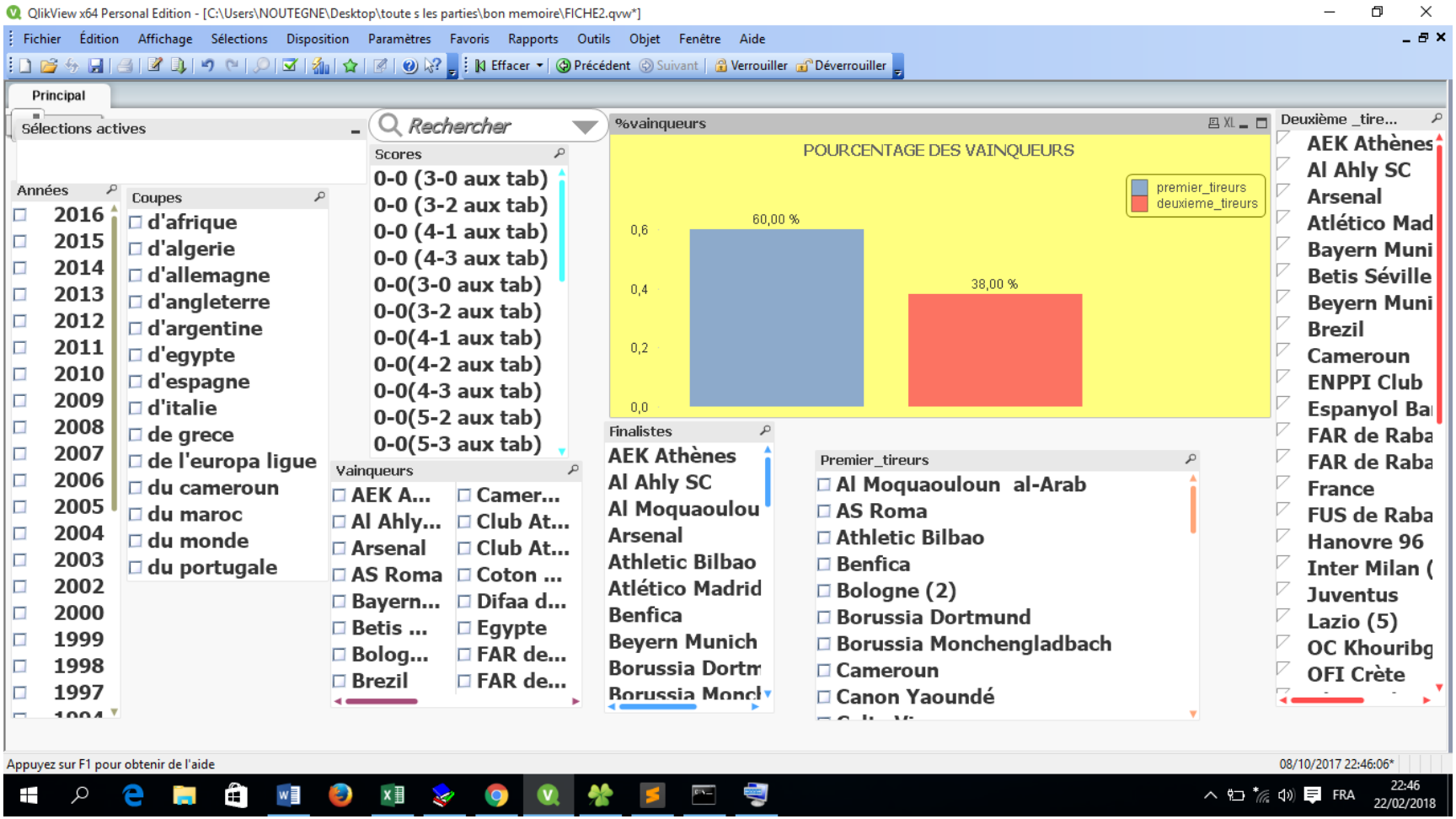


Figure : 9) Capture qui donne les pourcentages des vainqueurs

Web Scraping au service des tirs au but

Au regard de nos résultats

- Les psychologues du sport expliquent ;
- Ces différences par le stress ;
- Habituellement les équipes suivent le format « **AB**-**AB** » ;
- Le nouveau format « **AB**-**BA** » veut que :
  - L'équipe **A** débute la séance avant que ;
  - L'équipe **B** enchaîne deux tentatives et ainsi de suite ;
- **La séance des tirs n'est pas un hasard** ;
- Tenir compte de **l'ordre de passage**.

Concepts qui font les bases de algorithme Zhang et Shasha's

Notions de bases

- **Arbre T** : ensemble de nœuds et d'arcs reliés entre eux dans lequel Il existe un nœud particulier appelé racine ;
- **Forêt F** : ensemble finie ordonné d'arbres ;
- **Keyroots**(T) = **{root**(T)} U v $\in$**V**(T) $\big|$ v a un frère gauche**}** ;
- **lml**(T) : ensemble de tous les descendants gauche de T.

Théorème

Opérations d'édition

- *Etant donnees deux arbres T_1, T_2 et les couples (a, b) (λ,λ) ;*
- *On noté a→b, où a est soit λ ou une étiquette d'un nœud de T_1 et b est soit λ ou une étiquette d'un nœud de T_2 definissant :*
  - *une insertion si a = λ.*
  - *une suppression si b = λ.*
  - *une substitution si a≠λ et b≠λ.*

Distance d'édition entre arbres

- Distances de Levenshtein,1966 et étandue aux arbres par Tai,1979 ;
- Séquence d'opérations correspondantes à un **mapping ;**

Définition

La distance d'édition entre T_1 et T_2 noté treedist $(T_1 ; T_2)$ = **min** {**C(M) M mapping** de T_1 vers T_2} et repose sur le **Lemme** suivant :

Lemme

Soit F_1 et F_2 deux forêts ordonnées et **c** *la fonction coût ;*

Soit v et w les racines des arbres dans F_1 et F_2 On a :

1. $c(\theta,\theta) = 0$
2. $c(F_1,\theta) = c(F_1-v,\theta) + c_0(v \rightarrow \lambda)$
3. $c(\theta,F_2) = c(\theta,F_2-w) + c_0(\lambda \rightarrow w)$
4. $$c(F_1,F_2) = min \begin{cases} c(F_1-v,F_2) + c_0(v \rightarrow \lambda) \\ c(F_1,F_2-w) + c_0(\lambda \rightarrow w) \\ c(F_1[v],F_2[w]) + c(F_1-T_1[v],F_2-T_2[w]) + c_0(v \rightarrow w) \end{cases}$$

Mapping entre arbres

Mapping M

Couples de nœuds (i,j) dans $V(T_1)*V(T_2)$, pemettant de passer T_1 à T_2 comme suit :

1. $1 \leq i \leq |T_1|, 1 \leq j \leq |T_2|$
2. Pour toutes paires (i_1, j_1) et (i_2, j_2) dans **M** :
   1. $i_1 = i_2$ si et seulement si $j_1 = j_2$.
   2. i_1 est à gauche de i_2 si et seulement si j_1 est est à gauche de j_2.
   3. i_1 est un ancêtre de i_2 si et seulement si j_1 est un ancêtre de j_2.

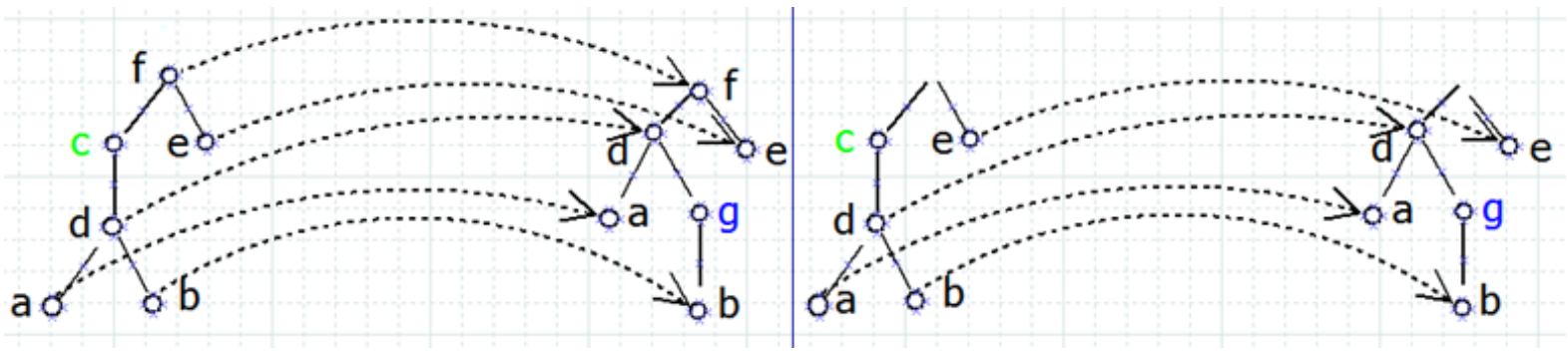


Figure : 10) **(à gauche)** mapping entre arbres et **(à droite)** mapping entre forêts

Algorithme de Zhang et Shasha's

- Améliorer l'algorithme de Tai ;
- S'interesse au calcul des **keyroots ;**

Algorthme

Begin
procédures :(lmd$_1$(),lmd$_2$(),keyroot(T_1)[], keyroot(T_2)[]).
for s : = 1 to $|$keyroot(T_1)$|$ do
for t : = 1 to $|$keyroot(T_2)$|$ do
i = keyroot(T_1)[s] ;
j = keyroot(T_2)[t] ;
treedist(i,j) ;
End

- On a la complexité :
  - $O(|T_1||T_2|)$ **en espace ;**
  - $O(|T_1||T2|minprofondeur(T_1),feuille(T_1)minprofondeur(T_2),feuille(T_2))$ **en temps.**

Appariement approximatif d'arbres

Procédure

- Dénote par **D** l'arbre des données et **P** l'arbre du patron ;
- Introduit l'opération de coupure comme montre la **Figure** : 11) ;
- L'insruction $c(F_1,\theta) = c(F_1 - v,\theta) + c_0(v \rightarrow \lambda)$ du **lemme1** est remplacé par $c(F_1,\theta) = 0$;

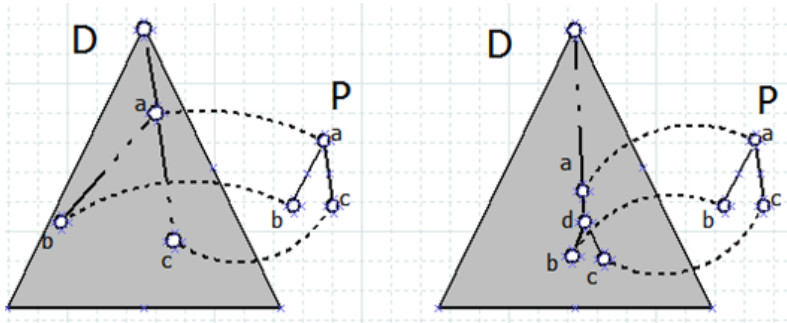


Figure : 11) Introduction de l'opération de coupure[9]

- Même complexité de temps et d'espace .

Algorithme dans un processus de Web Scraping

Rappelons que :

- HTML présente les informations sur les Pages Web ;
- Les changements sur une Page Web affectent son d'arbre ;
- Le patron est une structure de la partie des page HTML.

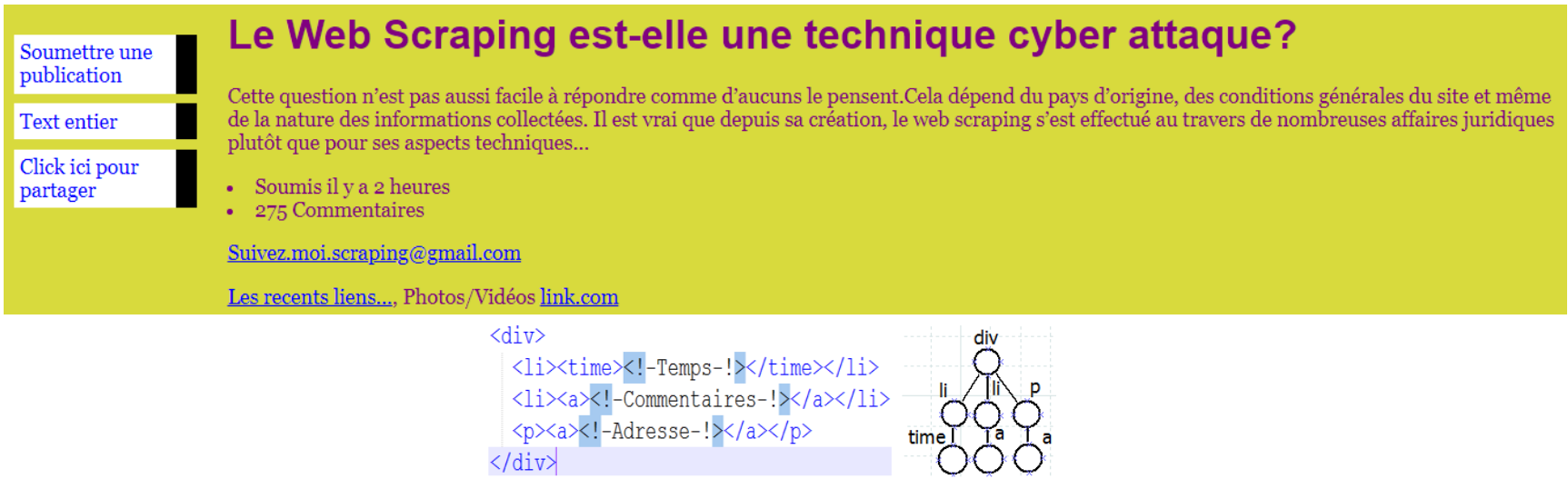


Figure : 12) Page sur link.net ainsi qu'un patron et son arbre

Comportement de l'algorithme

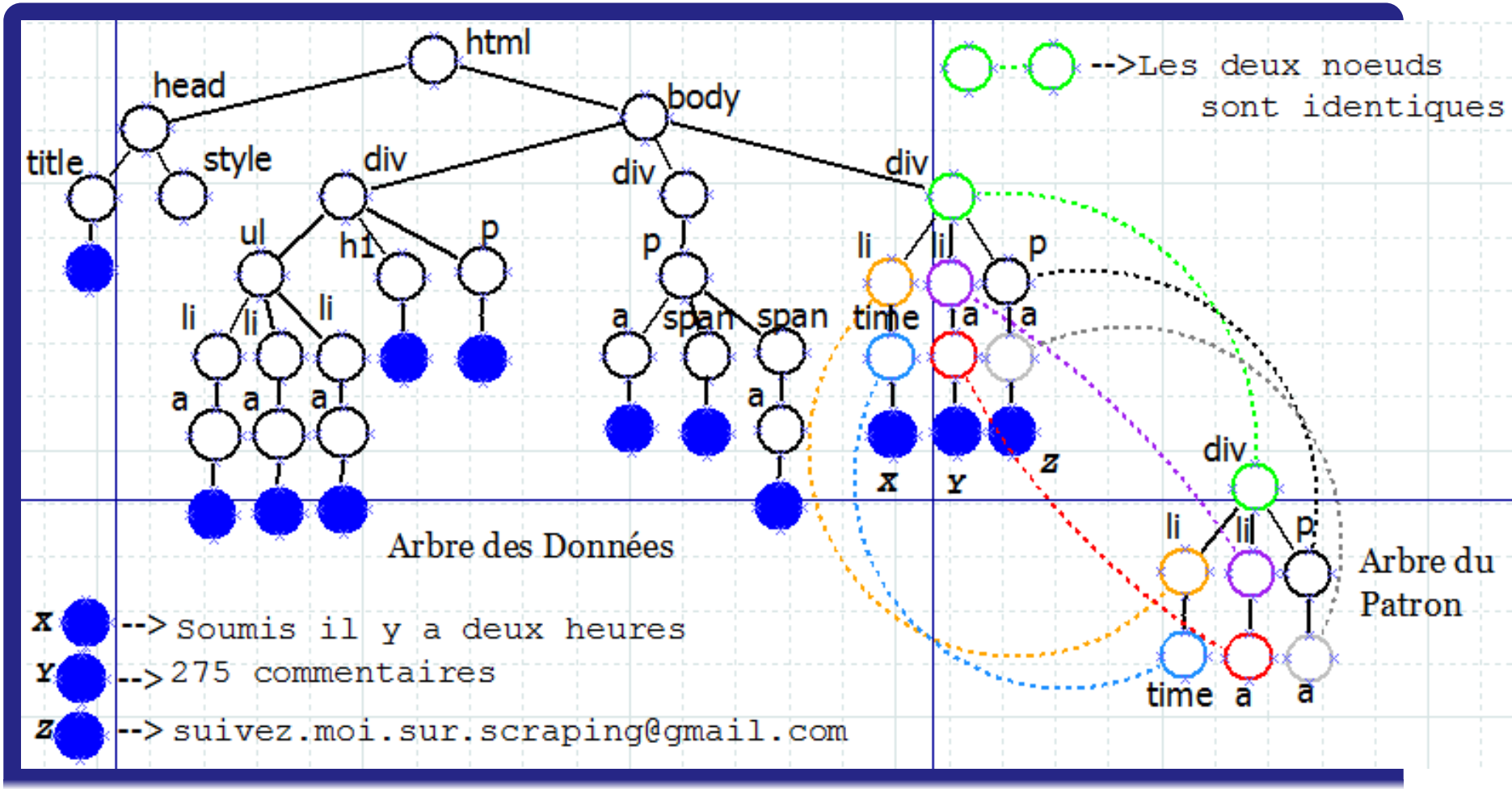


Figure : 13) Un mapping théorique

Comportement de l'algorithme

Figure : 14) Effort de similarité entre les deux arbres

- Supposons une représentation avec des balises <**a**>,<**time**>,<**li**> ;
- En considérant notre même patron de la **Figure** : 12) ;
- La transformation de l'arbre patron en arbre de données a changé.

Comportement de l'algorithme

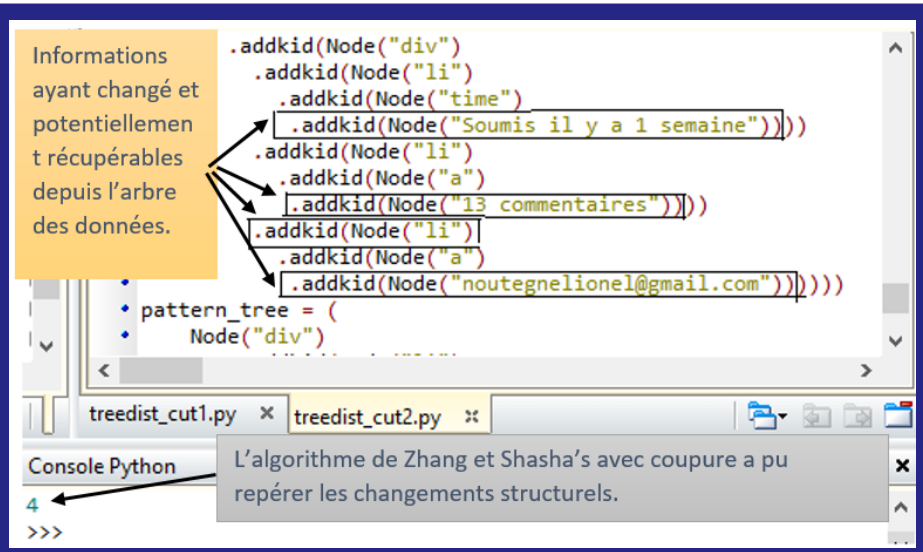


Figure : 15) Nouvelle représentation et effort structurel

Conclusion et perspectives

Problème

Détecter des **changements** au sein d'une page sans toutefois besoin de **régénérer** un **patron d'information**.

On peut dire

- La technique **d'appariement des structure arborescente** nous renseignent bien à cet effet ;
- **Zhang et Shasa's** permet de créer des **similarités** .

Perspectives

- Automatiser un nombre d'action ;
- Utiliser une structuration ;
- De la page web plus adaptée pour un accès plus de l'algorithme ;
- Pour un sortie des données potentiellement récupérable.


Merci pour
votre aimable
Attention !!!