Étude de la propagation au sein du Web à travers les liens hypertextes

José Rouillard, Jean Caelen

Étude de la propagation au sein du Web à travers les liens hypertextes

José Rouillard & Jean Caelen

Laboratoire CLIPS / IMAG Groupe GEOD
Campus Scientifique, BP 53
38041 Grenoble Cedex 9 - France
E-mail : Jose.Rouillard@imag.fr
E-mail : Jean.Caelen@imag.fr

RÉSUMÉ: Jusqu'où peut nous mener une page Web si l'on suit tous les liens hypertextes qu'elle comporte ? De quelle manière passe-t-on d'un pays à un autre ? La composition d'une page moyenne change-t-elle d'un continent à un autre ? Dans cet article, nous présentons une étude de la propagation sur le World Wide Web qui donne des éléments de réponse à ces questions. Pour cela, nous avons développé un robot informatique capable, à partir d'une page donnée, d'en étudier les composantes (nombre d'ancres, d'images, etc.), de les répertorier, et de naviguer automatiquement sur le réseau Internet en se projetant vers chacun des liens rencontrés.

ABSTRACT: How far can a Web page lead us if we follow all the hypertextuals links it contains ? How do we go from a country to another ? Does the composition of an average page change according to the continents ? The present paper describes a study of the propagation on the World Wide Web aiming at giving some answers to the above questions. To achieve that, we have developed a computer robot which is able to study the components (number of anchors, of pictures, etc.), of a given page, to record them, and to automatically navigate through the Internet network, plunging to each encountered link.

MOT-CLÉS : World Wide Web, études statistiques, navigation automatique, robot informatique.

KEY WORDS : World Wide Web, statistics studies, automatic navigation, computer robot.

1. Introduction

Dans cet article, nous présentons les résultats d'une étude concernant le contenu informationnel de pages au format HTML accessibles sur le World Wide Web et la propagation de liens en liens au sein du réseau Internet. Certains travaux ont déjà analysé la structure de documents types du web, en étudiant une collection de données récoltée par un puissant moteur de recherche [WOO 96], d'autres donnent un aperçu de ce à quoi ressemble une page 'moyenne' parmi les 50 millions de pages qui composent le WWW [BRA 96] ; mais, à notre connaissance, encore aucune expérimentation n'a était proposée pour montrer jusqu'où peut nous mener une page si l'on suit tous les liens hypertextes qu'elle comporte. Nous exposons dans les lignes qui suivent nos résultats concernant des statistiques comparatives faites sur des pages issues de différents continents à travers le monde et nous montrons la manière dont ces sites se font mutuellement référence.

2. Outils

Pour mener à bien notre projet, nous avons dû concevoir et développer un outil informatique (dit robot ou spider) capable, à partir d'une Uniform Resource Locator (URL) quelconque choisie par un utilisateur, de faire une analyse de cette page, d'en répertorier les principales caractéristiques, de sauver ces informations dans une base de données et de se projeter vers les liens référencés par ce document, et ainsi de suite pour toutes les nouvelles pages rencontrées. Ce robot a été écrit en langage Visual Basic version 4.0, pour PC sous Windows 95 et fonctionne grâce aux OCX HTML (HyperText Markup Language) et HTTP (HyperText Transfer Protocol) fournis par Microsoft. Les fichiers de données recueillies à chaque connexion sur un serveur ont été enregistrés au moment de l'étude sous forme standard ASCII, puis filtrés pour pouvoir être traités par un système de gestion de base de données (SGBD). Nous avons utilisé pour notre analyse (des requêtes principalement) le logiciel ACCESS version 7.0 de Microsoft. L'ordinateur que nous avons utilisé est un Pentium cadencé à 100 mégahertz, et équipé de 32 méga-octets de mémoire vive. La figure 1 ci-après donne un aperçu du robot conçu et développé pour notre étude.

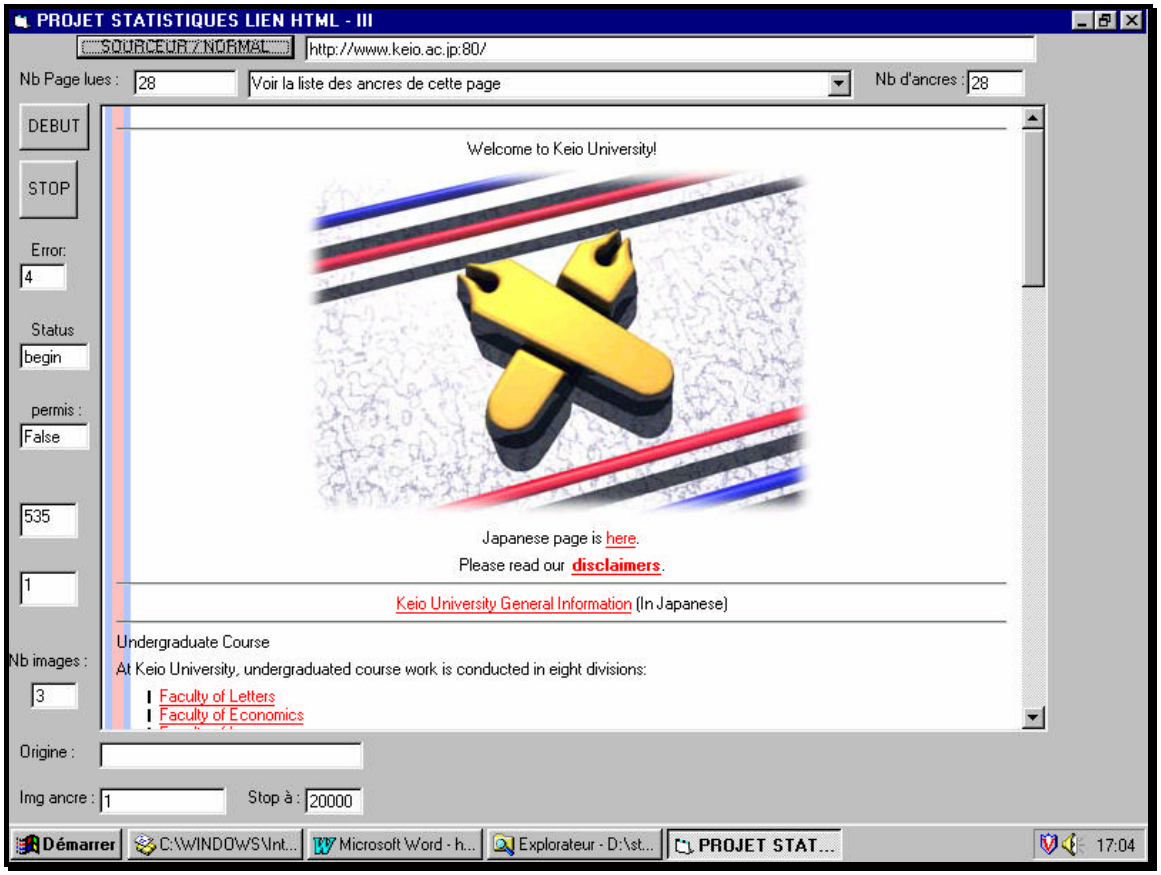

Figure 1 : Le robot utilisé pour notre étude

3. Méthodologie

Nous avons sélectionné comme URLs de départ six *homepages* parmi les pays suivants :
France (*Informatique et Mathématiques Appliquées de Grenoble*), **Allemagne** (Universität Augsburg), **États-Unis** d'Amérique (*Massachusetts Institute of Technology*), **Australie** (*Australien Broadcasting Corporation*), **Brésil** (*Université de São Paulo*) et **Chine** (*Université de Pékin*). Pour chaque page visitée, le robot sauvegarde dans une base de données[1] les informations suivantes : nombre d'ancres trouvées, nombre d'images, nombre d'images étant des liens, nombre de liens MAILTO, nombre de liens HTTP, nombre de liens FTP (File Transfer Protocol), nombre de liens NEWS, nombre de liens GOPHER, nombre de liens WAIS, nombre de liens TELNET, ainsi que l'origine de la page (pays ou institution qui a produit ce document hypermédia). Pour cette dernière information, le programme analyse l'URL et compare ses sous-chaînes de caractères avec une liste d'abréviations officielle pour les pays présents sur Internet. (fr pour France, it pour Italie, etc., [LIS 96]).

Une remarque importante doit être faite pour relativiser nos résultats : nous n'avons tenu compte dans ces statistiques que des URLs de formes absolues (adresse complète). De ce fait, notre robot accède davantage à des pages principales (home pages souvent plus riches et plus denses) qu'à des pages annexes, pour lesquelles les concepteurs se contentent la plupart du temps d'adresser les liens relativement à leur page de base. De même, les liens internes permettant d'accéder à un paragraphe ultérieur sur une même page HTML ont été ignorés.

---

1 On a donc créé 7 bases de données : BDUSA, BDFrance, BDAllemagne, BDAustralie, BDBrésil, BDChine et BDGlobale

D'autre part, nous avons fait en sorte qu'une page ne soit visitée qu'une seule fois : si en cours de route, une page fait référence à un lien déjà répertorié, ce dernier sera comptabilisé, mais ne sera pas envoyé vers la fonction d'appel, pour éviter des situations de bouclage.

Une contrainte de temps nous a également imposé de mettre en place un compteur incrémenté à chaque seconde. C'est ce compteur que vient tester une routine du robot, afin de vérifier si le document demandé ne pose pas de problème lors de la tentative de connexion ou de rapatriement des données (time out). Si c'est le cas, c'est à dire si l'on observe un dépassement du temps supposé nécessaire pour récupérer une page HTML, l'opération est annulée, et l'on passe au lien suivant. La figure 2 ci-dessous montre la méthode adoptée pour récolter les données de notre étude. C'est un parcours en largeur de l'arbre de liens :

Phase A: A partir d'un document initial, le robot détecte toutes les URLs disponibles sur cette page (exemple : les N premiers liens dans notre figure 2), tout en stockant dans une base de données les adresses rencontrées.

Phase B : Il explore toutes les pages répertoriées lors de l'analyse de la première page (URL1), soit les URLs 1.1 à 1.5 sur notre figure. Les nouveaux liens trouvés sont sauvegardés dans la base de données.

Phase C : Le robot appelle l'URL suivante dans la liste, donc l'URL 2 dans notre exemple, et fournit le même travail que précédemment jusqu'au lien N.

Phase D : On se dirige ensuite vers le premier lien (1.1.1) de la première page (1.1) de la première URL (URL 1) issue de la page originale. *Ad infinitum*.

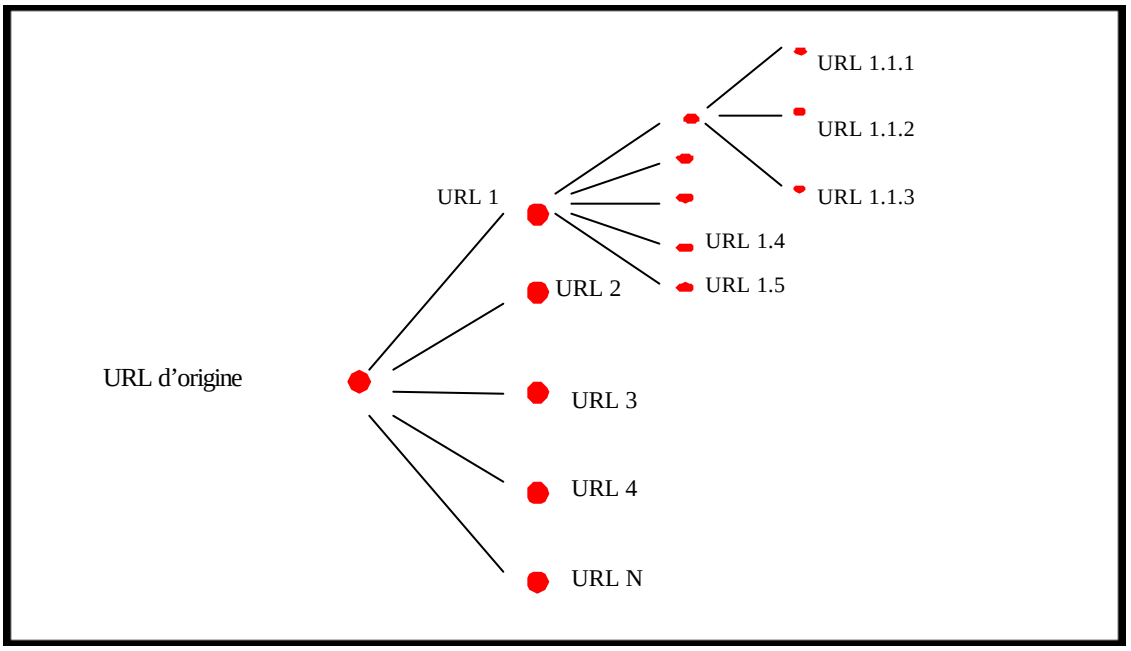

Figure 2 : Mode de propagation du robot sur le Web.

4. Résultats

L'expérience s'est déroulée de janvier à mars 1997. Le nombre total de pages visitées se situe autour de 50000. Nous présentons tout d'abord des résultats statistiques en fonction de la page d'origine pour les 6 pays étudiés, et ensuite des résultats globaux regroupant l'ensemble des données récoltées au cours de ces expériences.

4.1. Résultats en fonction du pays d'origine de la page primaire

A partir de :	France	Usa	Australie	Allemagne	Brésil	Chine	*Moyenne*
Moyenne ANCRE	33,43	11,66	19,45	13,20	11,06	9,63	16,41
Moyenne IMG	5,28	6,42	8,90	6,14	10,35	7,20	7,38
Moyenne IMG CLIC	2,68	3,10	4,45	2,97	6,20	3,44	3,81
Moyenne HTTP	32,30	10,14	17,82	11,78	10,11	8,71	15,14
Moyenne FTP	0,31	0,63	0,51	0,67	0,23	0,08	0,41
Moyenne MAILTO	0,62	0,60	1,00	0,51	0,51	0,74	0,66
Moyenne NEWS	0,04	0,06	0,08	0,07	0,04	0,01	0,05
Moyenne GOPHER	0,15	0,15	0,03	0,12	0,15	0,08	0,11
Moyenne WAIS	0,01	0,03	0,00	0,00	0,00	0,00	0,01
Moyenne TELNET	0,01	0,04	0,01	0,04	0,00	0,01	0,02

Tableau 1 : Moyenne pour chacun des éléments, par page et en fonction du pays d'origine

Le tableau 1 ci-dessus synthétise les résultats concernant le nombre moyen d'éléments présents dans une page, compte tenu du pays d'origine de la première page. Les éléments remarquables sont surlignés. La moyenne du nombre d'ancres par page dans la base de données France peut paraître très élevée.

L'échantillon pour cette statistique est de 11227 éléments, et l'écart type que nous avons calculé vaut 75,12. Il apparaît donc qu'il doit y avoir une répartition atypique autour de la moyenne du nombre d'ancres. C'est ce que montre la figure 3 ci-dessous : 3133 pages ne possèdent aucun lien, 1010 pages n'en ont qu'un seul, 593 en ont entre 11 et 15 et 738 pages abritent entre 51 et 100 ancres. On peut remarquer également que 1076 pages disposent de plus de 101 liens ; le maximum étant de 782 ancres dans une seule et même page sur la base de données France.

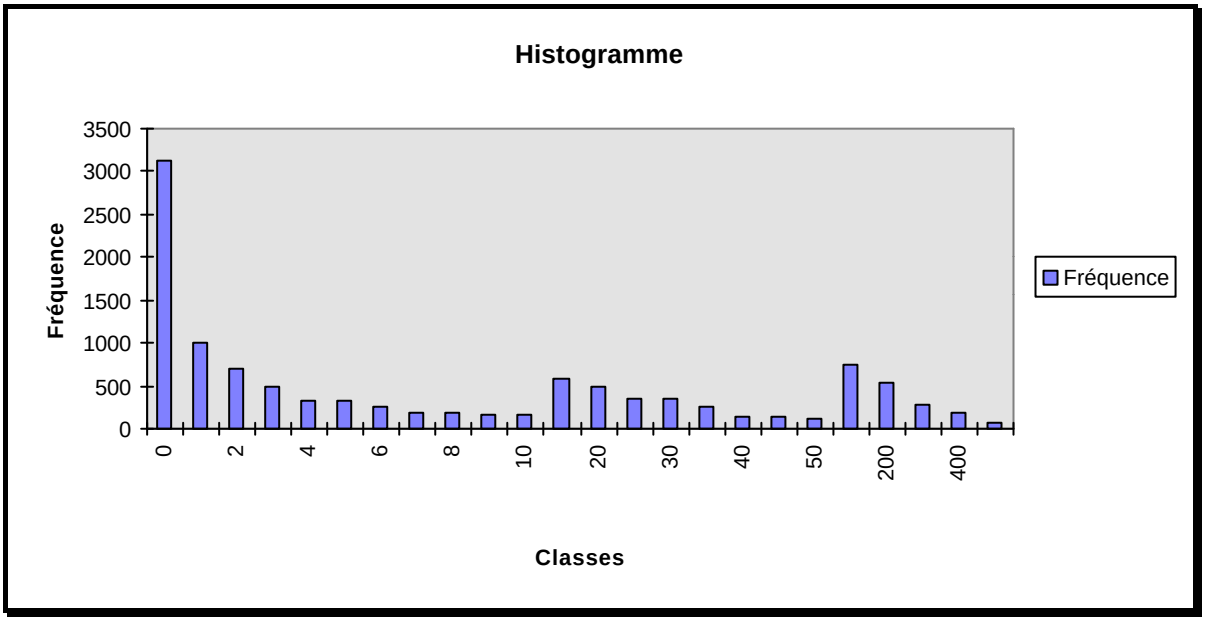

Figure 3 : Fréquence du nombre d'ancres dans la base France (de 0 à 10 avec un pas de 1, de 11 à 50 avec un pas de 5, plus de 51 avec un pas de 100)

On note donc immédiatement un certain clivage selon l'origine de la page primaire. Mais à ce stade de l'analyse, on ne peut pas encore dire qu'un pays utilise (par exemple) plus d'images qu'un autre dans la conception de ces pages HTML. En effet, on ne connaît pour le moment que le pays d'origine pour une page de base donnée, et rien ne nous permet d'affirmer que les pages auxquelles elle fait référence demeurent dans le même pays (ou institution). Pour cela, nous aurons à étudier le taux de propagation vers les autres pays pour chacune des 6 bases de données.

A partir de :	France	Usa	Australie	Allemagne	Brésil	Chine	*Maximum*
Max ANCRE	782	2667	5640	2944	1605	1311	5640
Max IMG	337	301	1050	276	4111	538	4111
Max IMG CLIC	294	295	810	186	4111	189	4111
Max HTTP	781	2400	5640	2944	1528	1311	5640
Max FTP	306	671	163	338	45	22	671
Max MAILTO	300	333	538	300	24	93	538
Max NEWS	160	97	174	141	10	23	174
Max GOPHER	93	92	105	44	53	32	105
Max WAIS	22	43	2	2	2	1	43
Max TELNET	21	40	18	32	4	8	40

Tableau 2 : Maximum pour chacun des éléments en fonction du pays d'origine

Des documents HTML ayant un très grand nombre de liens sont en fait des pages d'index pointant vers une multitude de sites au sein même du pays ou à l'étranger. De même, la page qui contient 4111 images cliquables (cf. tableau 2) n'a pas été créée « à la main » par un concepteur, mais est en réalité un accès à un répertoire de serveur (ces images sont toutes les mêmes : un logo permettant de charger un fichier parmi les 4111 disponibles).
Nous présentons à présent ci-dessous les résultats de propagation à partir de nos six pages de base choisies dans les pays suivants : France, USA, Allemagne, Australie, Brésil et Chine.

4.1.1. Propagation depuis la France vers les autres pays :

Suisse	35,54
Commerce US	16,37
Éducation US	11,20
Afrique du Sud	4,58
Organisation non commerciale	3,87
Royaume Uni	3,77

Allemagne	3,56
Réseau	2,90
France	2,77
Gouvernement US	2,63
Australie	2,08
Hong Kong	1,97
Canada	1,33

Japon	0,85
Pays-Bas	0,74
Inconnu	0,68
Singapour	0,62

Tableau 3 : Taux de répartition des 17 pays les plus référencés depuis une page originale en France

En partant de France, le robot a donc visité 64 pays différents (identifiables). 90% du cumul de répartition au sein des pays est atteint avec les 11 premiers pays. La France ne pointe sur elle-même qu'avec un taux de 2,77 % et se place ainsi au 9ème rang. On ne peut donc (ici) pas parler de phénomène centripète franco-français en matière de mise à disposition aux usagers de liens vers d'autres sites Internet. Le fait que la Suisse soit en tête de liste peut sans doute s'expliquer par le fait que le CERN (European Laboratory for Particle Physics) soit référencé par l'IMAG dans ses toutes premières pages. Le premier lien étranger atteint est en fait le moteur de recherche Altavista de Digital. Lorsque le robot n'a pas réussi à identifier la provenance d'une URL, il l'a note comme étant d'origine *Inconnue*. Typiquement, c'est une adresse de ce genre : http://140.251.5.102:80/. Il n'existe pas de sous chaîne de caractères identifiables dans cette URL (pourtant valide), ce qui empêche toute comparaison avec une liste d'abréviations de pays ou d'institutions.

4.1.2. Propagation depuis les USA vers les autres pays :

Éducation US	45,19
Commerce US	14,62
Gouvernement US	6,79
Organisation non commerciale	6,64
Allemagne	5,69
Canada	2,89

Royaume Uni	2,88
Réseau	1,84
Militaire US	1,67
Japon	1,65
France	1,63
Australie	1,30
Pays-Bas	1,01

Italie	0,83
Suisse	0,67
Inconnu	0,58
Espagne	0,55

Tableau 4 : Taux de répartition des 17 pays les plus référencés depuis une page originale aux USA

Depuis les États-Unis, le robot a visité 58 pays différents. Les 6 premiers liens cumulent plus de 81% des pays visités avec 66,60% de liens qui pointent sur les USA eux-mêmes. Les 407 et 408ème liens de la base de données USA sont les premiers à pointer à l'extérieur des États-Unis : l'un vers l'Australie et l'autre vers une université italienne.

4.1.3. Propagation depuis l'Allemagne vers les autres pays :

Allemagne	70,06
Commerce US	5,33
Autriche	4,98
Éducation US	4,75
Organisation non commerciale	2,66
Réseau	2,30

Royaume Uni	1,39
Suisse	1,11
France	0,77
Gouvernement US	0,74
Inconnu	0,69
Canada	0,56
Pays-Bas	0,49

Belgique	0,33
Australie	0,28
Japon	0,26
Italie	0,25

Tableau 5 : Taux de répartition des 17 pays les plus référencés depuis une page originale en Allemagne

Depuis la page de l'université d'Augsburg, 65 pays différents ont été visités. L'Allemagne fait référence à elle-même avec un très fort taux (plus de 70%), puis aux USA (environ 10%) et à d'autres pays européens avec des taux nettement moins importants. Le premier lien étranger rencontré se situe au 165ème rang et pointe vers l'Autriche (server of the Physics Department of the University of Innsbruck).

4.1.4. Propagation depuis l'Australie vers les autres pays :

Éducation US	40,22
Commerce US	30,49
Japon	6,18
Réseau	4,44
Allemagne	2,98
Organisation non commerciale	2,73

France	1,68
Royaume Uni	1,67
Inconnu	1,12
Canada	0,87
Gouvernement US	0,86
Pays-Bas	0,59
Australie	0,52

Italie	0,50
République Tchèque	0,49
Finlande	0,48
Suisse	0,41

Tableau 6 : Taux de répartition des 17 pays les plus référencés depuis une page originale en Australie

En partant d'Australie, 58 pays différents ont été visités par le spider. On observe que les 8 premiers pays cumulent déjà plus de 90% de la répartition des pays visités, dont environ 70% rien que pour les États-Unis. On remarque également qu'en partant d'une page australienne, on ne retrouve que 0,52% de liens qui pointent vers l'Australie. Cela peut être parce qu'une fois arrivé en Amérique, le phénomène de propagation USA-USA reprend le dessus. Le quatrième lien pointe déjà vers les États-Unis dans la base de données Australienne. Viennent ensuite l'Italie en 151ème position, la Suède (en 180), Les Bermudes (en 191) ; la France apparaît au 197ème lien (http://www.toplog.FR:80).

4.1.5. Propagation depuis le Brésil vers les autres pays :

Commerce US	41,55
Organisation non commerciale	16,12
Éducation US	12,56
Réseau	4,97
Brésil	4,55
Australie	3,43

Canada	2,44
Pays-Bas	2,44
Royaume Uni	2,27
Gouvernement US	2,27
Allemagne	1,16
Inconnu	1,07
Suède	0,81

USA	0,73
Luxembourg	0,51
Belgique	0,43
France	0,34

Tableau 7 : Taux de répartition des 17 pays les plus référencés depuis une page originale au Brésil

Depuis l'université de São Paulo au Brésil, le robot a traversé 39 pays différents. Le 14ème lien pointe vers le premier pays étranger : le Portugal. Ce n'est qu'au 5ème rang et avec un taux de 4,55% que le Brésil s'auto-référence. Les pays anglophones sont majoritaires dans ce classement.

4.1.6. Propagation depuis la Chine vers les autres pays :

Éducation US	40,17
Commerce US	26,68
Organisation non commerciale	24,23
Gouvernement US	2,90
Réseau	1,26
Inconnu	0,90

USA	0,69
Royaume Uni	0,42
China	0,39
Canada	0,30
Guatemala	0,24
Australie	0,21
Allemagne	0,21

Norvège	0,15
Pays-Bas	0,12
Afrique du Sud	0,12
Belgique	0,09

Tableau 8 : Taux de répartition des 17 pays les plus référencés depuis une page originale en Chine

Depuis l'université de Pékin, l'Irlande est le premier des 38 pays étrangers visités ; viennent ensuite la Belgique, les États-Unis et l'Australie. La proportion de liens chinois vers la Chine n'est que de 0,39% (9ème rang). En revanche, les USA sont pointés à plus de 70%.

La figure 4 ci-dessous montre quels sont les grandes relations entre les principaux pays étudiés. Le continent américain est le plus référencé dans notre étude. Non seulement la majorité des liens américains pointent vers d'autres liens aux USA, mais on note aussi que tous les autres pays, sans exception, proposent un accès aux pages des États-Unis. Le Brésil pointe donc surtout sur les USA, mais également vers l'Australie et dans une moindre mesure vers quelques pays européens. La Chine se comporte sensiblement de la même manière, tandis qu'en Europe, la France et l'Allemagne sont ceux qui ont désigné le plus grand nombre de pays dans le monde (respectivement 64 et 65). Globalement, on voit donc que d'où que l'on parte sur la planète, on a de grandes chances, grâce au Web, de faire le tour du monde, si l'on suit un à un les liens d'une page standard.

Figure 4 : Comment les pays se font référence majoritairement ?

4.2. Résultats comparatifs par rapport à l'ensemble des données

Au total, 81 pays ou institutions ont été visités par notre robot. Nous présentons sur la figure 5 ci-après les pays ou institutions les plus importantes dans notre base globale, en fonction du taux d'apparition. Les facultés (et le milieu universitaire en général) y sont les institutions les plus représentées avec 29,17% de taux d'apparition. Ce sont ensuite les documents à caractères commerciaux des États-Unis d'Amérique avec 20,84%, puis l'Allemagne avec 11,61%. La Suisse arrive au 4ème rang avec un taux de 8,30%. La France ne représente que 1,62% de la base de données globale, le Brésil 0,27% et la chine 0,05%.

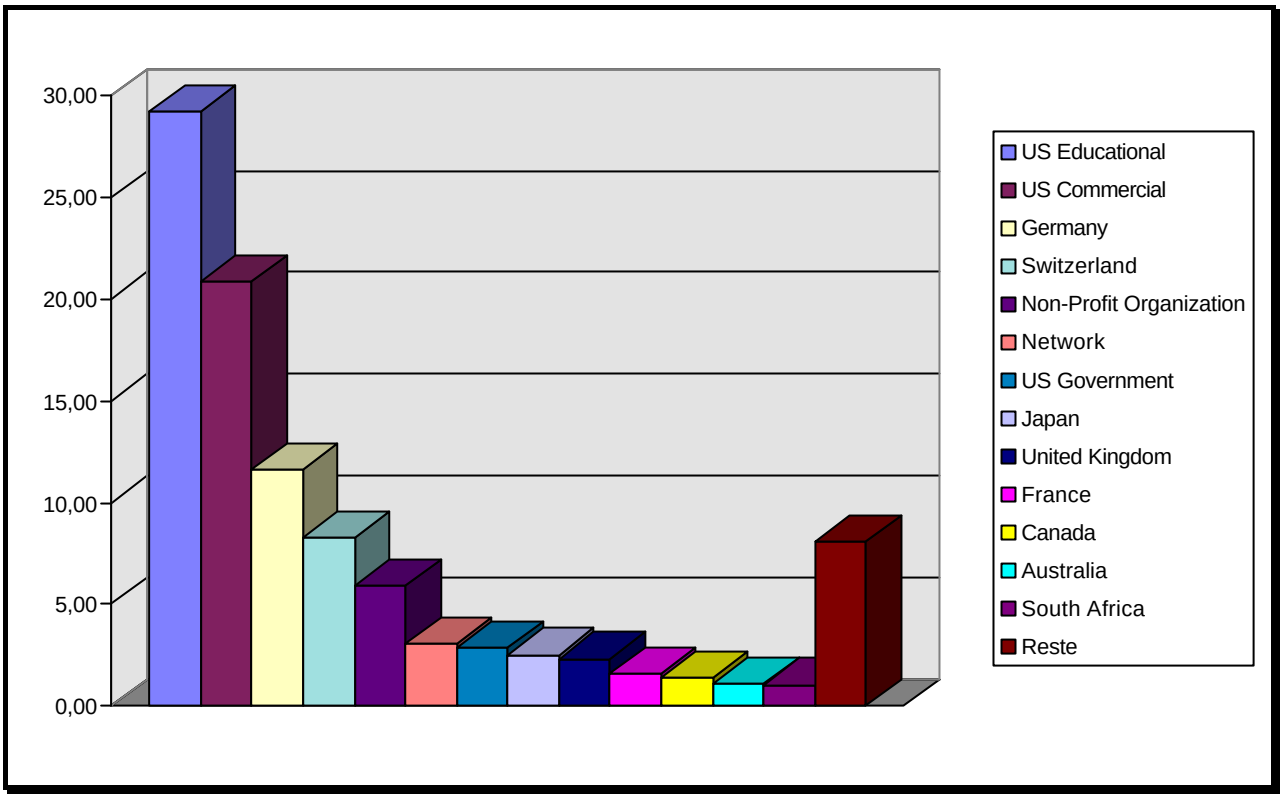

Figure 5 : Le pourcentage d'apparition des pays les plus présents dans la base de données globale

Grâce à la base de données globale, regroupant les informations des 6 autres bases de données, nous pouvons à présent déterminer quel est le nombre moyen d'éléments composant une page HTML 'standard' pour chacun des pays visités. Concernant le nombre moyen de liens hypertextuels par page, la Suisse détient le record avec plus de 71 ancres en moyenne par document. On trouve ensuite l'Australie avec 25 liens moyens par page, « Éducation US » (20 liens), l'Autriche, le Royaume Uni (12 liens). La France avec 6,66 liens en moyenne par page se place au 16ème rang.

Le nombre moyen d'images selon les pays varie de 11,58 pour « Réseau[2] » à 1 pour la Suisse. Dans l'ordre, on voit « Commerce US » avec 10,28 images par page, l'Autriche, les Pays-Bas, l'Afrique du Sud, le Japon, Hong Kong (environ 9 images) ; l'Australie, la France, le Canada, l'Angleterre et l'Allemagne (environ 6 images). On ne différencie pas ici une image de type « bouton rouge » (toute petite image très courante sur le Web pour attirer l'œil) d'une autre image plus importante (logo, image de fond, etc.).

La figure 6 ci-après, représente le nombre moyen d'images (couleur claire) des pays ou institutions les plus représentatifs de notre base de données. Nous montrons aussi le nombre d'images qui sont également des liens, autrement dit des images cliquables (couleur foncée sur la figure) qui envoient vers un autre document au sein d'Internet. Dans la plupart des sites visités, on observe que le nombre d'images cliquables est environ la moitié de celui des images simples. On voit donc bien que, comme le suggéraient certaines idées et réflexions du début des années 90 ([BAL 90] et [BER 91]), dans le domaine de l'hypermédia, l'image joue déjà un grand rôle en tant que lien entre documents. L'image n'est donc plus seulement un élément « décoratif » et attrayant pour capter l'attention de l'usager, elle donne également, presque une fois sur deux, la possibilité d'interagir avec la machine.

[2] Institution Network

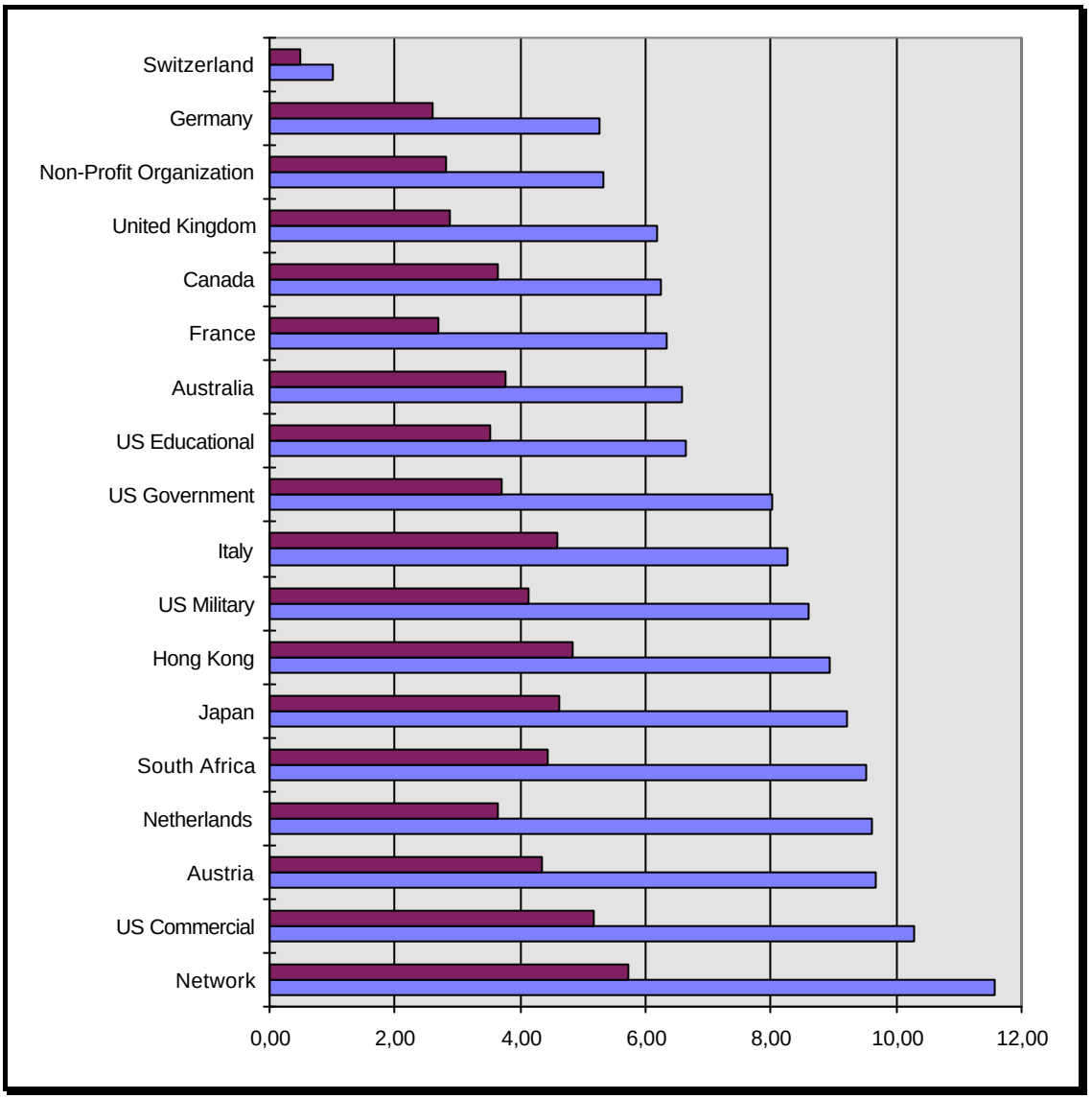

Figure 6 : Le nombre moyen d'images et d'images étant des liens, par page et par pays.

L'Australie possède le plus grand nombre moyen de tag « mailto » (3,54) au sein de ses pages HTML. On retrouve ensuite l'Afrique du Sud, avec 3,33 mailto par page et le Royaume Uni avec 1,43. Les États-Unis sont à 0,80 et tous les autres pays représentatifs oscillent entre 0,3 et 0,6. La France est à 0,60 et la Suisse n'a, en moyenne, que 0,12 lien de ce type (mailto) dans ses pages.

5. Conclusions et perspectives

Dans cette étude de la propagation au sein du Web grâce aux liens hypertextuels, nous avons montré les différents cheminements possibles auxquels nous mènent une URL standard et la manière dont on passe d'un pays à l'autre sur le Web. Il apparaît que la composition d'une page moyenne varie sensiblement selon son pays d'origine, en terme de nombres d'ancres, d'images et de références à d'autres types de liens. Wendy Hall écrit «In the WWW, highlighted words in HTML documents indicate that more information about that term is available at the click button.» [HAL 95], il faut ajouter à présent que le clic souris se fait également sur les images, et peut être même plus facilement que sur du texte, selon sa taille à l'écran.

Pour le domaine de l'interaction Homme-Machine, cette étude apporte quelques éléments intéressants à prendre en compte. On sait que les utilisateurs du Web peuvent être très vite perdus au milieu de tant de pages disponibles. En effet, selon Conklin [CON 87] les deux plus importants problèmes liés à l'accès aux informations par le biais d'interfaces hypermédia sont la désorientation et la surcharge cognitive. On voit ici qu'en plus, l'usager peut très facilement passer d'un site à un autre, et traverser les pays sans pour autant en prendre conscience. En effet, à l'heure actuelle et à notre connaissance, aucun

feuilleteur ne nous informe explicitement du fait que l'on passe d'un site à un autre, ni d'un pays à un autre. Une autre notion doit être prise en compte dans cette étude, c'est le fait que nous avons lancé notre robot essentiellement avec des pages HTML du milieu universitaire, et il est donc logique que l'on retrouve dans nos statistiques ces sites pointant les uns vers les autres. Cependant, compte tenu de la taille de l'échantillon des données étudiées, nous ne pouvons, ici, que donner de grandes lignes directives sans bien entendu pouvoir en conclure des stéréotypes fidèles pour chaque pays visitée. Nous avons également manqué de temps pour étudier des pages de base du continent Africain.

La première version du programme que nous avons créé pour explorer le réseau Internet automatiquement ne pouvait travailler que sur une seule base de données à la fois. Une amélioration a été apportée dans ce sens, afin de pouvoir lancer plusieurs robots simultanément. De même, nous avons mis au point, au sein de notre équipe de travail, un « parser » en langage Perl fonctionnant sous Unix, qui ne charge pas la page HTML en entier mais seulement son descriptif (le nom d'une image plutôt que l'image elle-même par exemple), ce qui améliore considérablement la vitesse du robot. Cet outil est couplé à un SGBD et les deux applications se font mutuellement référence. De cette manière, nous allons pouvoir orienter nos recherches, avec des requêtes spécifiques pour sélectionner les pays qui nous intéressent. Au lieu de suivre une URL comme nous l'avons fait dans la présente étude, nous pourrons sélectionner à l'avance des URLs dans la base de données pour les envoyer au robot-parser et obtenir des informations *ad hoc*.

Références

[BAL 90] BALPE J.P., «Hyperdocuments, hypertextes, hypermédias » , Eyrolles , Paris, 1990

[BER 91] BERK E., DEVLIN J., «Hypertext / Hypermedia handbook » , McGraw Hill , 1991

[BRA 96] BRAY T. « Measuring the Web » , Fifth International World Wide Web Conference, May 6-10, 1996, Paris

[CON 87] CONKLIN J., «*Hypertext: an introduction and survey*», IEEE Computer, pp. 17-41, September 1987.

[HAL 95] HALL W., «The role of hypermedia in multimedia information systems » ACM Computing surveys, vol 27, N° 4 Décembre 1995

[WOO 96] WOODRUFF A., AOKI P.M., BREWER E., GAUTHIER P., ROWE L.A., « An Investigation of Documents from the World Wide Web » , Fifth International World Wide Web Conference, May 6-10, 1996, Paris

[LIS 96] Liste des pays présents sur le Web : http://www.w3.org/pub/DataSources/WWW/Geographical.html