Visualisation de la perception d'un site web par ses utilisateurs[1]

Fabrice Rossi[2], Yves Lechevallier et Aïcha El Golli
Projet AxIS
INRIA Rocquencourt
Domaine de Voluceau, Rocquencourt, B.P. 105
78153 LE CHESNAY CEDEX – FRANCE
Email : {Fabrice.Rossi,Yves.Lechevallier,Aicha.ElGolli}@inria.fr

Résumé. Nous proposons dans cet article une méthode de visualisation de l'activité des utilisateurs d'un site web qui permet d'évaluer qualitativement l'adéquation entre son architecture logique et la perception de celle-ci par les internautes. Nous travaillons sur les parcours des internautes sur le site étudié, après reconstruction de ceux-ci grâce aux fichiers logs des serveurs concernés. Nous utilisons la structure logique des sites étudiés pour simplifier la représentation des parcours, en ne tenant pas compte de l'ordre de visite des catégories sémantiques du site. Les parcours simplifiés sont utilisés pour calculer une dissimilarité entre les catégories sémantiques qui sont ensuite représentées dans un plan par *Multi Dimensional Scaling*. Nous complétons cette visualisation d'ensemble par une représentation de l'arbre couvrant minimal des catégories sémantiques qui permet de mieux appréhender certaines interactions. Nous illustrons l'intérêt de la méthode en l'appliquant au site de l'INRIA.

1 Introduction

La construction puis la maintenance continue d'un site Web de taille importante demandent un travail considérable sans lequel le site perd peu à peu tout intérêt aux yeux du public. Le contenu lui-même doit bien entendu correspondre au public visé, mais cela ne suffit pas. L'organisation hyper-textuelle d'un site Web induit en effet un mode de parcours totalement différent de celui des médias traditionnels : il n'y a plus de début et de fin, l'utilisateur étant libre d'interrompre à tout moment un parcours linéaire pour suivre un hyperlien, puis revenir au document précédent grâce à l'opération "page précédente" de son navigateur.

A cette complexité du média, s'ajoute celle induite par les ressources externes. L'indexation d'un site par les moteurs de recherche de référence peut par exemple créer une structure de navigation totalement différente de celle envisagées par les concepteurs du site. L'inclusion du site dans des listes de favoris ou dans des annuaires thématiques peut créer des rapprochements incongrus ou de nouvelles structures de navigation.

Les responsables d'un site Web ne peuvent donc pas se contenter de simple statistiques d'accès pour comprendre l'utilisation de leur site par les internautes. Pour les

[1]Publié dans les actes de la conférence EGC'2005
Disponible à http://apiacoa.org/publications/2005/egc05.pdf
[2]Les coordonnées actuelles de Fabrice Rossi sont disponibles à l'URL http://apiacoa.org/

raisons évoquées au dessus, il est nécessaire en effet de confronter la conception du site à sa perception par les utilisateurs. Pour ce faire, il est possible d'utiliser les traces laissées par les visiteurs d'un site sous forme des fichiers logs des serveurs concernés. Il s'agit alors de réaliser une forme particulière de *Web Usage Mining* (WUM) dans laquelle on cherche à se focaliser sur la perception de l'organisation et du contenu d'un site par ses utilisateurs. Le WUM est utilisé depuis une dizaine d'années dans le but de comprendre et d'améliorer les sites Web (cf (Srivastava et al., 2000) par exemple pour une présentation synthétique des objectifs principaux du WUM).

Nous proposons dans cet article une méthode de visualisation d'un site Web inspirée des travaux sur la visualisation d'un domaine de connaissances (Chen and Paul, 2001; Noel et al., 2003). En utilisant la structure du site étudié, nous définissons des groupes de documents faciles à interpréter. Les données d'usage du site nous permettent de calculer des dissimilarités entre ces groupes que nous visualisons au moyen du *Multi Dimensional Scaling* et, comme dans (Chen, 1998), en utilisant l'arbre couvrant minimal induit par les dissimilarités.

La suite de cet article est organisée de la façon suivante. Nous présentons notre problème dans la section 2, en évoquant brièvement les solutions existantes et leurs limitations. Nous exposons la méthode proposée dans la section 3, puis nous l'illustrons sur le site Web de l'INRIA dans la section 4.

2 Visualiser un site Web

2.1 Le contenu d'un site

Un site Web est constitué de documents identifiés par des URLs (*Uniform Ressource Locators*, un cas particulier des *Uniform Ressource Identifiers* (Berners-Lee et al., 1998)). Un URL est de la forme simplifiée suivante : `http://<host>/<path>` (dans cet article, nous ne prendrons pas en compte la partie recherche qui peut terminer un URL). La partie `<host>` correspond au nom DNS du serveur considéré alors que la partie `<path>` correspond au chemin d'accès au document demandé sur le serveur. L'URL `http://www-sop.inria.fr/axis/` correspond ainsi au serveur `www-sop.inria.fr` et au document `axis/` sur ce serveur. Nous ne restreignons pas notre travail à l'analyse d'un site hébergé sur un seul serveur, i.e. d'une partie `<host>` unique. Pour prendre en compte les sites Web complexes utilisant plusieurs serveurs, nous considérons que l'`<host>` peut varier.

La plupart des documents d'un site Web sont des pages au format (X)HTML (Group, 2002; Raggett et al., 1999) qui contiennent des hyperliens, c'est-à-dire des références vers d'autres documents accessibles sur le Web (sous forme d'URLs). En raison des références internes, un site Web est donc un graphe dont les noeuds sont les documents et les arêtes les liens inclus dans les documents.

2.2 Visualisation

Un site Web de taille moyenne peut contenir des milliers de documents et il devient donc rapidement difficile d'en avoir une vue d'ensemble. C'est pourquoi de très

nombreuses méthodes de visualisation de site Web ont été proposées (cf Benford et al., 1999; Dodge, 2004). Nous nous intéressons dans cet article à une classe particulière de méthodes de visualisation : il s'agit de faire apparaître comment les utilisateurs s'approprient le site Web. Plus précisément, nous cherchons à visualiser la perception du contenu d'un site par ses utilisateurs en faisant ressortir les rapprochements entre documents opérés par les utilisateurs.

Bien que de nombreuses méthodes de visualisation aient été proposées dans le contexte du WUM, celles-ci sont assez mal adaptées au problème auquel nous nous intéressons. En effet, elles ont essentiellement été créées afin de visualiser le cheminement des internautes sur un site Web. Elles se focalisent en fait sur la représentation des chemins fréquents dans un site (e.g., Cadez et al., 2000; Chen et al., 2004; Chi, 2002; Chi et al., 1998; Cugini. and Scholtz, 1999; Dodge, 2001; Youssefi et al., 2004), c'est-à-dire des suites de pages visitées par une proportion importantes d'utilisateurs. Ces motifs fréquents sont très utiles en WUM, car ils constituent entres autres une base intéressante pour les algorithmes de personnalisation et de recommandation (e.g., Mobasher et al., 2000; Srivastava et al., 2000).

De plus, les méthodes de visualisation proposées partagent en général un élément important : la représentation d'un site consiste à placer des symboles correspondant aux documents d'une façon adaptée au tracé des liens entre ces documents tels qu'ils sont définis par la structure de graphe du site. Des enrichissements graphiques (couleurs, épaisseur des traits, etc.) sont utilisés pour mettre en avant les liens les plus actifs. Dans certaines visualisations (e.g., Chen, 1998; Chen et al., 2004; Chi, 2002) les données d'usage du site favorisent la visualisation de certains liens.

Contrairement à ces travaux, nous souhaitons faire apparaître des liens entre documents même si ceux-ci sont indirects et ne reposent pas sur la structure de graphe du site. De plus, nous souhaitons disposer les symboles représentant les documents de sorte que les proximités induites par les utilisateurs soient le plus apparentes possibles. Les liens hypertextuels ne nous semblent pas l'information la plus pertinente pour analyser la perception globale d'un site important par ses utilisateurs. L'ordre précis d'un parcours, le passage exact par certains liens, etc. ne sont pas des éléments déterminants pour comprendre la perception globale du site. Il faut au contraire avoir une vision plus grossière des navigations afin d'en extraire les rapprochements réalisés par les utilisateurs entre les différentes parties du site. Nous proposons donc de travailler directement sur les navigations réalisées par les utilisateurs sur le site afin d'en déduire une mesure de dissimilarité entre ses pages, sans tenir compte de la structure de graphe du site.

3 Principe de la méthode proposée

3.1 Préparation des données

3.1.1 Pré-traitements

Les données d'usage d'un site Web proviennent essentiellement des fichiers log des serveurs concernés. Ceux-ci sont généralement écrits dans le format CLF (Luotonen, 1995) ou dans sa version étendue qui comporte plus d'informations, en particulier le *User Agent* un élément très important pour la reconstruction des navigations (il

s'agit en général du nom du logiciel de navigation utilisé ainsi que d'une information sur le système d'exploitation, par exemple `"Mozilla/5.0 (X11 ; U ; Linux i686 ; rv :1.7.3) Gecko/20041001 Firefox/0.10.1"` correspond au logiciel Firefox utilisé sous Linux). Une des premières difficultés du WUM est de reconstruire le comportement de chaque utilisateur à partir des logs. Les logs sont en effet constitués de lignes indépendantes, ordonnées selon les dates des requêtes, et contenant entres autres l'adresse IP du client associé à une requête et son *User Agent*. Il faut donc combiner les lignes associées pour reconstruire l'historique d'un utilisateur. De plus, nous souhaitons réaliser des analyses multi-sites impliquant plusieurs serveurs. Certains utilisateurs passeront naturellement d'un site à un autre (cf la table 1 pour un exemple) : pour reconstituer la trajectoire d'un utilisateur, il faut donc fusionner les fichiers logs.

Nous ne détaillerons pas ici la méthode retenue pour la préparation des données : nous utilisons les algorithmes proposés dans (Tanasa and Trousse, 2004a,b). Ceux-ci permettent de travailler sur des données multi-sites, en supprimant les requêtes provenant de robots et en reconstruisant efficacement les navigations des utilisateurs (un utilisateur est défini par un couple adresse IP et *User Agent*). Nous supposerons donc dans la suite de cet article que nous disposons de données nettoyées et sous la forme suivante : pour chaque utilisateur du site, nous avons une liste d'URLs (les documents demandés) avec pour chacune d'eux la date de la requête correspondante. Nous ne conservons que les requêtes correctes, c'est-à-dire qui correspondent à un document effectivement accessible (statut 2xx). D'autre part, nous éliminons les requêtes vers des images pour nous focaliser sur les documents. Enfin, nous découpons l'historique de chaque utilisateur en navigations. Une navigation est une suite de requêtes d'un utilisateur séparées au plus de 30 minutes. Notre analyse est entièrement basée sur les navigations et ne tient pas compte du fait que plusieurs navigations peuvent provenir d'un même utilisateur (au sens indiqué précédemment). Ceci réduit les problèmes liés aux caches Web (*proxy*), aux adresses IP dynamiques et au partage d'ordinateurs.

3.1.2 Prise en compte de la structure du site

Pour obtenir une vue d'ensemble de la perception du site par ses utilisateurs, nous devons simplifier la représentation des navigations selon deux directions. Tout d'abord, l'ordre précis dans lequel un internaute parcourt les sites étudiés ne nous semble pas pertinent pour une analyse d'ensemble. Nous supprimerons donc l'aspect temporel des navigations, à l'image de (Mobasher et al., 2002). D'autre part, un site de taille important peut contenir des milliers de documents et il est peu probable de trouver de fortes ressemblances entre les navigations, excepté si on se focalise sur les motifs fréquents. Pour une analyse globale, il est donc nécessaire de regrouper les documents en classes d'éléments comparables.

Une solution simple, proposée dans (Fu et al., 2000), consiste à utiliser la structure hiérarchique des sites étudiés. Un URL est en effet organisé de façon hiérarchique : dans l'URL `http://www-sop.inria.fr/axis/Publications/`, on retrouve le serveur de l'unité de recherche de l'INRIA située à Sophia-Antipolis (`www-sop.inria.fr`), le projet de recherche AxIS (`axis`) et la liste de publications de ses membres (`Publications`). Pour simplifier l'analyse d'un ensemble de navigations, on peut donc remplacer les URLs des documents visités par une version "raccourcie" qui se base sur la structure

du site. Pour une vision de très haut niveau, la navigation de la table 1 pourra être

	URL
1	`http://www-sop.inria.fr/`
2	`http://www-sop.inria.fr/act_recherche/les_projets_fr.shtml`
3	`http://www.inria.fr/recherche/equipes/axis`
4	`http://www-sop.inria.fr/axis/`
5	`http://www-sop.inria.fr/axis/ra.html`
6	`http://www.inria.fr/rapportsactivite/RA2003/axis2003/axis_tf.html`

TAB. 1 – *Une navigation*

simplifiée par la représentation tabulaire de la table 2 si on se contente de deux niveaux dans l'arborescence des sites étudiés.

	Serveur	Niveau 1	Niveau 2
1	`www-sop.inria.fr`		
2	`www-sop.inria.fr`	`act_recherche`	`les_projets_fr.shtml`
3	`www.inria.fr`	`recherche`	`equipes`
4	`www-sop.inria.fr`	`axis`	
5	`www-sop.inria.fr`	`axis`	`ra.html`
6	`www.inria.fr`	`rapportsactivite`	`RA2003`

TAB. 2 – *Représentation tabulaire d'une navigation*

En pratique, on détermine p groupes d'URLs à partir de l'arborescence du site, $(u_1, \ldots, u_p)$. Chaque groupe correspond à un site, puis à un début d'URL jusqu'à un certain niveau. Chaque navigation est alors représentée par un vecteur $(x_1, \ldots, x_p)$. La valeur x_i correspond au nombre de requêtes de la navigation dont l'URL commence comme l'URL u_i. On retrouve de type de représentation simplifiée dans de nombreux travaux de WUM qui cherchent à caractériser les utilisateurs d'un site, comme dans (Fu et al., 2000) par exemple.

Notons que d'autres méthodes de regroupement d'URLs sont envisageables, en travaillant par exemple sur le contenu des pages ou encore sur la structure d'hyperlien du site. Cependant, il est important de conserver une visualisation exploitable. L'avantage de la méthode proposée est que le schéma de simplification est très simple : le groupe d'URLs est décrit d'une façon facile à comprendre par un humain puisqu'il s'agit d'un simple élagage de l'arbre associé à l'URL.

3.2 Visualisation

Après pré-traitements et simplifications, les données d'usage sont donc représentées sous forme d'un tableau à p lignes (les groupes d'URLs) et n colonnes (les navigations). En pratique, p est relativement modeste (une centaine de groupes d'URLs) pour une

visualisation aisée, alors que n peut être très grand (plusieurs dizaines voire centaines de milliers de navigation). Malgré le nombre très élevé de dimensions, il peut être tentant de travailler directement sur le tableau de données, en appliquant des méthodes classiques de visualisation, comme l'analyse en composantes principales (ACP). Cependant, comme nous le verrons dans la section 4, les résultats obtenus sont très décevants, la dimension intrinsèque des données étant vraisemblablement très élevée. De plus, nous n'observons pas ici une limitation de l'ACP mais bien un problème lié à la dimension des données car l'utilisation d'un algorithme de projection non linéaire comme Isomap (Tenenbaum et al., 2000) n'améliore en rien la qualité de la visualisation obtenue.

Nous proposons donc de visualiser les données de la façon suivante :

1. le tableau de données est transformé en un tableau de dissimilarités entre les groupes d'URLs
2. une méthode projection non linéaire est appliquée au tableau pour visualiser les dispositions relatives des groupes URLs
3. l'arbre de recouvrement minimal induit par le tableau de dissimilarités est tracé par un algorithme de représentation de graphes

Nous combinons donc deux visualisations concurrentes afin de mieux comprendre les données.

Comme l'utilisation de la distance euclidienne entre les groupes d'URLs dans l'espace des navigations ne donne pas des résultats satisfaisant, il est naturel d'utiliser une autre dissimilarité. Parmi les très nombreuses dissimilarités ont été proposées pour comparer des données de comptages, nous avons retenu l'indice de Jaccard (préconisé par (Foss et al., 2001) dans le cadre du WUM) car il ne tient pas compte du nombre de pages vues dans un groupe d'URLs. En ce sens, il favorise donc les rapprochements, ce qui est important dans les grands sites pour lesquels les groupes d'URLs sont souvent très isolés (cf la section 4).

La deuxième étape de la visualisation consiste à réaliser une représentation en deux (ou trois) dimensions du tableau de dissimilarités. Nous utilisons le *Multi Dimensional Scaling* (MDS) classique (Torgerson, 1952).

La visualisation ainsi obtenue est parfois trompeuse car les données de très grandes dimensions sont difficiles à projeter en deux ou trois dimensions si elles ne possèdent pas une dimension intrinsèque faible. Nous associons donc à la projection basée sur le respect de l'ensemble des dissimilarités, une représentation basée au contraire sur la conservation du minimum de la structure de relation. Nous construisons en effet l'arbre couvrant minimal associé à la matrice de dissimilarités et nous visualisons cet arbre grâce à un algorithme de représentation de graphes assez classique (Fruchterman and Reingold, 1991). Nous utilisons le programme `fdp` du logiciel libre graphviz (Ellson et al., 2003).

4 Application

4.1 Les données

Dans cette section, nous mettons en oeuvre la méthode proposée sur une partie du site Web de l'Institut National de Recherche en Informatique et Automatique (INRIA).

Le site de l'INRIA est reparti en plusieurs serveurs dont les rôles sont différents. Le site principal, `www.inria.fr` présente l'institut dans son ensemble, assure la diffusion des rapports de recherche, la promotion de l'institut, etc. Les Unités de Recherche (UR) qui correspondent grossièrement aux différentes implantations géographiques de l'INRIA possèdent aussi des serveurs. Nous nous sommes intéressés au serveur de l'UR de Sophia Antipolis, `www-sop.inria.fr`, et à celui de l'UR Futurs, `www-futurs.inria.fr` (qui correspond à plusieurs implantations géographiques). Comme l'illustre la navigation de la table 1, les différents serveurs de l'INRIA sont étroitement liés et le passage de l'un d'entres eux à un autre se fait de façon totalement transparente pour l'utilisateur. Une analyse multi-serveurs est donc indispensable dans ce contexte.

Nous étudions les accès effectués sur les serveurs pendant les 15 premiers jours de l'année 2003. Nous observons pendant cette période un total de 446 014 requêtes correctes (statut 2xx). Nous appliquons la simplification proposée à la section 3.1.2 en ne conservant que le serveur et le niveau 1 de l'URL, ce qui nous donne 136 groupes d'URLs. Les requêtes sont regroupées en 152 826 navigations. Nous ne retenons que les navigations qui comportent entre 5 et 400 requêtes, ce qui réduit le nombre de requêtes considérées à 199 173 et le nombre de navigations à 16 717. Après ce premier filtrage, nous ne conservons que 107 groupes d'URLs, ceux dont les URLs ont été visités par au moins 5 navigations différentes, ce qui réduit le nombre de requêtes à 199 096.

Serveur	Groupes d'URLs	Requêtes	Navigations
`www.inria.fr`	25	115 155	11 159
`www-sop.inria.fr`	77	83 880	8 933
`www-futurs.inria.fr`	5	61	18

TAB. 3 – *Statistiques de visite des serveurs*

On constate sur la table 3 que les serveurs n'ont pas une fréquentation équilibrée. La très faible fréquentation du serveur de Futurs s'explique par le fait que l'UR a été créée au début de l'année 2002 et n'a démarré véritablement qu'en 2003. La somme des navigations est supérieure à 16 717 car les serveurs sont fortement liés entre eux et beaucoup de navigations concernent plusieurs sites. On compte en effet 3 375 navigations qui contiennent des URLs de `www.inria.fr` et de `www-sop.inria.fr`, 10 concernant Sophia et Futurs, et 17 concernant Futurs et le site principal. 9 navigations sont passées par les trois sites.

De plus, comme dans tout grand site, les groupes d'URLs sont assez isolés. 40 % des navigations (6 739) ne concernent qu'un seul groupe d'URLs. On dénombre seulement 2 516 navigations (15 %) passant par au moins 5 groupes d'URLs. Il est donc particulièrement important d'utiliser une dissimilarité qui fait ressortir les points communs entre les navigations

4.2 Analyses classiques

Une ACP réalisée sur le tableau $107 \times 16\,717$ ne donne pas des résultats très utiles. Les deux premières composantes n'expliquent que 37 % de la variance, et il

faut retenir 22 composantes avant d'atteindre 80 % de variance expliquée. En fait l'ACP est dominée par l'effet taille, puisque qu'elle fait ressortir les groupes d'URLs `www.inria.fr/rapportsactivite` et `www.inria.fr/travailler` qui sont les plus visités (44 728 requêtes pour le premier et 15 919 pour le second). Le seul point qui ressort clairement de la représentation par ACP est que le serveur `www.inria.fr` est beaucoup plus visité que les autres. L'utilisation de Isomap (Tenenbaum et al., 2000) n'améliore que marginalement la qualité de la visualisation qui reste dominée par le poids relatif des différents groupes d'URLs. De plus, une analyse des correspondances, qui permet de s'affranchir des problèmes liés aux effectifs, n'améliore pas sensiblement la qualité de la visualisation. Il semble donc que la dimension intrinsèque des données (étudiées avec la métrique euclidienne ou celle du χ^2) soit élevée et peu propice à une visualisation simple.

4.3 Projection

La figure 1 représente le résultat du MDS appliqué aux données d'usage comparées selon l'indice de Jaccard. On constate une assez claire séparation entre les trois serveurs, avec un point de contact (au centre) entre `www.inria.fr` et `www-sop.inria.fr`. Les groupes d'URLs correspondant à la présentation générale de l'UR de Sophia sont

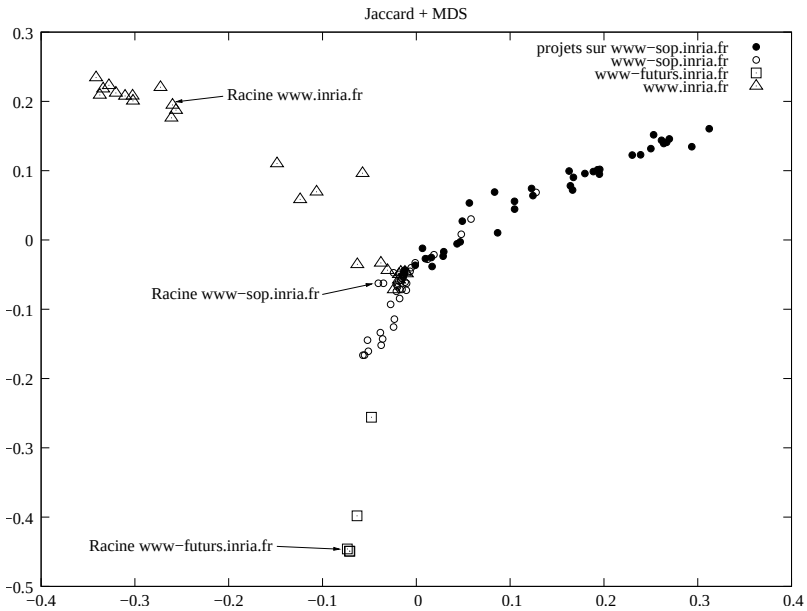

FIG. 1 – *Les serveurs de l'INRIA*

localisées au centre du graphique, à proximité de la racine de `www-sop.inria.fr`. La branche de droite contient presque exclusivement des groupes d'URLs qui décrivent des projets de recherche associés à cette UR.

En dehors de cette vue d'ensemble qui fait ressortir les grands groupes d'URLs, on peut se focaliser sur certaines zones de l'affichage. La figure 2 correspond à un zoom sur la partie `www.inria.fr`. On identifie aisément une zone institutionnelle proche de la racine du site, en particulier les groupes d'URLs `www.inria.fr/presse`, `www.inria.fr/valorisation`, `www.inria.fr/acualites`, etc. Le groupe `www.inria.fr/rapportsactivite` est plus au centre car le serveur principal héberge tous les rapports d'activités des projets, même si ceux-ci dépendent de l'UR de Sophia, par exemple. Il est donc logique que le groupe d'URLs concerné soit plus proche des groupes du serveur `www-sop.inria.fr`.

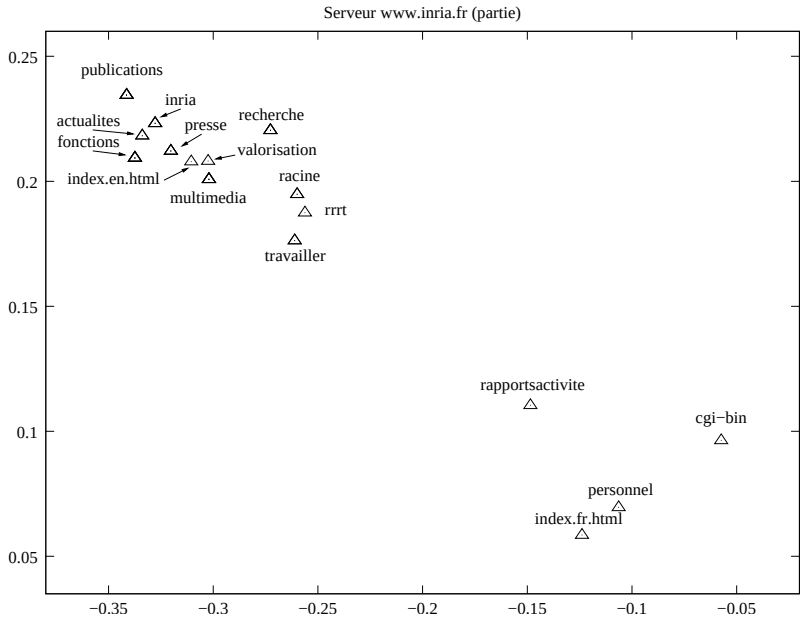

FIG. 2 – *Une partie du serveur* `www.inria.fr`

4.4 Arbre couvrant minimal

Bien que la vue d'ensemble proposée par le MDS soit intéressante, elle n'est pas toujours très précise et peut induire en erreur car les données d'origine ont une dimension intrinsèque élevée. Pour compléter la vue d'ensemble, on utilise une représentation de l'arbre couvrant minimal qui fait ressortir la structure de voisinage entre les groupes d'URLs (figure 3).

Plusieurs éléments ressortent clairement de cette représentation. On constate par exemple qu'un groupe de pages internes (majoritairement sur le serveur `www-sop.inria.fr`) se détache tout en étant lié au serveur concerné. Il s'agit vraisemblablement de navigations vers des services destinés au personnel de l'INRIA. Les URLs concernées apparaissent aussi de façon regroupées au centre de la figure 1 (après un zoom), mais pas

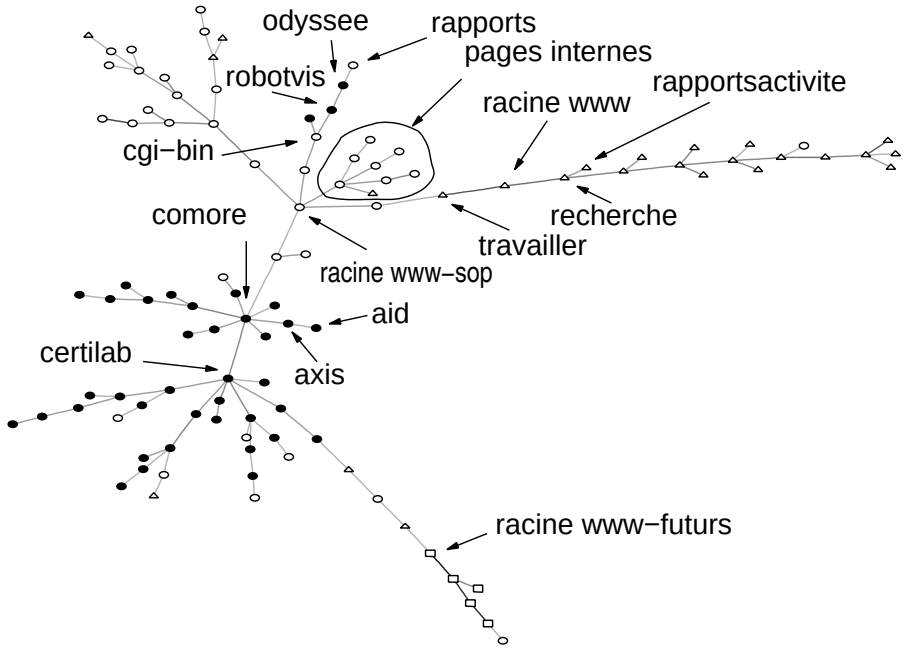

FIG. 3 – *Arbre couvrant minimal*

de façon isolée comme ici.

On remarque aussi que la racine du serveur central (`www.inria.fr`) joue bien son rôle puisqu'elle est liée au groupe d'URLs `travailler` (offres d'emplois et concours) et au groupe `recherche`, lui-même lié aux rapports d'activités (`rapportsactivite`). L'arbre confirme la séparation entre les trois serveurs, ainsi que celle qui existe entre les pages des projets de recherche et le reste du site. On remarque que le projet AxIS est relié au projet aid, ce qui est parfaitement normal : le second a disparu au profit du premier.

Un autre exemple de l'intérêt de cette représentation apparaît dans l'étude de la branche partant de `www-sop.inria.fr` vers `cgi-bin`, `robotvis`, etc. Le projet Robotvis est celui qui engendre le plus d'accès dans les logs étudiés. Il est naturellement lié au projet Odyssée (qui le remplace maintenant). De plus, le site de Robotvis contient de très nombreuses démonstrations en ligne des algorithmes développés par le projet. Ces algorithmes sont implémentés sous forme d'extension sur le serveur de l'UR de Sophia, ce qui engendre de très nombreuses requêtes vers le groupe d'URLs `cgi-bin`. D'autre part, les pages du projet Odyssée contiennent de très nombreux liens vers des rapports de recherches hébergés dans le groupe d'URLs `www-sop.inria.fr/rapports`. Tous ces rapprochements sont invisibles sur la représentation par MDS.

Par contre, le fait de ne conserver que quelques liens donne parfois une impression trompeuse. Les projets Comore et Certilab occupent par exemple des positions centrales dans l'arbre couvrant minimal, alors qu'ils sont assez peu visités (Certilab est le 2-

ième projet le moins visité et Comore 7-ième le moins visité parmi 39 projets). Ces projets apparaissent vraisemblablement dans des navigations de type exploratoire qui balayent un large ensemble de projets : ils sont relativement proches des autres projets. Au contraire, le projet Robotvis, qui est le plus visité, est plutôt éloigné des autres projets car il engendre beaucoup de visites spécifiques. Ces éléments se retrouvent sur la représentation par MDS (Comore et Certilab sont les points les plus extrêmes du nuage représentant les projets, alors que Robotvis est assez central), mais pas dans l'arbre couvrant minimal. Les deux visualisations sont donc complémentaires.

5 Conclusion

La combinaison d'un regroupement simple des URLs d'un site, d'une dissimilarité adaptée à la comparaison de navigations et d'une représentation double par MDS et arbre couvrant minimal, permet de visualiser efficacement les données d'usage d'un site Web. Les outils proposés doivent maintenant être validés auprès de concepteurs et d'animateurs de sites pour montrer qu'ils permettent de confronter la vision éditoriale avec celle des internautes et comprendre les modes d'utilisations du site.

Références

Benford, S., Taylor, I., Brailsford, D., Koleva, B., Craven, M., Fraser, M., Reynard, G., Greenhalgh, C., December 1999. Three dimensional visualisation of the world wide web. ACM Computing Surveys 31 (4es).

Berners-Lee, T., Fielding, R., Masinter, L., August 1998. Uniform Resource Identifiers (URI) : Generic Syntax. RFC 2396, The Internet Society, `http://www.ietf.org/rfc/rfc2396.txt`.

Cadez, I., Heckerman, D., Meek, C., Smyth, P., White, S., 2000. Visualization of navigation patterns on a web site using model based clustering. In : Proceedings of the KDD 2000. pp. 280–284.

Chen, C., 1998. Generalized similarity analysis and pathfinder network scaling. Interacting with Computers 10, 107–128.

Chen, C., Paul, R. J., 2001. Visualizing a knowledge domain's intellectual structure. IEEE Computer 34 (3), 65–71.

Chen, J., Sun, L., Zaiane, O. R., Goebel, R., 2004. Visualizing and discovering web navigational patterns. In : Proceedings of the 7th International Workshop on the Web and Databases : colocated with ACM SIGMOD/PODS 2004. Paris, France, pp. 13–18.

Chi, E. H., March 2002. Improving web usability through visualization. IEEE Internet Computing , 64–71.

Chi, E. H., Pitkow, J., Mackinlay, J., Pirolli, P., Grossweiler, R., Card, S. K., 1998. Visualizing the evolution of web ecologies. In : Proc. of ACM CHI 98 Conference on Human Factors in Computing Systems. ACM Press, Los Angeles, California, pp. 400–407, 644–645.

Cugini., J., Scholtz, J., September 1999. VISVIP : 3D visualization of paths through web sites. In : Proceedings of the International Workshop on Web-Based Information Visualization (WebVis'99) (in conjunction with DEXA'99, Tenth International Workshop on Database and Expert Systems Applications, eds A.M. Tjoa, A. Cammelli, R.R. Wagner). IEEE Computer Society, Florence, Italy, pp. 259–263.

Dodge, M., May 2001. Mapping how people use a website. Mappa.Mundi Magazine `http://mappa.mundi.net/maps/maps_022/`.

Dodge, M., 2004. An atlas of cyberspaces. `http://www.cybergeography.org/atlas/atlas.html`.

Ellson, J., Gansner, E., Koutsofios, E., North, S., Woodhull, G., 2003. Graphviz and dynagraph – static and dynamic graph drawing tools. In : Junger, M., Mutzel, P. (Eds.), Graph Drawing Software. Springer-Verlag, pp. 127–148, `http://www.graphviz.org`.

Foss, A., Wang, W., Zaïane, O. R., April 2001. A non-parametric approach to web log analysis. In : Proc. of Workshop on Web Mining in First International SIAM Conference on Data Mining (SDM2001). Chicago, IL, pp. 41–50.

Fruchterman, T. M., Reingold, E. M., 1991. Graph drawing by force-directed placement. Software - Practice and Experience 21 (11), 1129–1164.

Fu, Y., Sandhu, K., Shih, M.-Y., 2000. A generalization-based approach to clustering of web usage sessions. In : Masand, Spiliopoulou (Eds.), Web Usage Analysis and User Profiling. Vol. 1836 of Lecture Notes in Artificial Intelligence. Springer.

Group, W. H. W., August 2002. XHTML 1.0 the Extensible HyperText Markup Language. W3C recommandation, W3C, second Edition. `http://www.w3.org/TR/xhtml1/`.

Luotonen, A., 1995. The common logfile format. `http://www.w3.org/pub/WWW/Daemon/User/Config/Logging.html`.

Mobasher, B., Cooley, R., Srivastava, J., August 2000. Automatic personalization based on web usage mining. Communication of ACM 43 (8), 142–151.

Mobasher, B., Dai, H., Luo, T., Nakagawa, M., January 2002. Discovery and evaluation of aggregate usage profiles for web personalization. Data Mining and Knowledge Discovery 6 (1), 61–82.

Noel, S., Chu, C.-H. H., Raghavan, V., 2003. Co-citation count versus correlation for influence network visualization. nformation Visualization, 2 (3).

Raggett, D., Le Hors, A., Jacobs, I., December 1999. HTML 4.01 specification. W3C recommandation, W3C, `http://www.w3.org/TR/html4/`.

Srivastava, J., Cooley, R., Deshpande, M., Tan, P.-N., 2000. Web usage mining : Discovery and applications of usage patterns from web data. SIGKDD Explorations 1 (2), 12–23.

Tanasa, D., Trousse, B., March-April 2004a. Advanced data preprocessing for intersites web usage mining. IEEE Intelligent Systems 19 (2), 59–65.

Tanasa, D., Trousse, B., August-September 2004b. Data preprocessing for wum. IEEE Potentials 23 (3), 22–25.

Tenenbaum, J. B., de Silva, V., Langford, J. C., December 2000. A global geometric framework for nonlinear dimensionality reduction. Science 290 (5500), 2319–2323.

Torgerson, W. S., 1952. Multidimensional scaling : I. theory and method. Psychometrika 17, 401–419.

Youssefi, A. H., Duke, J. D., Zaki, M. J., Glinert, E. P., May 2004. Visual web mining. In : Proc. of the13th International World Wide Web Conference. New York, NY.

Summary

This article introduces a visualization method for web usage mining that enables to confront the logical and semantical organization of a web site with the perception and understanding of this organization by the users. The method is based on user trajectories reconstruction from the log files produced to the web site servers. The logical and hierarchical organization of the site is used to simplify trajectory representation, especially by removing the temporal aspect. Simplify trajectories are used to calculate dissimilarities between URL groups defined thanks to the set hierarchy. URL groups are then projected in two dimensions thanks to the Multi Dimensional Scaling algorithm. This visualization is associated to a complementary representation of the minimum spanning tree induced by the dissimilarity matrix. In order to demonstrate its practical interest the method is applied to real world data : the INRIA web site.