Comparaison de dissimilarités pour l'analyse de l'usage d'un site web

Fabrice Rossi*, Francisco De Carvalho**, Yves Lechevallier*, Alzennyr Da Silva*,**

*Projet AxIS, INRIA Rocquencourt
Domaine de Voluceau, Rocquencourt, B.P. 105, 78153 Le Chesnay Cedex – France
**Centro de Informatica - CIn/UFPE
Caixa Postal 7851, CEP 50732-970, Recife (PE) – Brésil

Résumé. L'obtention d'une classification des pages d'un site web en fonction des navigations extraites des fichiers "logs" du serveur peut s'avérer très utile pour évaluer l'adéquation entre la structure du site et l'attente des utilisateurs. On construit une telle typologie en s'appuyant une mesure de dissimilarité entre les pages, définie à partir des navigations. Le choix de la mesure la plus appropriée à l'analyse du site est donc fondamental. Dans cet article, nous présentons un site de petite taille dont les pages sont classées en catégories sémantiques par un expert. Nous confrontons ce classement aux partitions obtenues à partir de diverses dissimilarités afin d'en étudier les avantages et inconvénients.

1 Introduction

La conception, la réalisation et la maintenance d'un site web volumineux sont des tâches difficiles, en particulier quand le site est écrit par plusieurs rédacteurs. Pour améliorer le site, il est alors important d'analyser les comportements de ses utilisateurs, afin de découvrir notamment les incohérences entre sa structure *a priori* et les schémas d'utilisation dominants. Les utilisateurs contournent en effet souvent les limitations du site en navigant (parfois laborieusement) entre les parties qui les intéressent, alors que celles-ci ne sont pas directement liées aux yeux des concepteurs. A l'opposée, certains liens sont très peu utilisés et ne font qu'encombrer la structure hyper textuelle du site.

Une méthode d'analyse dirigée par l'usage consiste à réaliser une classification du contenu du site à partir des navigations enregistrées dans les logs du serveur. Les classes ainsi obtenues sont constituées de pages qui ont tendance à être visitées ensembles. Elles traduisent donc les préférences des utilisateurs. La principale difficulté de cette approche réside dans la nature des observations (les navigations). Comme celles-ci sont de taille variable, on peut en déduire de nombreuses mesures de dissimilarité entre les pages visitées, selon qu'on tient compte de la durée de la visite, du nombre de fois que la page est vue, etc. Dans le contexte de la classification, il est alors difficile de choisir *a priori* quelle mesure de dissimilarité est la plus adaptée à l'analyse du site.

Dans cet article, nous étudions un site web peu volumineux (91 pages), très bien structuré, et au contenu sémantique bien défini. Grâce à cet exemple de référence, nous comparons différentes dissimilarités afin de mesurer leur aptitude à révéler ce contenu sémantique.

2 Données d'usage

2.1 Préparation des données

Les données d'usage d'un site Web proviennent essentiellement des fichiers log des serveurs concernés. Chaque ligne du fichier log décrit une requête reçue par le serveur associé : elle indique ainsi le document demandé, la provenance de la requête, la date de la demande, etc. Diverses techniques de pré-traitement permettent d'extraire des logs des *navigations*, comme par exemple celles de Tanasa et Trousse (2004), utilisées dans le présent article. Une navigation est une suite de requêtes provenant d'un même utilisateur et séparées au plus de 30 minutes. Elle constitue donc la trajectoire d'un utilisateur sur le site. Nous nous contenterons dans cet article de considérer chaque navigation comme la liste ordonnée des pages demandées par un utilisateur, sans chercher à tenir compte du temps passé sur chaque page.

2.2 Analyse du contenu du site

L'un des buts de l'analyse de l'usage est d'améliorer le site web étudié. Pour ce faire, il est important de déterminer si les *a priori* sémantiques des concepteurs sont validés par les usages. On cherche en particulier à savoir si les catégorisations choisies *a priori* sont perçues comme telles par les utilisateurs. Pour ce faire, on construit à partir des données d'usage une mesure de (dis)similarités entre les pages du site. Les pages qui sont souvent ensembles dans les navigations seront ainsi considérées comme proches. En classant le contenu du site selon une dissimilarité induite par l'usage, on détermine des groupes de pages liées qu'on peut confronter aux hypothèses des concepteurs du site.

Toute la difficulté réside dans la nature des données. On peut en effet décrire chaque page par l'intermédiaire des navigations, sous forme de données complexes, en considérant les pages comme des individus et les navigations comme des variables. Une page p_i est ainsi décrite par les variables $n_1, \ldots, n_N$ (pour N navigations). La valeur de la variable n_k est l'ensemble des positions de la page p_i dans la navigation n_k. Si un site contient quatre pages, A, B, C et D, et est visité par la navigation $n_1 = (A, B, A, C, D)$, la variable n_1 vaut $\{1, 3\}$ pour la page A, $\{2\}$ pour la page B, $\{3\}$ pour la page C, et $\{4\}$ pour la page D.

Ces données sont difficiles à analyser car les variables n'ont pas des valeurs numériques mais plutôt ensemblistes. De plus, le nombre de colonnes du tableau peut être très élevé. En outre, si le nombre de lignes est élevé (site volumineux), le tableau est alors en général très creux (i.e., contient beaucoup d'ensembles vides).

2.3 Dissimilarités

De nombreuses solutions sont envisageables pour traiter ce type de tableau de données. La plus simple consiste à binariser les valeurs des variables, en remplaçant un ensemble vide par la valeur 0 et un ensemble non vide par un 1. De nombreux indices de (dis)similarités ont été définis pour de telles données binaires (cf e.g. Gower et Legendre (1986)). Nous retenons la dissimilarité basée sur l'indice de Jaccard, définie comme suit

$$d_J(p_i, p_j) = \frac{|\{k | n_{ik} \neq n_{jk}\}|}{|\{k | n_{ik} \neq 0 \text{ ou } n_{jk} \neq 0\}|}, \quad (1)$$

où n_{ik} vaut 1 si et seulement si la page p_i est visitée par la navigation k et où $|U|$ désigne le cardinal de l'ensemble U. Pour cette dissimilarité, deux pages sont proches dès que la plupart des navigations qui passent par l'une passent par l'autre.

Le défaut principal de la dissimilarité de Jaccard est qu'elle ne tient pas compte du nombre de passages par une page. Pour palier ce problème, on remplace le tableau binaire n_{ik} par un tableau d'entiers positifs m_{ik} : la valeur de m_{ik} est le nombre de passages de la navigation k par la page i. On utilise ensuite une des nombreuses dissimilarités adaptées à ce type de données, comme par exemple la dissimilarité "cosinus" définie par

$$d_{\cos}(p_i, p_j) = 1 - \frac{\sum_{k=1}^{N} m_{ik} m_{jk}}{\sqrt{\left(\sum_{k=1}^{N} m_{ik}^2\right)\left(\sum_{k=1}^{N} m_{jk}^2\right)}}, \tag{2}$$

où N désigne le nombre total de navigations. Une autre dissimilarité intéressante est donnée par le modèle tf$\times$idf, dans lequel on pondère une navigation en tenant compte à la fois du nombre de passages dans une page donnée, mais aussi de la longueur de la navigation : une navigation longue passe par beaucoup de pages et la similarité sémantique entre les pages n'est donc pas garantie, contrairement au cas des navigations courtes. La dissimilarité est donnée par

$$d_{\mathrm{tf}\times\mathrm{idf}}(p_i, p_j) = 1 - \sum_{k=1}^{N} w_{ik} w_{jk}, \text{ avec } w_{ik} = \frac{m_{ik} \log \frac{P}{P_k}}{\sqrt{\sum_{l=1}^{N} m_{il}^2 \log\left(\frac{P}{P_l}\right)^2}}, \tag{3}$$

où P désigne le nombre de pages et P_k le nombre pages distinctes par lesquelles la navigation k est passée (cf par exemple Chen (1998)).

3 Site de référence

3.1 Présentation

Pour comparer les similarités retenues, nous utilisons le site du CIn, le laboratoire de deux d'entre nous. Ce site est réalisé par l'intermédiaire d'un ensemble de servlets programmées en Java. Les URL des pages sont assez complexes (car elles correspondent à l'appel des servlets) et il est extrêmement peu probable qu'un utilisateur accède directement à une page en saisissant l'URL (une URL complète pour le site comprend une centaine de caractères). Nous supposons donc que les utilisateurs accèdent au site en entrant par la page principale, puis en suivant les liens proposés dans un menu contextuel situé à gauche de la page.

Le site est petit (91 pages) et très bien organisé, sous la forme d'un arbre de hauteur 5. L'information est essentiellement située dans les feuilles de l'arbre (75 pages) alors que les noeuds internes jouent le rôle de pivots. Le site est très dense en liens, car le menu contextuel comporte toujours la page principale, les 10 pages de premier niveau, ainsi que les parents et les soeurs de la page courante. On dénombre ainsi parfois plus de 20 liens dans le menu lui-même, ce qui ne facilite pas toujours la navigation.

Nous avons étudié les accès au site du 26 Juin 2002 au 26 Juin 2003 (le fichier de logs contient environ 2Go de données brutes). Après pré-traitement et nettoyage, ceci représente 113 784 navigations.

3.2 Sémantique de référence

Nous avons étudié le contenu du site et construit manuellement une partition des pages en 13 classes, allant des pages recensant les publications du CIn à celles destinées à l'inscription des étudiants en mastère. Pour comparer les dissimilarités, nous utilisons un algorithme de classification et nous comparons les classes obtenues aux classes *a priori*.

1	2	3	4	5	6	7
Publications	Recherche	Partenariats	Cycles 1 & 2	Objectifs	Présentation	Annuaire
8	9	10	11	12	13	
Équipe	Options	Archives	3ème cycle	Nouvelles	Divers	

4 Classifications

4.1 Algorithmes et critères

Pour comparer les dissimilarités présentées dans la section 2.3, nous produisons des classes homogènes de pages, puis nous comparons ces classes à celles issues de l'analyse experte du site de référence. Pour la classification, nous utilisons un algorithme de type nuées dynamiques applicable à un tableau de dissimilarités (cf Celeux et al. (1989)) et une classification hiérarchique basée sur le lien moyen. D'autres algorithmes sont bien entendu envisageables.

Pour analyser les résultats, nous utilisons deux critères. Pour une analyse classe par classe, nous étudions la F mesure de van Rijsbergen (1979) associée à chaque classe *a priori* : il s'agit de retrouver au mieux une classe experte dans l'ensemble de classes produites par un algorithme. Pour une analyse globale, nous utilisons l'indice de Rand corrigé (cf Hubert et Arabie (1985)) qui permet de comparer deux partitions. Pour les deux indices, une valeur de 0 correspond à une absence totale de correspondance entre la structure *a priori* et la structure obtenue, alors qu'une valeur de 1 indique une correspondance parfaite.

4.2 Nuées dynamiques

L'algorithme des nuées dynamiques demande de choisir *a priori* un nombre de classes. Pour limiter les effets de ce choix, nous étudions les partitions produites pour un nombre de classes allant de 2 à 20. Nous obtenons les résultats suivants :

Dissimilarité	Indice de Rand	Classes retrouvées	F mesure minimale
Jaccard	0.5698 (9 classes)	6	0.4444
Tf$\times$idf	0.5789 (16 classes)	7	0.5
Cosinus	0.3422 (16 classes)	4	0.3

Pour l'analyse globale (indice de Rand corrigé), nous indiquons la taille de la partition maximisant le critère. On constate que tf$\times$idf et Jaccard donnent des résultats assez proches (léger avantage pour la première) alors que cosinus obtient des résultats sensiblement moins bons.

Pour l'analyse fine, nous cherchons pour chaque classe *a priori* une classe correspondante (au sens de la F mesure) dans l'ensemble des classes produites en faisant varier la taille de la partition, toujours de 2 à 20. Nous indiquons le nombre de classes parfaitement retrouvées et la plus mauvaise F mesure pour les classes non retrouvées. La mesure tf$\times$idf apparaît comme la plus performante. Les classes retrouvées parfaitement par les autres dissimilarités le sont aussi

par tf×idf (qui retrouve les classes 3, 4, 5, 7, 8, 9 et 12). On constate cependant que les classes parfaitement retrouvées le sont dans des partitions différentes, ce qui explique les indices de Rand relativement mauvais, comparativement aux résultats classe par classe.

4.3 Classification hiérarchique

Nous reprenons l'analyse conduite pour les nuées dynamiques dans le cas de la classification hiérarchique. Nous faisons varier ici le nombre de classes *a posteriori*, en étudiant tous les niveaux de coupure possible dans le dendrogramme. Nous obtenons les résultats suivants :

Dissimilarité	Indice de Rand	Classes retrouvées	F mesure minimale
Jaccard	0.6757 (11 classes)	3	0.5
Tf×idf	0.4441 (15 classes)	3	0.4
Cosinus	0.2659 (11 classes)	5	0.4

Au niveau global, on constate une nette domination de Jaccard et une amélioration des résultats pour celle-ci. Le critère du lien moyen utilisé ici, ainsi que la structure hiérarchique, semble permettre une meilleure exploitation de la dissimilarité de Jaccard, alors que les résultats sont nettement dégradés pour les autres mesures.

Les résultats classes par classes sont plus difficiles à analyser et semblent ne pas dépendre de la mesure. Cependant, les "bonnes" performances de tf×idf et de cosinus correspondent à une bonne approximation des classes pour des niveaux de coupure très différents dans le dendrogramme : il n'est donc pas possible d'obtenir avec ces mesures, une bonne récupération de l'ensemble des classes, alors que Jaccard se comporte globalement mieux.

5 Discussion et conclusion

La dissimilarité de Jaccard apparaît globalement comme la plus performante pour retrouver la sémantique *a priori* du site de référence, à condition d'être utilisée avec une classification hiérarchique. Tf×idf donne des résultats satisfaisants alors que cosinus semble incapable de retrouver les classes. Il est cependant important de confronter ces conclusions au contenu du site.

Il s'avère en effet que le site du CIn est un peu particulier à cause de sa très grande densité de liens, mais aussi en raison de la présence de pages de navigation, proches de "tables de matières". La classe 6 par exemple, contient 8 pages de description du laboratoire, alors que la classe 5 contient 6 pages présentant ses objectifs. Une des pages de la classe 5 est une description générale des objectifs, qui se contente pour l'essentiel d'orienter le lecteur vers les autres pages. En ce sens, elle pourrait faire partie de la classe 6. Cette page joue donc le rôle de pont entre la classe 6 et la classe 5. De plus la structure du site fait que le passage par cette page est obligatoire pour atteindre les autres pages de la classe 6. Ce problème est en fait assez généralisé dans le site, en raison de sa structure d'arbre. Il a des conséquences sur les dissimilarités car il force une certaine proximité entre des pages qui n'ont pas été placées dans une même classe *a priori*. La dissimilarité de Jaccard est particulièrement sensible à ce problème : elle a tendance, en particulier dans la classification hiérarchique, à regrouper par exemple les classes 5 et 6, excepté la page de présentation générale du CIn (qui est un pont entre la pagre principale du site et celles de la classe 6) et une page indique comme se rendre au CIn.

D'autre part, la fréquentation du site n'est pas uniforme, certaines pages étant très visitées d'autres peu. Or, les pages peu visitées le sont souvent dans des navigations longues qui ont un poids relatif plus faible que les courtes dans tf$\times$idf. Ceci explique en partie les problèmes rencontrés par cette dissimilarité, qui semble par contre plus immunisée que Jaccard aux problèmes induits par la structure du site (elle propose ainsi une séparation entre les classes 5 et 6, même si la classe 6 n'est pas parfaitement retrouvée).

Il semblerait donc intéressant d'explorer des variantes de tf$\times$idf (en particulier en terme de normalisation des pondérations) afin d'essayer de limiter la rétrogradation des navigations longues, ceci afin d'améliorer la sensibilité de la dissimilarité sur les pages peu fréquentées. Il serait aussi intéressant de tenir compte de la structure du site pour pondérer les données d'usage afin de minimiser l'impact de liens obligatoires : s'il n'existe qu'un chemin pour passer d'une page à une autre, le fait qu'il soit emprunté est moins pertinent que s'il existe plusieurs chemins.

Le cas de cosinus demande une investigation spécifique, puisque les mauvaises performances semblent être liées à la création d'une classe volumineuse sans aucune cohérence sémantique. Vue la nature de la dissimilarité, il est possible que ce problème soit une conséquence de la présence de plusieurs classes d'utilisateurs bien distinctes dont les navigations ne seraient pas du tout corrélées.

On note aussi une influence importante de l'algorithme de classification sur les résultats obtenus et même sur le classement des dissimilarités. Une analyse de l'adéquation entre le critère optimisé par l'algorithme et la dissimilarité utilisé semble donc important.

Références

Celeux, G., E. Diday, G. Govaert, Y. Lechevallier, et H. Ralambondrainy (1989). *Classification Automatique des Données*. Paris : Bordas.

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

Gower, J. et P. Legendre (1986). Metric and euclidean properties of dissimilarity coefficients. *Journal of Classification 3*, 5–48.

Hubert, L. et P. Arabie (1985). Comparing partitions. *Journal of Classification 2*, 193–218.

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

van Rijsbergen, C. J. (1979). *Information Retrieval* (second ed.). London : Butterworths.

Summary

Clustering the pages of a web site according to the browsing patterns identified in its server log files can be very useful for the analysis of the organization of the site and of its adequacy to user needs. Such a set of homogeneous classes is often obtained from a dissimilarity measure between the pages defined via the visits extracted from the logs. Choosing the right dissimilarity is therefore extremely important. This paper presents an analysis of different dissimilarity measures based on the comparison between the semantic structure of a benchmark site identified by experts and the clustering constructed with standard algorithms applied to the dissimilarity matrices generated by the chosen measures.