Mesure d'audience sur Internet par populations de fourmis artificielles

Nicolas Labroche

UPMC, LIP6, Pole IA, 8 rue du Capitaine Scott
75015 Paris
nicolas.labroche@lip6.fr
http://lofti.lip6.fr

Résumé. Nous présentons dans ce travail un outil pour la mesure d'audience sur Internet, reposant sur l'extraction de profils de navigation représentatifs de l'activité des internautes sur les sites. Ces profils sont obtenus par l'application d'un algorithme de classification non supervisée – inspiré du système de reconnaissance chimique des fourmis – sur des sessions de navigations construites à partir des fichiers log du site étudié. Cet algorithme de classification a été associé à une représentation multimodale des sessions utilisateurs permettant d'employer l'ensemble des informations à disposition dans les fichiers log (impacts sur les pages, heure de connexion, durée, séquence des pages, …), ainsi qu'à une mesure de similarité adaptée pour créer les profils de chacun des clusters obtenus. Il reste cependant d'autres modalités (basées sur le contenu des documents accédés) qui pourraient améliorer la capacité de l'outil à donner du sens aux profils découverts.

1 Introduction

La mesure d'audience sur Internet s'attache à extraire et à donner du sens aux navigations des internautes. Ses champs d'applications sont nombreux : personnalisation automatique des sites Web en fonction des pages accédées (Mobasher et al., 1999) ou encore recommandation dynamique de pages aux internautes en fonction de leur navigation passée (Yan et al., 1996). Dans ce cadre, la représentation des navigations des internautes est cruciale car elle détermine le type de recherche d'informations qui pourra être conduit par la suite.

Généralement, les informations de navigation des internautes sur un site Internet sont extraites d'un fichier log, présent sur le serveur Web, qui recense, pour simplifier et de manière idéale, l'ensemble des demandes de pages du site de la part des internautes. Ces requêtes clientes sont ensuite triées, filtrées et regroupées en sessions qui constituent pour chaque internaute, l'ensemble des informations issues de leur navigation sur un site à un moment donné (Cooley et al., 1999).

Plusieurs représentations des sessions ont été utilisées dans la littérature. Par exemple, le système WebMiner (Cooley et al., 1999) utilise un vecteur de transactions qui indique pour chacune des pages du site si elle a été accédée au moins une fois durant la session de façon à extraire des règles d'association. Dans Masseglia et al. (1999), les auteurs conservent les dates d'accès à chacune des pages pour extraire des règles séquentielles. D'autres représentations ont été utilisées comme la durée de visite ou le nombre d'impacts par page

comme dans les travaux de Yan et al. (1996) en raison de leur facilité de mise en œuvre dans les calculs de similarité des algorithmes de classification notamment. Enfin dans Heer et Chi (2001), les auteurs décrivent une session comme la somme de vecteurs représentatifs des pages visitées. Chaque « vecteur de page » est composé de sous vecteurs concaténés, chacun correspondant à une modalité comme le contenu de la page, les liens entrants ou sortants, …

Nous présentons dans ce travail une représentation multimodale des sessions de navigation des internautes reposant sur l'utilisation conjointe de différentes sources d'informations (modalités), issues du fichier log comme les impacts par page, la séquence des pages visitées, les dates d'accès aux pages, la durée totale de la session, l'adresse IP et l'identifiant de l'internaute, … L'intérêt de l'approche est de pouvoir moduler l'influence de chacune de ces modalités lors du calcul de la similarité entre deux sessions. Cette représentation multimodale associée à une mesure de similarité adaptée a été utilisée avec l'algorithme de classification non-supervisée AntClust sur un fichier log du site Internet d'une formation universitaire en informatique pour en extraire des comportements de navigation types.

Cet article s'articule donc comme suit : la section 2 décrit l'ensemble des modalités utilisées dans notre représentation ainsi que les mesures de similarité nécessaires à leur utilisation. La section 3 présente rapidement l'algorithme de classification AntClust, le fichier log de test ainsi que les types d'expérimentations dont il a fait l'objet. La section 4 décrit les résultats obtenus et la section 5 conclut en discutant de la nécessité d'adjoindre de nouvelles modalités relatives au contenu des documents accédés pour améliorer la compréhension des profils de navigation obtenus.

2 Représentation multimodale des sessions

La navigation de chaque internaute est décrite par un vecteur composé de m modalités pondérées. Dans le cas général, chaque modalité possède son propre type t (vecteur numérique, date, adresse IP, …). Nous exprimons donc la mesure globale de similarité $\Delta(s^1, s^2)$ entre deux sessions s^1 et s^2 comme étant la somme pondérée des mesures de similarité associées au type de chacune des m modalités qui les composent. On a donc :

$$\Delta(s^1, s^2) = \frac{\sum_{\forall i \in [1,m]} \omega_i \times \Delta_{t(i)}\left(s^1(i), s^2(i)\right)}{m \times \sum_{\forall i \in [1,m]} \omega_i}$$

Avec : $\Delta_{t(i)}\left(s^1(i), s^2(i)\right)$ la mesure de similarité entre les i^{èmes} modalités de type $t(i)$ des sessions s^1 et s^2 et ω_i le poids associé à cette modalité.

L'adresse IP et l'identifiant de l'internaute sont représentés par une chaîne de caractères : la mesure de similarité est ramenée à un simple test d'égalité entre les valeurs comparées.

La date et l'heure de connexion sont traduites en un nombre de secondes, puis normalisées en fonction de la première et de la dernière date apparaissant dans le fichier log. La mesure de similarité entre deux « dates » t_1 et t_2 devient donc :

$$\Delta_{Tps}(t_1, t_2) = 1 - |t_1 - t_2|$$

La même mesure de similarité est utilisée pour comparer les durées normalisées de deux sessions. La modalité de durée n'est toutefois pas préconisée, le fichier log ne permettant pas d'estimer le temps passé sur la dernière page visitée (symboliquement fixé à 0 seconde).

L'historique des pages visitées a été modélisé comme une séquence de pages du site. On recherche la plus longue sous séquence commune de pages entre les deux historiques h_1 et h_2 notée $Plsc(h_1,h_2)$ (Cormen et al., 1994) pour définir la mesure de similarité suivante :

$$\Delta_{Hist}(h_1,h_2) = |Plsc(h_1,h_2)| / \max(|h_1|,|h_2|)$$

Le vecteur de transactions est un vecteur de taille p indiquant pour chacune des pages du site si elle a été accédée au moins une fois durant la session. La mesure de similarité $\Delta_{Trans}(t_1,t_2)$ entre deux vecteurs de transactions t_1 et t_2 est égale à la proportion de composantes égales entre les deux vecteurs :

$$\Delta_{Trans}(t_1,t_2) = \frac{1}{p} \times \sum_{i=1}^{p} \delta_{t_1(i)=t_2(i)} \text{ , avec la fonction unité } \delta_{c_1=c_2} = \begin{cases} 1 & \text{si } c_1 = c_2 \\ 0 & \text{sinon} \end{cases}$$

Les vecteurs des impacts et des durées par page sont des vecteurs normalisés ayant également p composantes et qui indiquent pour chacune des pages du site le nombre de fois qu'elles ont été accédées ou le temps écoulé dessus. Nous avons utilisé la mesure de similarité suivante, inspirée d'une mesure de Minkowsky à l'ordre 2 entre deux vecteurs v_1 et v_2 :

$$\Delta_{Vec}(v_1,v_2) = 1 - \left(\sum_{i=1}^{p} |v_1(i) - v_2(i)|^2 / p \right)^2$$

3 Expérimentations

Nous présentons l'algorithme de classification AntClust utilisé pour créer les clusters de sessions desquels les statistiques de navigations sont extraites. Nous décrivons également le jeu de données employé ainsi que les détails des expérimentations dont il a fait l'objet.

3.1 L'algorithme de classification AntClust

AntClust (Labroche et al., 2004) est un algorithme de classification inspiré du système de reconnaissance chimique des fourmis. Dans celui-ci, chaque fourmi possède une odeur propre appelée label, partiellement définie par son génome, qui lui permet d'être acceptée au sein de son nid, ainsi qu'un modèle de reconnaissance de ce que doit être l'odeur d'un membre du nid, appelé template. Ces « odeurs » individuelles évoluent lors des rencontres entre fourmis et définissent à l'échelle du nid une odeur coloniale « moyenne ».

Similairement, dans notre modèle, le génome de chaque fourmi artificielle est associé à une session de navigation extraite du fichier log analysé. Chaque fourmi apprend ensuite un seuil d'acceptation (son template), défini par les similarités observées entre son génome et celui d'autres fourmis choisies aléatoirement. Chaque fourmi va ensuite réaliser des rencontres aléatoires de façon à déterminer le label (son appartenance à un nid ou cluster) qui correspond le mieux à son génome (une session) grâce à un ensemble de règles

comportementales individuelles. A la fin, les fourmis ayant des génomes similaires sont réparties dans les mêmes nids, ce qui forme la partition désirée des sessions de navigation.

3.2 Jeu de données et expérimentations conduites

Notre fichier log regroupe des requêtes soumises sur le site Web du DUAC informatique de l'Université de Tours. Ce dernier propose le contenu de la plupart des enseignements dispensés sous forme html et est fortement structuré avec peu de passerelles entre chacun des cours. L'architecture est simple et repose sur une page d'index servant de sommaire au reste du site ; le site n'est pas dynamique et une url ne renvoie qu'à un seul contenu, ce qui permettra, par la suite, d'inférer les motivations des internautes à partir des urls qu'ils auront accédées. Le fichier log a été enregistré sur une période d'environ 1 mois en octobre 2001. Le nombre de sessions reconstruites est de 1064 dont 667 sont uniques (i.e. proviennent d'une adresse IP qui n'apparaît qu'une fois dans le fichier log). Les impacts sont enregistrés sur 248 documents html différents. Si on ne conserve que les sessions ayant strictement plus d'un impact, leur nombre chute à 376.

Nous avons conduits principalement des expériences visant à étudier les partitions obtenues en fonction des pondérations données aux différentes modalités utilisées pour représenter les sessions : (1) poids de 1 pour toutes les modalités, (2) poids de 1 pour la modalité « impacts/page » (et 0 pour les autres) et enfin, (3) poids de 1 pour la modalité « transactions » (et 0 pour les autres).

4 Résultats

Les résultats des trois expérimentations sont représentés sous la forme d'un histogramme cumulé dont chacune des barres représente le « contenu » d'un groupe ou cluster. Ce contenu est exprimé en fonction de la proportion d'impact dans chacun des cours du site (/duac/java par exemple) pour l'ensemble des sessions de ce groupe de façon à donner du sens aux groupes découverts plus facilement.

Les résultats obtenus par l'ensemble des modalités (FIG 1) montrent que sur les 17 clusters découverts, certains auraient mérités d'être regroupés du fait de leur attachement à un même enseignement (par ex. le cours OMT, groupes 4, 15 et 17). La séparation de sessions ayant trait aux mêmes cours est probablement due à l'introduction des autres modalités (comme l'adresse IP, la date de début ou la durée de la session). Par ailleurs, certaines modalités sont redondantes (comme les impacts par page et le vecteur de transactions) et donnent ainsi trop de poids aux légères différences de fréquentation au sein d'un même enseignement entre deux sessions pourtant semblables.

L'utilisation de la modalité du nombre d'impacts par page (FIG 2) donne de bons résultats rapidement : les clusters sont biens séparés et même si plusieurs groupes s'intéressent au même enseignement, une étude plus approfondie des accès montre que ce ne sont pas les mêmes documents qui ont été demandés.

Enfin, l'utilisation du vecteur de transactions (FIG 3), pourtant très répandue, ne permet pas, dans notre cas de révéler une information de navigation puisque la très grande majorité des sessions est concentrée dans un seul cluster.

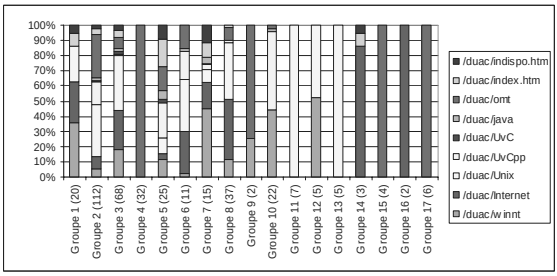

FIG. 1 – *Résultats de l'expérience 1 avec l'ensemble des modalités employées*

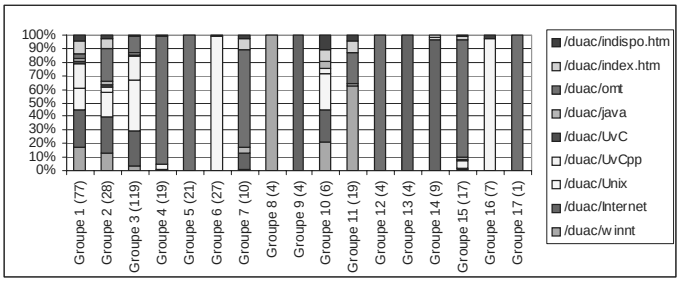

FIG. 2 – *Résultats de l'expérience 2 avec la modalité des impacts par page*

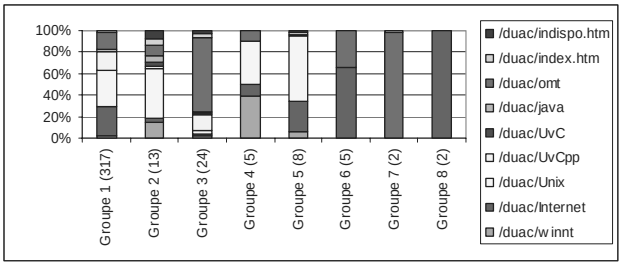

FIG. 3 – *Résultats de l'expérience 3 avec la modalité de vecteurs de transactions*

5 Conclusion et perspectives

Nous nous sommes intéressés dans ce travail à la mesure de l'audience sur les sites Internet et nous avons proposé à cet effet une nouvelle représentation multimodale de la navigation des internautes permettant l'utilisation conjointe et pondérée des différentes sources d'informations à disposition dans les fichiers log des serveurs Web. Nous avons associé cette modélisation utilisateur à un algorithme de classification non supervisée inspiré du système de reconnaissance chimique des fourmis nommé AntClust. Les expériences menées sur le site d'une formation informatique montrent toutefois que l'emploi de toutes les modalités n'est pas la meilleure approche à cause des temps de calculs rédhibitoires et de la perte de lisibilité des groupes trouvés (du fait des contradictions ou des redondances entre les modalités employées). On préfèrera donc une approche en deux temps : (1) extraction des profils de navigation, soit sous la forme de représentants de groupes de sessions soit sous la

forme de séquences typiques de pages visitées et (2) recherche du sens de ces profils. Dans le cas d'une approche par algorithme de classification, on se limitera aux modalités qui autorisent les calculs les plus rapides (comme la nombre d'impacts par page), en les couplant à un algorithme capable de déterminer automatiquement le nombre de groupes recherchés et de manipuler des volumes de données importants (comme l'algorithme BIRCH (Fu et al., 1999)(Zhang et al., 1996)). Pour l'extraction du sens des profils, il apparaît nécessaire de s'appuyer sur le contenu des documents (texte, images, …), sur la topologie du site étudié voire sur une ontologie des documents qui composent le site (obtenue par classification sémantique des documents). Enfin, il est envisageable d'utiliser des informations plus personnelles concernant l'internaute (via son historique notamment) qui permettraient de définir le contexte dans lequel il a navigué : niveau de connaissance sur le sujet, état d'esprit, mise en lumière de différences entre les navigations passées et actuelle sur un même site.

Références

Cooley R., Mobasher B. et Srivastava J. (1999), Data preparation for mining world wide web browsing patterns, Knowledge and Information Systems, Vol. 1, pages 5-32, 1999.

Cormen T., Leiserson C. et Rivest R. (1994), Introduction à l'algorithmique, pages 308-313, Dunod, 1994.

Fu Y., Sandhu K. et Shih M. (1999), Clustering of Web users based on access patterns, KDD Workshop on Web Mining, San Diego, CA, 1999.

Heer J. et Chi E. (2001), Identification of web user traffic composition using multi-modal clustering and information scent, Workshop on web mining, SIAM Conference on Data Mining, pages 51-58, Chicago IL, April 2001.

Labroche N., Guinot C. et Venturini G. (2004), Fast Unsupervised Clustering with Artificial Ants, Parallel Problem Solving from Nature Conference, Birmingham (UK), 2004.

Masseglia F., Poncelet P. et Cicchetti R. (1999), WebTool : An integrated framework for data mining, Database and Expert Systems Applications, pages 892-901, 1999.

Mobasher B., Cooley R. et Srivastava J. (1999), Creating adaptative Web sites through usage-based clustering of urls, 1999.

Yan T.W., Jacobsen M., Garcia-Molina H. et Dayal U. (1996), From user access patterns to dynamic hypertext linking, 5th WWW Conference, pages 1007-1014, 1996.

Zhang T., Ramakrishnan R. et Livny M. (1996), BIRCH an efficient data clustering method for very large databases, ACM SIGMOD International Conference on Management of Data, pages 103-114, Montreal, Canada, 1996.

Summary

In this work, we present a new Web usage mining tool that extracts navigation profiles representative of the users' activities on the Web sites. These profiles are computed by an unsupervised clustering algorithm – inspired by the chemical recognition system of ants – that is applied on Web navigation sessions from the Web servers log files. This clustering algorithm is associated with a multi-modal representation of the Web sessions and an adapted similarity measure to create the expected profiles from the Web sessions clusters. In conclusion, we discuss the interest of adding modalities related to the content of the documents accessed by the Web users to help explain the discovered navigation profiles.