Clusters d'ensembles de données larges dans le Web Log Mining

Gabriella Schoier et Giuseppe Melfi

Dipartimento di Scienze economiche e statistiche
Università di Trieste
Piazzale Europa, 1
I-34127, Trieste, Italie
Gabriella.Schoier@econ.units.it

Groupe de Statistique
Université de Neuchâtel
Espace de l'Europe, 4
CH-2002, Neuchâtel, Suisse
Giuseppe.Melfi@unine.ch

RÉSUMÉ. *Nous présentons une solution du problème de l'identification de clusters denses dans l'analyse de données concernant les accès à l'internet d'un ensemble d'utilisateurs. L'algorithme utilisé ici est une modification d'un algorithme proposé pour un problème de même nature concernant les reseaux sociaux.*

MOTS-CLÉS : *Data Mining, Web Log Mining, Clusters*

1. Introduction

Il y a une compétition intense entre les sociétés basés sur l'Internet pour acquérir de nouveaux clients et retenir les clients existants ; pour cette raison la personnalisation dans le web est devenue une partie indispensable de l'e-commerce. En particulier la personnalisation basée sur le Web Usage Mining ou le Web Log Mining, développée pour extraire des modèles intéressants dans les accès sur le Web a plusieurs avantages sur les techniques plus traditionnelles [SRI 00, MOB 02]. Considérons une série finie d'unités (les adresses I.P. des ordinateurs des utilisateurs) sur lesquelles deux variables relationelles ont été mesurés (ayant visité au moins M pages en commun ; étant resté le même intervalle de temps sur la même page) ; ceci forme un réseau N (la série d'unités et de relations associées) [WAS 94]. Dans le but d'analyser un tel réseau on peut considérer les résultats dérivant de deux théories de réseaux sociaux classiques : la théorie *small-world* [KOC 89] et la théorie *peer influence* [FRI 98]. La première a montré qu'il y a un haut degré de clustering local dans les réseaux, donc une approche pour étudier la structure de grands réseaux impliquerait l'identification de clusters locaux et l'analyse des relations à l'intérieur et entre les clusters. La seconde a prouvé que, en se basant sur un procédé d'influence endogène, les unités proches ont une tendance à converger sur des attitudes similaires et ainsi les clusters dans un réseau small-world doivent être similaires le long de multiples dimensions.

Dans ce papier nous présentons une solution au problème d'identification de clusters denses dans l'analyse des enregistrement d'accès au web, en considérant une modification d'un algorithme connu de l'analyse de réseaux sociaux [MOO 01]. L'avantage de cette approche est une structure réduite et plus flexible sur laquelle des techniques différentes telles que le blockmodelling [SCH 02] peuvent être utilisées. Nous comparons aussi les résultats de l'algorithme avec ceux de Batagelj et Mrvar [BAT 02a, BAT 02b] basée sur la méthode k-core.

2. Sur l'identification de clusters denses dans les données de Web Usage Mining

Le point de départ de l'analyse sont les fichiers d'enregitrement d'accès (web access logs) d'utilisateurs du site web réel, *www.girotondo.com*, un portail pour les enfants. Dans ce site il y a sept rubriques différentes : *Bacheca (Lanterne)*, *Corso (Cours)*, *Favolando (Fables)*, *Giochi (Jeux)*, *Links (Liens)*, *News (Nouvelles)*, *Percome (Comment)*, et il y a 362 pages de jhtml. La période d'observation est du 29/11/2000 au 18/01/2001.

Un tel fichier presente les données dans une forme brute. Dans la Table 1, un extrait est presenté.

Table 1 - Fichiers d'enregistrement d'accès

130.93.25.19	20/DEC/2000 :10 :19 :44+0100	"GET/mappa/01.jhtml HTTP/1.0"	200	2472	Mozilla/4.0
235.58.54.78	20/DEC/2000 :10 :19 :41+0100	"GET/news/archivio.jhtml HTTP/1.0"	200	115	Mozilla/4.0
267.12.83.56	20/DEC/2000 :10 :19 :40+0100	"GET/news/01/01/01.jhtml HTTP/1.0"	200	793	Mozilla/4.0
241.27.83.61	20/DEC/2000 :10 :19 :37+0100	"GET/favolando/01.jhtml HTTP/1.0"	200	88	Mozilla/4.0

Les fichiers d'enregistrement d'accès de serveur contiennent : le nom du domaine (ou l'adresse I.P.) de la demande ; la date et le temps de la demande ; la méthode de la demande (GET ou POST) ; l'URL de la page demandée ; le résultat de la demande (le succès, l'échec, l'erreur, etc.) ; la taille des données du fichier ; l'identification de l'agent client.

Une entrée dans le fichier des accès est automatiquement ajoutée à chaque fois qu'une demande pour une ressource atteint un serveur. Les enregistrements de fichiers contenant de l'information de n'importe quel objet (avec extension .gif, .jpeg, etc.) qui n'est pas une adresse internet est annulée pour obtenir ainsi un nouveau fichier. De cette façon nous avons un fichier indiquant l'adresse d'Internet pour chaque page visitée. Nous avons ensuite éliminé les pages visitées par moins de cinq adresses I.P. et ainsi 117 pages ont été considérées. Après le prétraitement un fichier de 1000 enregistrements a été utilisé. Les données consistent en une série d'adresses I.P. sur lesquelles deux variables relationelles (ayant visité au moins $M = 35$ pages en commun, étant resté le même intervalle de temps sur la même page pour des intervalles de temps fixés à l'avance) ont été mesurées ; pour chacune de ces deux variables les données sont représententées dans une matrice à deux modes (les adresses I.P. $\times$ pages). Les deux matrices sont changées en une matrice à un mode (I.P. $\times$ I.P.) en utilisant le programme UCINET [BOR 99] ; la matrice qui en résulte, appelée matrice des adiacences, est composée de zéros et de uns. Les coefficients de la matrice sont 1 si les utilisateurs (assimilés aux adresses I.P.) ont visité au moins 35 pages en commun et sont resté au moins 30 minutes dans les pages visitées en commun, 0 autrement (voir Table 2).

Table 2 - Matrice des adiacences

	138.222.202.11	151.15.169.130	151.2.15.154	.....
138.222.202.11	-	0	1	...
151.15.169.130	0	-	0	...
151.2.15.154	1	0	-	...
.....	...	...	...	

Maintenant nous introduisons une matrice $N \times m$ des influences, Y, où pour chacune des N adresses I.P. (lignes) corresponds un vecteur à m composantes qui décrit les influences. La matrice Y a autant de lignes que les adresses I.P. d'utilisateurs et un nombre de colonnes correspondant au nombres d'influences directes auxquelles chaque individu est sujet. Pour l'exemple qui suit nous avons décidé de utiliser une matrice Y à trois colonnes ($m = 3$). Ceci corresponds à assumer que chaque utilisateur peut avoir jusqu'à trois influences, en négligéant d'ultérieures éventuelles influences.

Pour construire cette matrice nous utilisons une version modifiée de l'algorithme de la moyenne du voisinage récursif (Recursive Neighbourhood Mean algorithm, RNM) proposé par Moody [MOO 01] et écrit en SAS. La modification, (Modified Recursive Neighbourhood Mean algorithm, MRNM) consiste dans le calcul pondéré de la moyenne après un certain nombres d'itérations et generalize l'algorithme RNM. L'algorithme peut être décrit comme suit :

1) Assigner à chaque adresse I.P. dans le réseau un nombre aléatoire uniforme entre 0 et 1 pour chacune des m variables. On obtient ainsi une matrice $Y^{(0)}$ $(N \times m)$ de nombres aléatoires.

2) La matrice $Y^{(t+1)}$ est définie par la formule

$$Y_{ik}^{(t+1)} = \frac{\sum_{j \in L_i} Y_{jk}^{(t)} N_{ij}}{\sum_{j \in L_i} N_{ij}} \qquad k = 1, \ldots, m, \qquad i = 1, \ldots, N,$$

où L_i est le sous-ensemble de $1, \ldots, N$ correspondant aux adresses I.P. qui sont en relation avec i, et N_{ij} est le nombre de pages en commun visitées par i et j.

3) Répéter le pas 2) n fois.

REMARQUE. — Pour $N_{ij} = 1$ pour tout $i, j = 1, \ldots, N$, l'algorithme ci-dessus corresponds à l'algorithme RNM classique [MOO 01].

Cette procédure demande dans l'input, la liste des adjacences, c'est-à-dire les paires de points entre lesquels une relation existe. A ce point l'algorithme RNM modifié est appliqué. Le résultat est la matrice Y.

Dans une situation idéale, $Y = \lim_{n \to \infty} Y^{(n)}$. Toutefois dans notre cas $n = 7$ a suffit pour obtenir des resultats tout à fait satisfaisant.

Sur les trois variables de position une "Ward's minimum variance cluster analysis" est exécuté. De telle façon nous obtenons un clustering clair qui révèle une structure de trois clusters entre les unités appartenant au réseau.

Table 3 - Table des résultats

I.P.	cluster	var1	var2	var3
1	1	0.48816	0.42557	0.53592
2	3	0.48822	0.42593	0.53589
3	1	0.48816	0.42557	0.53592
...	.....	.....	.....	.....

Le premier cluster, le plus nombreux est formé par les adresses I.P. qui ont une haute fréquence de relations ; le second est identifié par les adresses I.P. qui n'ont pas beaucoup de relations tandis que le tiers par les adresses I.P. qui ont peu de relations. Deux adresses I.P. ne sont classés nulle part, ce qui correspond à ce que l'on s'attendait car il s'agit des addresses I.P. du webmaster du site.

Pour la visualisation du réseau le programme PAJEK [BAT 02a] a été appliqué

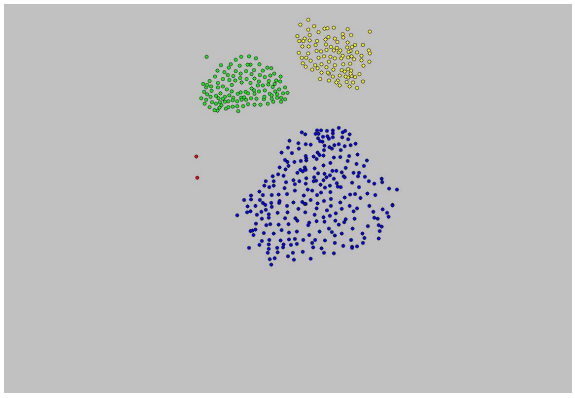

Fig. 1. Partition selon la méthode MRNM.

Si nous comparons les résultats de l'application de la procédure RNM modifiée avec la liste des adjacences, les adresses I.P. sont classés correctement.

Les résultats ont été comparés avec la méthode k-core [BAT 02b].

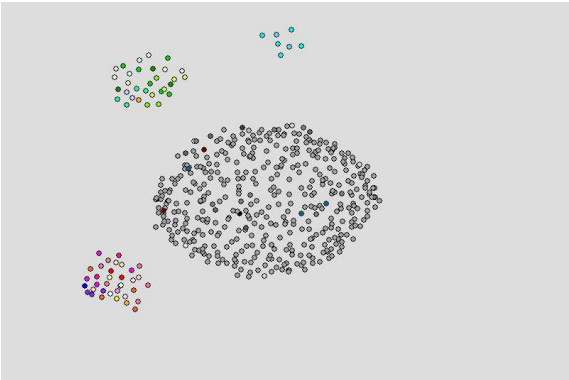

Fig. 2. Partition selon la méthode k-core.

Dans ce cas certains éléments ne sont pas correctement classés, car les deux éléments n'ayant pas de rélations avec les autres ne sont pas individués.

3. Conclusions

Dans cet article nous avons présenté une solution au problème d'identification de clusters denses dans l'analyse de fichiers d'enregistrement d'accès, en considérant une modification d'un algorithme connu de l'analyse de réseaux sociaux. Ainsi nous avons obtenu un outil pour étudier les clients dans les termes de leur comportement et leur information personnelle. Ceci permet l'accumulation d'éléments utiles pour l'amélioration de sites web et le développement de systèmes quand les séries de données sont grandes ou même énormes.

4. Bibliographie

[BAT 02a] BATAGELJ V., MRVAR A., *PAJEK : Program for large Network Analysis*, http ://www.vlado.fmf.uni-lj.si/pub/networks/pajek/, 2002.

[BAT 02b] BATAGELJ V., MRVAR A., *Partitioning approach to visualization of large graphs*, http ://www.vlado.fmf.uni-lj.si/pub/, 2002.

[BOR 99] BORGATTI S., EVERETT M., FREEMAN L., *Ucinet for Windows Software for Social Network Analysis*, Harvard : Analytic Technologies, http ://www.analytictech.com, 1999.

[FRI 98] FRIEDKIN N., E.C. J., Social position in influence networks, *Social Networks*, vol. 19, 1998, p. 122-143.

[KOC 89] KOCHEN M., *The small World*, Ablex Publishing Corporation, Norwood, NJ, 1989.

[MOB 02] MOBASHER B., DAI H., LUO T., SUNG Y., ZHU J., Integrating Web Usage and Content Mining for more Effective Personalization, *http ://www.maya.cs.depaul.edu/˜mobasher/personalization*, , 2002.

[MOO 01] MOODY J., Peer influence groups : identifying dense clusters in large networks, *Social Networks*, vol. 23, 2001, p. 261-283.

[SCH 02] SCHOIER G., Blockmodeling Techniques for Web Mining, W. H., B. R., Eds., *Proceedings of Compstat 2002*, Berlin, 2002, Springer and Verlag.

[SRI 00] SRIVASTAVA J., COLLEY J., DESHPANDE M., TON P., *Web Usage Mining : Discovery and Applications of Usage Patterns from Web Data*, http ://www.maya.cs.depaul.edu/˜mobasher/personalization, 2000.

[WAS 94] WASSERMAN S., FAUST K., *Social Network Analysis : Methods and Applications*, Cambridge University Press, New York, 1994.