Catégorisation des pages

Groupe Sensnet

(ce document au format PDF)

1 Sommaire

2 Fichiers de travail:

- Etat n°1 : EchantillondespagesCategorisation-manuelle.zip
- Etat n°2 : Echantillon-apprentissage.zip
- Etat n°3 : Echantillon_apprentissage_v2.zip
- Etat n°4 : validationDesCategories_fichierVierge.zip
- Etat n°5 : Pages Valides.zip

3 Journal "catégorisation manuelle des pages"

3.1 Le 03.02.2004

Voici le classeur des pages à catégoriser à la main. La 1ère page du classeur est le premier tirage au sort et la deuxième page du classeur correspond au second tirage.

Echantillon des pages pour catégorisation manuelle

Typologie de pages :

Page de lien interne / une (page d'accueil à contenu) / page de liens externes / communication (forums,chat, livres d'or, contact) / page de catalogue / page de réponse à une requête /page de login / page de redirection / page d'erreur / pages intraitables / pages de contenu / autre / rien

3.2 16.03.2004

Catégorisation manuelle :

- regroupement sur 65 a 75%
- certaines catégories sont sous représentées : enrichir à la main
- beaucoup de page inaccessible : présence de la page dans la base à vérifier
- pb dans la constitution de l'échantillon (le 3 en particulier) : Arnaud doit vérifier

300 pages à re-catégoriser par personne pour le 23.03

Apres cette catégorisation : classification a mettre en œuvre (identification des traits, construction de traits), choix des méthodes de classif (plusieurs en parallèle...)

3.3 16.03.2004

"Ci joint les pages à catégoriser manuellement. Les catégories sont rappelées dans le classeur."

- Benoît de 77 à 376
- Arnaud de 377 à 676
- Houssem de 677 à 976
- Valérie de 977 à 1276
- Serge de 1277 à 1576
- Laurent de 1577 à 1876
- Nikolay de 1877 à 2176
- Thomas de 2177 à 2476
- Martine de 2477 à 2776
- Michèle de 2777 à 3076

3.4 Le 17.03.2004

"*Il me semble que le fichier joint est invalide : après avoir trouvé beaucoup d'erreurs (404, 500, etc.) au cours du taggage, j'ai regardé un peu dans la base si c'était normal et si l'aspiration avait réellement fonctionné. Au final, je constate une mauvaise correspondance entre identifiants de pages (champ pag_id) et URL : en clair, nous ne taggons pas les bonnes adresses. Arnaud est absent jusqu'à la fin de la semaine, et je n'arrive pas à retrouver mes petits dans la base de données pour sortir une nouvelle sélection de pages à tagguer. Nous attendrons donc le retour d'Arnaud pour mettre de l'ordre dans tout ça… en attendant, ne tagguez pas !! Thomas*"

3.5 Le 17.03.2004

"*Grâce à l'assistance d'un mystérieux "Arnonyme", nous revoilà sur nos pieds, et prêts à catégoriser vaillamment. Pour mémoire, les URL du premier fichier Excel étaient les bonnes, mais les identifiants de pages non : les pages non visibles aujourd'hui ont bien donné lieu à une aspiration il y a quelques mois. Le nouveau fichier tout propre : Echantillon_apprentissage_v2.xls.*

Pour la répartition des URL, c'est pareil qu'avant (cf ci-dessous), mais j'ai rajouté les noms dans le fichier pour qu'on s'y retrouve (des fois qu'emporté par l'élan, on déborde sur les url du voisin...). A noter également : une colonne indiquant si vous avez un doute ou non sur l'attribution d'une catégorie à une page. Mise par défaut à 0, elle vaudra 1 si vous souhaitez que la catégorisation soit vérifiée par quelqu'un d'autre. Bonne catégorisation ! "

3.6 23 03.04

-Résultats categorisations manuelles :

- Idée de BH : repartir les catégories produites sur chacun pour vérifier l'homogénéïté de celles-ci
- Pb sur plusieurs labels de la catégorisation : formulaire, catalogue, annuaire/liste, page de re-direction automatique... nouvelles rubriques pour affiner le classement
- Chacun homogénéïse sa propre catégorisation avec les nouvelles catégories pour le 29.03 matin

3.7 30 03.04

- AB : met les document de la catégorisation des parcours (nos différentes phases de catégorisation et les synthèses) dans l'espace partagé
- VB et HA : font leur révision de catégorisation pour vendredi.
- Jessica, stagiaire de Benoît, va essayer de déterminer les traits visuels qui distinguent nos catégories, en constituant une grille.

Processus de catégorisation des pages

- Etape 1 : catégorisation de 300 pages par personne, avec marquage des doutes
- Etape 2 : synthèse et distribution des cas 'avec doute' à faire évaluer par d'autres
- Etape 3 : l'idée consiste à examiner l'ensemble des pages classées dans une catégorie pour en examiner la cohérence. On élimine les pages qui paraissent vraiment bizarres, hétérogènes par rapport au groupe. Des binômes sont constitués pour chaque catégorie :

  BH et HA pour 1,2,3 et 4
  SF et VB pour 5, 9, 11 et 19
  MHP et TB pour 8
  LA pour 6, 7 et 20
  JB 13, 18 et 99
  NG : 12 et 14

- Etape 4 : sélection des pages les plus homogènes dans chaque catégorie; éventuellement regroupement de certaines catégories.
- Utilisation des traits webxref, utilisation des méthodes de catégorisation supervisée avec apprentissage.
- Pour le 05.04.2004 : fin de l'étape 2 et passage à l'étape 3 : distribution des cas bizarres et des groupements à examiner pour les devoirs de vacances.

3.8 05 04.04

- Préparation de la validation de la catégorisation : cf répartition faite le 30.03
- Envoi du fichier à retraiter le 09.04 et validation à faire pour le 26.04.2004
- Prévoir test de plusieurs processus de classification : kheops, KNN...

3.9 Le 13.04

"Voici la nouvelle passe de validation des pages visitées. Pour mémoire, il s'agit ici de confirmer ou non les rattachements déjà faits => dans le *fichier ci-joint*, remplir la colonne "Cette page est bien classée" : 1 pour dire "oui, c'est bien classé, je suis tout à fait d'accord avec la catégorisation déjà effectuée par mes collègues, j'abonde en leur sens", et 0 pour... etc. La répartition du travail (qui vérifie que les catégories) est donnée dans le CR de la dernière réunion, que voici ci-dessous (point 5) ; vous reverrez le résultat à Jean-François, dont l'adresse SMTP-intelligible est: jvincent.ext@rd.francetelecom.com ".

3.10 Le 27.04

"*Merci à ceux qui n'ont pas encore eu le temps de le faire de valider les catégorisations déjà faites pour le 30 avril au soir et les envoyer à Jean-François (jvincent.ext@rd.francetelecom.com). Il faut que Jean-François puisse faire la synthèse lundi 3 mai en vue de la réunion de mardi 4 prochain. Cf mail de Thomas pour les instructions.*"

BH et HA pour les catégories1, 2, 3 et 4
SF et VB pour 5, 9, 11 et 19
MHP et TB pour 8
LA et MJ pour 6, 7 et 20
JB : 13, 18 et 99
NG et MJ: 10, 12 et 14

3.11 Le 30.04

"Quelques remarques pour aider aux choix entre catégories :

Il me semble que la distinction entre 2 et 3 est souvent difficile. Le type d'interaction (orientation dans des rubriques) est du même ordre, même si on n'est pas au même endroit dans l'arborescence. En outre, on ne voit pas toujours où on est dans l'arborescence. Et inversement, certains sites (pornos) mettent comme une page interne une copie presque conforme (photos exclues) de la page d'accueil du site. Je proposerais donc de fusionner 2 et 3. Mais pas forcément dans une première étape.

Il y a pas mal de déchet dans 4. Pour moi, 4 suppose une forme d'incomplétude : on voit qu'il manque quelque chose (le centre de l'écran, le bas, la droite). Mais il y a aussi des pages de rubriques très creuses qui peuvent ressembler à des 4. Je m'interroge sur le rôle de cette catégorie : il s'agit plutôt de "déchets" qu'il faut retirer (ce ne sont pas des "vraies" pages). Est-ce qu'on ne veut pas apprendre cette catégorie précisément pour pouvoir enlever les pages qui en relèvent à un moment donné.

La notion de pas de porte (1) n'est pas simple. Il y a le choix initial entre langues par exemple. Ou une entrée en matière (image). Pour les sites pornos, c'est la notion de "Dites que vous avez bien plus de 18 ans pour entrer". Mais en même temps, pour ces sites, il y a des rubriques qui en font aussi des pages de type 2.

BH."

3.12 Le 05.05

Voici le fichier Excel contenant la catégorisation définitive,

Pour rappel :

- les catégories 14,15,16,17,18 et 99 sont détruites,
- les catégories 2 et 3 ont fusionnées, la nouvelle catégorie est étiquetée 2,
- les catégories 6 et 20 ont fusionnées, la nouvelle catégorie est étiquetée 6,
- les catégories 9 et 19 ont fusionnées, la nouvelle catégorie est étiquetée 9.

3.13 04.05.04

Catégorisation / réduction du nombre de catégorie

- Catégories à conserver pour apprentissage:

1 2 3 4 5 6 7 8 9 10 11 12 13 19 20

- Regroupements

2+3
9+19
6+20

Remarques de BH sur la catégorisation :

"*Quelques remarques pour aider aux choix entre catégories :.*

- ✍ *Il me semble que la distinction entre 2 et 3 est souvent difficile. Le type d'interaction (orientation dans des rubriques) est du même ordre, même si on n'est pas au même endroit dans l'arborescence. En outre, on ne voit pas toujours où on est dans l'arborescence. Et inversement, certains sites (pornos) mettent comme une page interne une copie presque conforme (photos exclues) de la page d'accueil du site. Je proposerais donc de fusionner 2 et 3. Mais pas forcément dans une première étape.*
- ✍ *Il y a pas mal de déchet dans 4. Pour moi, 4 suppose une forme d'incomplétude : on voit qu'il manque quelque chose (le centre de l'écran, le bas, la droite). Mais il y a aussi des pages de rubriques*

très creuses qui peuvent ressembler à des 4.Je m'interroge sur le rôle de cette catégorie : il s'agit plutôt de "déchets" qu'il faut retirer (ce ne sont pas des "vraies" pages). Est-ce qu'on ne veut pas apprendre cette catégorie précisément pour pouvoir enlever les pages qui en relèvent à un moment donné.

- ✍ *La notion de pas de porte (1) n'est pas simple. Il y a le choix initial entre langues par exemple. Ou une entrée en matière (image). Pour les sites pornos, c'est la notion de "Dites que vous avez bien plus de 18 ans pour entrer". Mais en même temps, pour ces sites, il y a des rubriques qui en font aussi des pages de type 2*."

4 Tableau des catégories

Rappel des catégories:	Identifiants des catégories:
Pas de porte (Seuil du site), par exemple avertissement ou intro	1
Page d'accueil (Page de liens internes)	2
Page d'accueil de rubrique, accueil interne	3
Page de sommaire (Frame Gauche)	4
Page de liens externes	5
Page de communication (forum, mail, chat)	6
Page de contact au site, adresse, info sur l'entreprise	7
Page de contenu (au sens feuille de l'arborescence du site) à contenu	8
Page de catalogue	9
Page de recherche (Formulaire)	10
Page de réponse à une requête type recherche	11
Page de login	12
Page d'erreur	13
Page de redirection	14
Page d'applet ou de flash	15

Page de plan du site	16
Page Autre	17
Page Rien (NOK) (Les images, ou morceaux de page qui ne font pas sens seuls)	18
Annuaire de catégories	19
Formulaire (inscriptions, mails avec formulaires)	20
Pages inaccessibles	99

5 Synthèse