LAGHZALI Hind
Technologies de l'Information pour la Santé
Rapport de stage
2008/2009

DETECTION AUTOMATIQUE DE LA QUALITE SUR LE WEB SANTE

Annexe

2008/2009

Table des Matières

Index des Illustrations et des tables

A. Evaluation du corpus

L'étude suivante est réalisée dans le but de permettre l'identification de critères plus spécifiques qui caractérisent le corpus de données. Ces caractéristiques pourront être utilisées pour améliorer les performances du système de classification de la qualité mis en place [cf. Tome principal].

1. Critères d'évaluation du corpus

On souhaite vérifier l'hypothèse selon laquelle les documents appartenant à la catégorie 'Mauvais' seraient plus longs. En effet, durant la phase de préparation, on a constaté que certain de ces documents contenaient une information redondante et répétitive. Le nombre de mots ainsi que le ratio de compression de chaque document sont calculés. La commande gzip de PERL est utilisée pour compresser les documents du corpus. Le ratio de compression est ensuite calculé selon la formule suivante :

$$= \frac{\qquad\qquad}{\text{è}}$$

On a également remarqué que certains 'mauvais' documents étaient des témoignages d'internautes en faveur d'un traitement ou d'une méthode donnée. On a alors supposé que les pages contenant ce type d'informations auraient un caractère plus émotif. Une manière d'évaluer cela est de compter la proportion d'adjectifs et d'adverbes dans chaque document. C'est le ratio de compression calculé ci-dessous. Pour identifier les adjectifs et adverbes on utilise les ressources Internet.

$$\text{R} \qquad \text{é} = \frac{\qquad\qquad}{\qquad\qquad}$$

2. Résultats

Les valeurs du ratio de compression obtenues sont représentées dans le graphique suivant avec en abscisse l'échelle du ratio comprise entre 0.2 et 8.6 et en ordonnées le nombre de documents correspondant. On constate que la plupart des documents ont un ratio de compression compris entre 2.2 et 4.2. Ces résultats ne mettent pas en évidence une différence entre les deux catégories de documents pour ce critère.

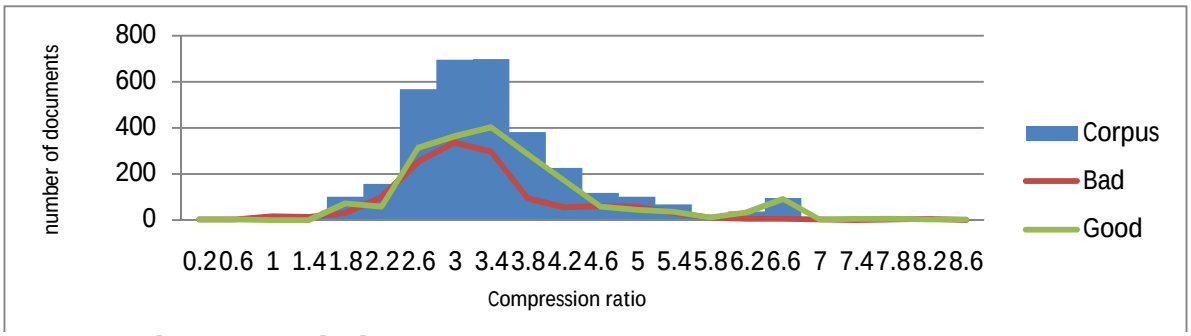

Figure 1:ratio de compression des documents

Le graphique suivant représente le nombre de mots par document. Pour un nombre de mots inférieur à 2700, les documents de la catégorie 'Bon' sont plus représentés que ceux de la catégorie 'Mauvais'. Cependant, ce résultat n'est pas suffisant pour affirmer le caractère discriminatif de ce critère.

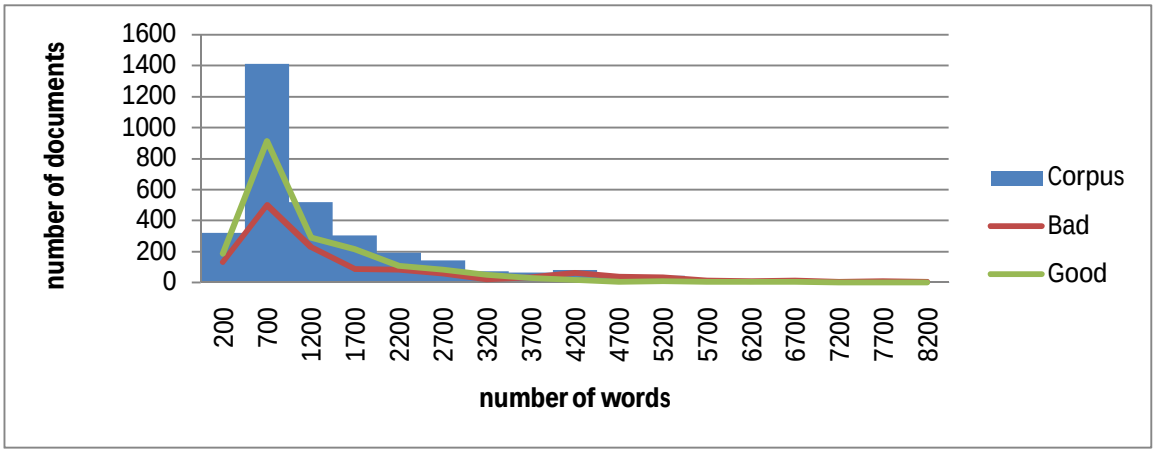

Figure 2: Nombre de mots des documents

Le graphique suivant présente le ratio de sensibilité des différents documents. Contrairement à l'hypothèse de départ, la plupart des documents présentent un faible ratio de sensibilité. Pour un ratio compris entre 0.015 et 0.065, les 'Bon' documents sont plus représentés.

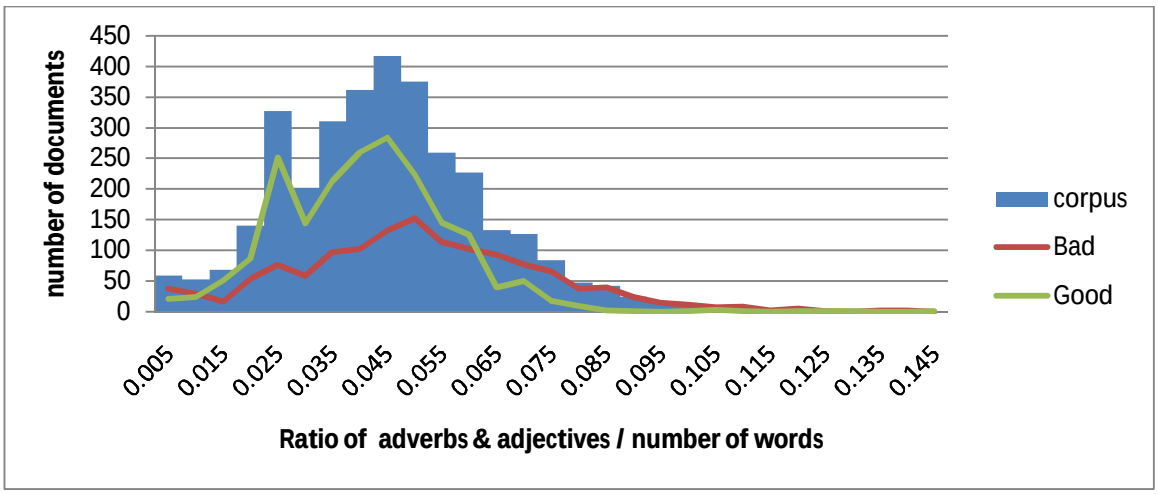

Figure 3: ratio de sensibilité des documents

L'ensemble des résultats obtenus ne permet pas une exploitation des critères évalués ci-dessus pour l'amélioration du système de détection de qualité.

***

B. Classification par domaine

Afin de mettre en place un système de classification de qualité plus précis, un apprentissage par domaine est envisagé. Cet apprentissage doit permettre l'identification du thème de la page puis l'évaluation de son contenu en faisant appel à un système de détection de qualité spécifique à ce domaine.

Nous commençons par réaliser un système de détection du domaine. Pour cela on utilise le même corpus de départ en changeant seulement la labellisation des documents. En effet, on attribue à chaque document l'un des dix labels suivants : « Laetrile cancer, Magnesium restless legs syndrome, Hydrogen peroxide cancer, Growth hormone anti-aging, Shark cartilage cancer, Pills weight loss, Cure diabetes, Cure multiple sclerosis, Alternative treatment asthma, Atkins diet » [voir Tome principal pour le choix des labels].

Ensuite, l'apprentissage est réalisé selon un protocole similaire à celui utilisé pour le système de détection de qualité décrit dans le tome principal.

Les résultats obtenus [ci dessous] montre une bonne performance du système pour détecter le domaine.

maR	maP	maF1	miR	miP	miF1	Err
0.9546	0.9109	0.9150	0.961	0.9537	0.9537	0.0085

Table 1: Performances du système de classification de domaines

La deuxième partie consiste à réaliser des classificateurs de la qualité pour chacun des domaines précédemment cités. Nous avons ainsi 10 mini corpus formés à partir du corpus d'origine.

Pour chaque domaine, les documents correspondants sont soumis au même protocole d'apprentissage. Les premiers résultats de l'évaluation ont montré une mauvaise performance du système. Cela est dû au sous-apprentissage du système. En effet, en séparant les documents par domaine, le volume du corpus est divisé par 10, ce qui rend le nombre de documents insuffisant pour permettre un bon apprentissage.

Pour résoudre ce problème on devrait ajouter suffisamment de documents au corpus pour permettre un bon apprentissage. Malheureusement, faute de temps cette solution ne sera pas réalisée.

***

C. Principes HONcode

- **Autorité** : Indiquer la qualification des rédacteurs : tout avis médical fourni sur le site sera donné uniquement par du personnel spécialisé (diplômé) du domaine médical et des professionnels qualifiés, à moins qu'une déclaration explicite ne précise que certains avis proviennent de personnes ou d'organisations non médicales.

- **Complémentarité** : Complémenter et non remplacer la relation patient-médecin : l'information diffusée sur le site est destinée à encourager, et non à remplacer, les relations existantes entre patient et médecin.

- **Confidentialité** : Préserver la confidentialité des informations personnelles soumises par les visiteurs du site : les informations personnelles concernant les patients et les visiteurs d'un site médical, y compris leur identité, sont confidentielles. Le responsable du site s'engage sur l'honneur à respecter les conditions légales de confidentialité des informations médicales applicables dans le pays dans lequel le serveur (ainsi que les éventuels sites- miroir) est situé.

- **Attribution** : Citer la/les source(s) des informations publiées et dater les pages de santé : la source des données diffusées sur le site est explicitement citée avec, si possible, un hyperlien vers cette source. La date de la dernière modification doit apparaître clairement sur la page Web (par exemple: en bas de chaque page).

- **Justification** : Justifier toute affirmation sur les bienfaits ou les inconvénients de produits ou traitements : toute affirmation relative au bénéfice ou à la performance d'un traitement donné, d'un produit ou d'un service commercial, sera associée à des éléments de preuve appropriés et pondérés selon le principe 4. ci-dessus.

- **Professionnalisme** : Rendre l’information la plus accessible possible, identifier le webmestre, et fournir une adresse de contact : les créateurs du site s'efforceront de fournir l'information de la façon la plus claire possible, et fourniront une adresse de contact pour les utilisateurs qui désireraient obtenir des détails ou du soutien. Cette adresse (e-mail) doit être clairement affichée sur les pages du site.

- **Transparence du financement** : Présenter les sources de financements : le support d'un site doit être clairement identifié, y compris les identités d'organisations commerciales et non-commerciales qui contribuent au financement, services ou matériel du site.

- **Honnêteté dans la publicité et la politique éditoriale** : Séparer la politique publicitaire de la politique éditoriale : si la publicité est une source de revenu du site, cela sera clairement établi. Le propriétaire du site fournira une brève description de la règle publicitaire adoptée. Tout apport promotionnel ou publicitaire sera présenté à l'utilisateur de façon claire afin de le différencier de l'apport uniquement créé par l'institution gérant le site.

***