Utilisation de méthodes d'apprentissage pour la détection des critères de qualité des documents en ligne du Web médical

Arnaud Gaudinat, Natalia Grabar, Célia Boyer

Fondation Health on the Net, Hôpitaux Universitaires de Genève, Genève, Suisse
{arnaud.gaudinat;natalia.grabar;celia.boyer}@healthonnet.org

Abstract

Objectives : The amount of health information is constantly growing on the Web and users can easily access it. However, the quality of this information is very difficult to control. The objective of this project is to propose automated methods to detect the quality of health documents.

Methods and material : Methods proposed rely on machine learning algorithms (*i.e.*, Support Vector Machine, Naive Bayes, Decision Trees) and, in particular, on the database of quality accredited websites compiled, since 1996, when the Health on the Net (HON) Foundation was established. This database contains websites that respect HONcode's ethical principles. Different settings of learning algorithms are tested.

Results : The initial evaluation shows promising results. Currently, the system shows 0,78 of microprecision and 0,73 of f-measure, the error rate is 0,06.

Conclusions : The applied methods appear to be suitable for the detection of the HONcode's ethical principles on health webpages, and current results are encouraging. Nevertheless, there is room for improvements.

Keywords :

Medical Web, Quality of health information, Automatic categorisation, Natural Language Processing

1 Évaluation de la qualité des informations sur l'Internet

La masse des informations sur l'Internet augmente continuellement. Dans le domaine de santé, par exemple, un utilisateur atteint sans difficulté des millards de pages dans lesquelles il pourra certainement trouver des réponses à ses questions. Si l'accès aux informations de santé devient facile et aisé, il reste difficile de contrôler et de garantir leur qualité. Ceci essentiellement pour deux raisons : (1) la notion de *qualité* des documents de santé est difficile à définir, (2) le très grand nombre de documents disponibles sur Internet.

La notion de qualité des documents médicaux peut se présenter en effet sous plusieurs facettes et, donc, être abordée de différents points de vue. Citons par exemple, l'analyse du contenu médical des documents, l'analyse de la transparence des informations fournies, le jugement

sur la fiabilité du fournisseur des documents ou encore la compréhension du contenu pour les usagers non professionnels. Par ailleurs, le jugement sur la qualité d'un document peut être basé sur l'opinion (implicite) d'un utilisateur, professionnel du domaine de santé ou non, ou bien identifié (explicitement) en fonction des principes clairement définis. Il est évident que, dans ces différentes situations, le jugement même sur la fiabilité des informations pourra être considéré comme plus ou moins fiable.

Quant au problème lié à la quantité des informations, il demande une réponse à la dimension de l'Internet. En effet, la sélection ou la révision de pages de santé par des experts humains devient utopiste aux vues de la production quotidienne, du dynamisme et des différences culturelles existantes sur l'Internet. Actuellement, quatre grands types de systèmes sont en usage sur l'Internet pour gérer l'aspect qualitatif de l'information : sélection ou référencement, certification, indice de popularité, collaboration des utilisateurs. Il existe, bien entendu, d'autres initiatives particulières mais seules celles utilisées à une large échelle ou en devenir sont présentées ici.

1.1 Sélection ou référencement des pages Internet

Historiquement, les premières initiatives concernant la qualité des informations sur l'Internet datent de la naissance de la *sélection*, ou du *référencement*, comme le fait par exemple l'annuaire de Yahoo ![1] Le principe consiste à publier une sélection de pages par rapport à des domaines organisés plus ou moins hiérarchiquement. Ainsi, la rubrique *Health* de Yahoo ! regroupe environ 50 sous-rubriques parmi lesquelles *Alternative Medicine*, *Procedures and Therapies*, *Nutrition*, *Weight Issues*, *Fitness* et *Pet Health*. La sélection est en général opérée par des experts du domaine, ce qui peut garantir la qualité des ressources répertoriées dans un annuaire généraliste comme Yahoo ! La sélection des rubriques dépend des préférences des experts. Parmi les annuaires ou les portails médicaux, on peut citer MedlinePlus[2] et CISMeF[3]. Dans le cas de CISMeF, les rubriques, selon lesquelles les documents sont organisés, correspondent aux catégories et mots-clés MeSH[4] (NLM, 2001), une terminologie dédiée à l'indexation et recherche d'information dans le domaine biomédical.

Une des limitations de cette initiative concerne la qualité réelle des experts, les jugements qu'ils effectuent sur le contenu des pages Internet rencontrées et sélectionnées et, finalement, la difficulté de traiter les masses d'informations actuellement disponibles.

1.2 Certification des pages Internet

La *certification* est un autre type d'initiative, qui vise en particulier à encourager le développement de la qualité de l'information sur l'Internet. Dans cette approche, les sites doivent respecter une liste de principes ou critères afin de bénéficier de la certification. Les initiative de certification les plus populaires dans le domaine de la santé sur Internet au niveau mondial sont le HONcode[5] (Boyer *et al.*, 1999) et URAC[6].

[1] *search.yahoo.com/dir*
[2] *medlineplus.gov*
[3] *www.chu-rouen.fr/cismef*
[4] *www.nlm.nih.gov/mesh/meshhome.html*
[5] *www.hon.ch/HONcode*
[6] *www.urac.org*

URAC a défini un ensemble de 16 types de documents pour lesquels les principes de certification sont fournis. Parmi ces types de documents, citons *Claims Processing* (*Traitement des plaintes*), *Disease Management* (*Traitement des maladies*), *Health Web Site* (*Site Internet de santé*) et *Privacy* (*Confidentialité*). La certification effectuée par URAC est un service payant et forcément élitiste. Quant aux principes proposés par HONcode, ils couvrent actuellement le code éthique des sites Internet et peuvent s'appliquer à tout type de documents. Ces principes sont au nombre de huit : *Authority* (*Autorité*), *Complementarity* (*Complémentarité*), *Privacy* (*Confidentialité*), *Reference* (*Référence*), *Justifiability* (*Justification*), *Authorship* (*Précision de l'auteur*), *Sponsorship* (*Précisions sur les subventions*), *Advertising* (*Publicité*). Contrairement à URAC, la certification proposée par HON est gratuite et volontaire.

Qu'il s'agisse de la certification par URAC ou par HONcode, les principes définis doivent être respectés dans les sites Internet pour que ceux-ci soient certifiés. Pour ceci, les experts examinent les pages et les sites avant d'accorder le certificat. Une des limitations de cette initiative consiste dans l'impossibilité de traiter les grosses masses d'informations que nous offre actuellement l'Internet.

1.3 Popularité des pages Internet

Une autre initiative est liée à l'utilisation de la *popularité* des pages Internet à travers le modèle Page Rank (Page *et al.*, 1998) ou les modèles similaires. Cette initiative a été principalement mise au point pour répondre au problème de la surabondance de l'information sur l'Internet. Elle est directement basée sur le jugement des créateurs des pages Internet lorsqu'ils décident de faire un lien vers d'autres pages, qu'ils considèrent alors comme intéressantes et « dignes » d'être référencées. Dans le cas de cette initiative, le classement d'une page est élevé si elle est souvent référencée par d'autres pages. Il existe en effet une hypothèse selon laquelle il y a une forte corrélation entre la popularité d'une page et sa qualité. Cette initiative aussi montre des limitations.

Une des limitations est liée au fait que les créateurs des pages participent à la formation de l'opinion experte sur la valeur des documents de l'Internet. Leur impartialité et compétence lors du référencement des pages Internet prennent ainsi une place importante. Une autre limitation est liée à l'effet de *retranchement* (*entrenchment*), qui remet en cause le principe même du modèle Page Rank. Selon (Pandey *et al.*, 2005), la relation entre la popularité et la qualité est très faible dans le cas des pages nouvellement créées. En effet, ces pages n'ont pas encore reçu beaucoup de visites de l'extérieur et ne sont pas, de ce fait, très connues. La découverte de ces nouveaux contenus est donc basée sur les moteurs de recherche. Dans ce cas, puisque les utilisateurs se concentrent essentiellement sur les premiers résultats, les pages nouvellement créées sont rarement visitées. En conséquence, elles mettent beaucoup de temps avant de devenir populaires et être référencées par d'autres pages. Dans ce cercle de vie, leur valeur de Page Rank restera donc faible assez longtemps. Pour tenter de régler ce problème, Google « gonfle » artificiellement le Page Rank de manière arbitraire.

1.4 Collaboration des utilisateurs

La quatrième initiative est plus récente et peut être caractérisée comme une *initiative sociale* ou de *collaboration*. Elle consiste à utiliser le jugement de tout internaute pour caractériser une

page et, aussi, à créer des réseaux d'utilisateurs de confiance par domaine d'expertise. Outfoxed ou Lijit[7] et, plus récemment, Google co-op[8] sont des initiatives de ce type. Elles semblent ouvrir des perspectives intéressantes à condition de motiver suffisamment d'utilisateurs et de contributeurs. Là encore, la masse d'informations sur l'Internet et la qualité des contributeurs représentent une limitation.

2 Contexte de travail

La fondation Health on the Net (HON)[9] travaille depuis 1995 pour contribuer à l'amélioration de la qualité des documents de santé sur l'Internet. HON aborde la question de qualité à travers l'élaboration des critères de qualité, qui forment le HONcode du comportement éthique (Selby *et al.*, 1996; Boyer *et al.*, 1997), et à leur implémentation pour la certification des documents de santé. HON se positionne donc parmi les initiatives de *certification* (et non auto-certification) des sites de santé sur l'Internet par les experts du domaine de santé (sec. 1.2). En prenant les critères éthiques comme base pour l'évaluation, HON fait l'hypothèse que si les informations en relation avec les critères éthiques (identification des créateurs du contenu de sites, financement du site, confidentialité des informations fournies par l'utilisateur, datation, etc.) sont indiquées clairement et correctement cela garantit la traçabilité et qualité des informations médicales, et l'utilisateur ne prend donc pas de risque lorsqu'il consulte ce site. Les critères de qualité, qui composent actuellement le HONcode, sont au nombre de huit :

1. *Autorité*. L'avis médical est donné par du personnel spécialisé du domaine médical.
2. *Complémentarité*. L'information proposée ne remplace pas les relations avec un médecin.
3. *Confidentialité*. L'information personnelle concernant les utilisateurs du site, y compris leur identité, est confidentielle.
4. *Attribution*. Les sources des données diffusées sur le site sont explicitement citées.
5. *Justification*. Toute affirmation relative à la performance d'un traitement donné, d'un produit ou d'un service commercial, est justifiée.
6. *Transparence de l'auteur*. L'auteur du contenu est identifié.
7. *Transparence des sponsors*. Les sources de financement sont clairement identifiées.
8. *Politique publicitaire et éditoriale*. Si la publicité est une source de revenu du site, cela est clairement indiqué.

L'objectif principal de HON consiste donc à offrir, grâce au HONcode, un cadre simple garantissant et renforçant une meilleure transparence et qualité des informations médicales présentes sur l'Internet. La soumission de demandes pour la certification est gratuite et volontaire. Les sites sont évalués par les réviseurs. Le sceau HONcode est attribué aux sites lorsqu'ils respectent les critères du code. Si certains critères du code ne sont pas satisfaits, un travail pédagogique est mené avec les webmasters afin de renforcer la transparence éthique et la qualité de leur site. Pour l'apposition du sceau HONcode, HON utilise des techniques de *sceaux actifs*, qui permettent aux utilisateurs de vérifier en temps réel et aisément la conformité de la certification d'un site par rapport à la base de données de HON. La vérification est effectuée par les navigateurs et par

[7] *www.lijit.com*
[8] *www.google.fr/coop*
[9] *www.hon.ch*

l'intermédiaire des outils spécifiques développés par HON. Les sites déjà certifiés sont réévalués tous les ans et, éventuellement, d'une manière aléatoire au cours de l'année. Actuellement, plus de 5 000 sites, répartis dans 72 pays, sont certifiés. Cela représente une estimation d'environ 1 200 000 pages sur Google. Cette base de données de sites certifiés est unique en son genre. Actuellement, la langue la plus présente est l'anglais mais d'autres langues, tels que le français, est très bien représentées. Les sites certifiés HONcode sont accessibles à travers le service *Google co-op health*,[10] ce qui garantit leur accès et leur visibilité dans le monde entier.

La certification des sites est effectuée par les réviseurs manuellement. Cela garantit une haute qualité du résultat, et donc, une sécurité pour les utilisateurs. Mais une telle approche représente aussi une limitation (voir sec. 1.2), qui réside dans l'impossibilité physique d'évaluer et de réévaluer un nombre très important de sites de santé que nous propose actuellement l'Internet et de répondre rapidement aux demandes de certification adressées à HON. L'objectif du travail, que nous présentons dans cet article, consiste à dépasser cette limitation et à proposer des méthodes automatiques d'aide à la détection, au sein des sites de santé, des pages en relation avec les critères HONcode. Par la suite, ces outils assisteront les réviseurs dans leur travail quotidien. En l'occurrence, ils permettront de systématiser et d'accélérer le processus de certification.

3 Méthodes

Pour la détection des pages en relation avec les critères du HONcode, nous avons choisi d'appliquer les méthodes statistiques : les algorithmes d'apprentissage automatique. Parmi les avantages de ces algorithmes, citons l'efficacité dans la tâche, le gain de temps au niveau de l'utilisation d'expert, la maturité des modèles et l'indépendance du domaine. De plus, l'habilité de généralisation de ces méthodes permet de capturer des événements difficilement identifiables pour un expert et rend donc des méthodes basées sur des mots clés (Wang & Liu, 2006) ou des patterns obsolètes du point de vue de la qualité des résultats. Parmi les méthodes d'apprentissage, on distingue habituellement les méthodes supervisées et non supervisées. Les méthodes supervisées nécessitent d'avoir une base d'apprentissage. Elle sont basées sur une étape d'entraînement ou de description. Tandis que les méthodes non supervisées n'ont pas besoin de disposer d'une base d'apprentissage. Elles sont guidées par les propriétés internes des données. Comme, dans notre travail, nous voulons identifier les contenus en relation avec les huit critères du HONcode, nous avons choisi d'appliquer les méthodes d'apprentissage supervisées. En effet, elles nous permettent de mieux caractériser et contraindre les catégories attendues. Longuement évaluées et optimisées sur des Corpus tel que Reuters (Grobelnik & Mladenić, 1998; Manevitz & Yousef, 2001), ces techniques ont montré leur intérêt pour catégoriser les documents selon un jeu d'étiquettes, plus ou moins grand. Par la suite, ces techniques ont été employées avec succès pour traiter des documents textuels : pour la détection de messages hostiles (Spertus, 1997), de contenus racistes (Vinot *et al.*, 2003), ou pour le filtrage des « spams » (Carreras & Márquez, 2001).

L'application de la méthode pour la catégorisation des documents de santé selon des critères du HONcode est effectuée selon les étapes et les points méthodologiques décrits dans la suite de cette section.

Algorithmes d'apprentissage. Les algorithmes d'apprentissages appliqués dans cette étude sont ceux offerts par la plateforme d'apprentissage utilisée (Williams & Calvo, 2002). En l'oc-

[10] *www.google.fr/coop*

currence, nous disposons de quatre classifieurs : *Naive Bayse* (NB), *Support Vector Machine* (SVM), *K Nearest Neighbour* (KNN) et *Decision Tree* (DT). Les algorithmes d'apprentissage représentent les documents sous forme de vecteurs dans un espace vectoriel. La dimension de cet espace est déterminée par le nombre d'attributs priss en compte (mots, *n-gram*, expressions). La taille de chaque vecteur est déterminée par la fréquence d'apparition des attributs dans un document. Dans le travail présenté ici, nous avons choisi de baser la catégorisation sur les phrases et non les documents entiers : l'information véhiculée par une phrase est plus homogène que celle contenue dans un document. Il nous semble donc que, au moins au stade initial de cette étude, le travail sur les phrases est plus raisonnable.

Pré-traitement des documents. Puisque l'identification des critères est réalisée au niveau des phrases, il est nécessaire de pré-traiter les documents et de les segmenter en phrases. La segmentation en phrases est basée sur des expressions régulières, qui prennent en comptes les signes de ponctuation et les balises `HTML`.

Attributs. Plusieurs types d'attributs de traitement ont été envisagés et testés dans notre étude : (1) utilisation ou non de « mots vides » (mots outils et grammaticaux comme *avec*, *de* ou *pour*) afin de réduire les mots les plus fréquents ; (2) utilisation ou non de l'algorithme de désuffixation type Porter (Porter, 1980) afin de réduire la variation lexicale et normaliser les mots sous un même attribut (par exemple, *sténoses* -> *sténose*) ; (3) utilisation d'un groupe de mots qui se suivent : *n-gram* de mots dans une fenêtre de 1 à 4 mots ; (4) utilisation de mots, qui apparaissent dans la même phrase, sous forme de « sacs de mots » (mots de la phrase triés dans l'ordre alphabétique).

Langues. Les langues traitées sont celles qui sont le mieux représentées dans les bases de sites certifiés avec le HONcode : anglais, français, espagnol et italien. Comme les sites en langue anglaise sont les plus nombreux et représentent donc un corpus d'apprentissage le plus important, c'est sur cette langue que nous avons testé différents réglages (combinaisons de traits et d'algorithmes d'apprentissage). Pour les autres langues, seuls les réglages donnant de meilleurs résultats sur l'anglais ont été appliqués.

Pondération des attributs. L'approche la plus simple consiste à utiliser la fréquence d'apparition d'un attribut au sein du document (phrase). Pour des raisons d'optimisation des performances il peut s'avérer intéressant de pondérer la fréquence de l'attribut par rapport à sa distribution au sein de la collection de documents. La pondération est alors composée de trois éléments (Salton, 1991; Singhal *et al.*, 1996) : fréquence de l'attribut (term frequency), fréquence inverse de documents (inverse document frequency) et normalisation de la longueur (length normalization). C'est cette dernière version de la pondération que nous utilisons.

Sélection des attributs. La sélection des attributs en classification de texte a été étudiée de façon intensive ces dernières années (Koller & Sahami, 1996). Elle poursuit un double objectif : (1) réduire la dimension de l'espace vectoriel, c'est-à-dire le nombre d'attributs, en en sélectionnant une partie, et (2) obtenir de meilleurs résultats en essayant d'éliminer les attributs, qui produisent le plus de bruit, et en ne gardant que les attributs les plus discriminants. Dans cette étude, nous nous sommes limités à réaliser une sélection d'attributs à partir du critère de fréquence de document (document frequency, DF) qui favorise les attributs qui sont distribués dans le plus grand nombre de documents (phrases). L'avantage du DF est d'offrir de bons résultats pour une réduction de plus de 80 % du coût de calcul (Yang & Liu, 1999).

Corpus d'entraînement et de test. Les corpus d'entraînement et de test sont composés, respectivement, de 90 % et 10 % de documents disponibles.

Critères	Anglais	Français	Espagnol	Italien
Autorité	1685	188	123	230
Complémentarité	1738	182	119	190
Confidentialité	1561	128	106	187
Attribution	1039	112	71	128
Justification	323	25	17	28
Auteur	1813	177	120	201
Sponsor	1473	163	101	163
Publicité	1030	103	86	142

TAB. 1 – Nombre de passages sélectionnés selon les langues et les critères du HONcode

Évaluation. L'évaluation est réalisée en appliquant les mesures suivantes, dans leurs versions micro et macro : précision, rappel and F-mesure. La macro précision (maP) permet d'évaluer la distribution correcte au niveau des catégories (critères du HONcode), tandis que la micro précision (miP) permet d'évaluer la distribution correcte au niveau des phrases.

4 Matériel

Nous utilisons deux types de matériel : critères du HONcode et sites de santé catégorisés en fonction de ces critères. Les critères du HONcode, spécifiés dans la section 2, sont au nombre de huit : *Autorité*, *Complémentarité*, *Confidentialité*, *Attribution*, *Justification*, *Transparence de l'auteur*, *Transparence des sponsors* et *Politique publicitaire et éditoriale*. En ce qui concerne les sites certifiés, nous avons utilisé les passages extraits à partir des sites. En effet, lors de la certification d'un nouveau site ou de la réévaluation annuelle des sites certifiés, les réviseurs sont invités à identifier et sélectionner, dans le texte, les passages représentant le ou les critères. Cette sélection est réalisée pour chaque critère de qualité. Ces données sont ensuite stockées dans une base de données, accompagnées de l'adresse URL correspondante. La table 1 indique les nombres de passages enregistrés par langue et pour chaque critère. Nous pouvons voir que l'anglais est le mieux représenté dans cette collection de données : les volumes pour cette langue sont environ 10 fois plus importants que dans d'autres langues. Par ailleurs, à l'instar du HONcode, nous avons distingué, au sein du critère *Attribution*, les données relatives au sous-critère *Date* comme neuvième critère. Nous disposons donc de neuf corpus dans chaque langue : *Autorité*, *Complémentarité*, *Confidentialité*, *Attribution Ref* , *Justification*, *Auteur*, *Sponsor*, *Publicité* et *Attribution Date*.

5 Résultats

La méthode décrite dans la section 3 a été appliquée à la catégorisation des documents, représentés dans la table 1, selon les huit critères du HONcode et le sous-critère *Attribution Date*. La catégorisation a été effectuée en quatre langues : anglais, français, espagnol et italien.

La table 2 illustre les résultats globaux en terme de macro et micro précision moyenne, de rappel, de F-mesure et du taux d'erreurs. Les quatre premières colonnes de la table indiquent les différents réglages du système. La première colonne *Lang* indique la langue de la collection

Lang.	Attr.	Méth.	Pondér.	maR	maP	maF1	miR	miP	miF1	Err
Eng	w1	NB	nnn	0,72	0,67	0,66	0,81	0,65	0,72	0,07
Eng	w1	NB	ann	0,71	0,71	0,65	0,81	0,65	0,72	0,07
Eng	w1	NB	ntn	0,76	0,59	0,66	0,79	0,64	0,71	0,07
Eng	w1	NB	nnc	0,65	0,61	0,61	0,77	0,61	0,68	0,08
Eng	w1	NB	atn	0,75	0,59	0,66	0,80	0,64	0,71	0,07
Eng	w1	NB	atc	0,67	0,68	0,61	0,77	0,61	0,68	0,08
Eng	w1	NB	lnn	0,72	0,67	0,65	0,81	0,65	0,72	0,07
Eng	w1	NB	ltn	0,76	0,59	0,65	0,79	0,64	0,71	0,07
Eng	s1	NB	nnn	0,77	0,60	0,67	0,82	0,63	0,71	0,07
Eng	cooc	NB	nnn	0,69	0,74	0,69	0,75	0,73	0,74	0,05
Eng	cooc	NB	atn	0,70	0,72	0,70	0,75	0,73	0,74	0,06
Eng	cooc	NB	atc	0,57	0,63	0,58	0,68	0,65	0,66	0,07
Eng	cooc	NB	ann	0,64	0,77	0,65	0,74	0,70	0,72	0,06
Eng	w2	NB	nnn	0,68	0,77	0,67	0,78	0,70	0,74	0,06
Eng	w2	NB	atn	0,73	0,67	0,69	0,78	0,71	0,74	0,06
Eng	w2	NB	ann	0,67	0,76	0,66	0,77	0,70	0,73	0,06
Eng	w3	NB	nnn	0,66	0,77	0,67	0,76	0,71	0,73	0,06
Eng	w4	NB	nnn	0,66	0,75	0,67	0,75	0,71	0,73	0,06
Eng	w1	SVM	nnn	0,64	0,73	0,68	0,69	0,78	0,73	0,06
Eng	cooc	SVM	ann	0,71	0,60	0,65	0,76	0,63	0,69	0,07
Eng	w1	KNN	nnn	0,42	0,73	0,52	0,45	0,84	0,59	0,07
Fra	w1	NB	nnn	0,80	0,71	0,74	0,81	0,61	0,70	0,08
Fra	cooc	SVM	nnn	0,69	0,78	0,73	0,64	0,75	0,69	0,06
Fra	w1	SVM	nnn	0,70	0,82	0,75	0,65	0,80	0,72	0,06
Esp	w1	NB	nnn	0,54	0,47	0,50	0,67	0,5	0,58	0,10
Esp	cooc	SVM	nnn	0,42	0,50	0,45	0,50	0,57	0,53	0,10
Ita	w1	NB	nnn	0,67	0,54	0,58	0,81	0,60	0,70	0,08
Ita	cooc	SVM	nnn	0,54	0,64	0,57	0,63	0,76	0,69	0,06

TAB. 2 – Évaluation des résultats selon la précision (*P*), le rappel (*R*) et la F-mesure (*F1*) dans leurs versions micro (*mi*) et macro (*ma*)

évaluée. La seconde colonne *Attr.* décrit les attributs : *w1* pour mot unique, *w2* pour bigram, *w3* pour trigram, *cooc* pour cooccurrences de mots, *s1* pour mot unique désuffixé (avec l'algorithme Porter). La troisième colonne *Méth.* indique l'algorithme d'apprentissage utilisé : *NB* pour Naive Bayes, *SVM* pour Support Vector Machine, *KNN* pour k-Nearest Neighbour et *DT* pour Decision Tree. La quatrième colonne *Pondér.* indique la pondération des attributs : le premier caractère est lié à la fréquence de l'attribut (*n* pour *natural*, *l* for *logarithmic* et *a* pour *augmented*), le second caractère représente la prise en compte ou non de la fréquence inverse de documents (*n* pour *none*, *t* pour *full*), le troisième caractère est lié à la longueur du document considéré (*n* pour *none*, *c* pour *cosine*). Les résultats d'évaluation, indiqués dans cette table, montrent qu'en anglais les performances de différents réglages sont assez équivalentes entre elles : entre 0,59 et 0,82 toute mesure d'évaluation confondue. L'algorithme *kNN* sort de cette fourchette car il obtient les valeurs du rappel très faibles (0,42 et 0,45), même si les valeurs de la précision restent élevées (0,73 et 0,84). Un des meilleurs résultats est fourni par l'algorithme *SVM* appliqué sur les mots uniques *w1* sans aucune pondération *nnn*. Appliqué à d'autres

	Autorité	Compl.	Conf.	Attr.	Justif.	Auteur	Sponsor	Pub.	Date
Autorit	0,64/0,72	0,05/0,05	0,01/0,01	0,19/0,34	0,01/0,09	0,04/0,13	0,04/0,09	0,01/0,01	0,00/0,01
Compl.	0,05/0,05	0,80/0,82	0,05/0,03	0,01/0,02	0,06/0,44	0,00/0,00	0,03/0,05	0,00/0,01	0,00/0,00
Conf.	0,02/0,03	0,02/0,04	0,92/0,90	0,00/0,01	0,00/0,03	0,01/0,02	0,01/0,02	0,02/0,06	0,00/0,00
Attr.	0,24/0,12	0,03/0,02	0,03/0,01	0,64/0,57	0,02/0,08	0,01/0,01	0,02/0,02	0,00/0,00	0,01/0,02
Justif.	0,06/0,01	0,32/0,03	0,06/0,00	0,06/0,01	0,45/0,33	0,02/0,01	0,00/0,00	0,00/0,00	0,00/0,00
Auteur	0,06/0,02	0,02/0,01	0,08/0,02	0,02/0,01	0,00/0,00	0,81/0,81	0,00/0,00	0,01/0,00	0,00/0,00
Sponsor	0,05/0,03	0,04/0,02	0,02/0,01	0,01/0,01	0,00/0,02	0,02/0,02	0,69/0,69	0,16/0,17	0,00/0,00
Pub.	0,01/0,01	0,02/0,01	0,05/0,01	0,00/0,00	0,00/0,02	0,00/0,00	0,13/0,12	0,77/0,73	0,00/0,00
Date	0,00/0,00	0,01/0,00	0,01/0,00	0,06/0,03	0,00/0,00	0,00/0,00	0,01/0,01	0,01/0,01	0,90/0,98

TAB. 3 – Contingence Précision/Rappel de critères de qualité du HONcode. Réglage du système : langue *anglaise*, algorithme *SVM*, attribut mot unique *w1*, pas de pondération *nnn*

langues, ce réglage de *SVM* montre des résulats intéressants surtout en français.

La table 3 illustre la contingence des valeurs de la précision et du rappel. La contingence est indiquée pour les neufs critères indiqués plus haut. La contingence est présentée pour le réglage du système d'apprentissage que nous avons considéré comme un des meilleurs : algorithme *SVM* appliqué aux mots uniques *w1* sans pondération *nnn*. Il s'agit des tests sur la langue anglaise, où la collection des données est la plus fournie. Ce sont les chiffres de la diagonale de la table qui illustrent la contingence d'un critère donné. Plus les valeurs sont élevées plus le critère en question est aisé à être détecté. Il est ainsi pour les critères *Confidentialité* et *Date*, avec une contingence de 0,92/0,90 et 0,90/0,98 respectivement. Le *DT* n'apparaît pas dans le tableau de résultats car il est trop lent dans son implémentation actuelle.

6 Discussion and Perspectives

Tous les réglages possibles du système (certains d'entre eux sont indiqués dans la table 2) ont été testés avec les données en anglais : ils ne montrent pas de différences notables, sauf l'algorithme *kNN* qui a des performances plus faibles. Nous considérons que la précision, qui correspond au pourcentage des catégorisations correctes effectuées par le système, donne le jugement le plus précis sur ses performances. Et, entre les deux précisions, c'est la micro précision (*miP*) que nous privilégions car elle indique la justesse avec laquelle une phrase est assignée à un critère. Nous considérons donc que c'est sur cette valeur qu'il faut baser le jugement, car c'est cette valeur qui indique la confiance que les réviseurs peuvent accorder au système dans leur travail quotidien. De ce point de vue, l'algorithme *SVM*, lancé sur les mots uniques *w1* sans pondération *nnn*, semble montrer une des meilleures performances : 0,78 de micro précision (la plus élevée), 0,69 du rappel (un des plus faibles), et 0,73 de F-mesure (une des plus élevées). Ce réglage a donc été appliqué aux autres langues, où il montre des performances intéressantes pour les critères avec des données suffisantes. Notons que ce jugement est appuyé, à l'étape actuelle de notre travail, par le fait que ce sont les valeurs de précision qui sont plus faciles à évaluer. Par contre, dans le future, le rappel pourra être privilégié, surtout en prenant en compte que les résultats du système seront validés par les experts. En ce qui concerne la prise en compte de la version micro des mesures (assignation correcte au niveau des phrases), elles sont également plus faciles à maîtriser car influencent moins les résultats globaux.

L'analyse de la contingence (tab. 3) indique que deux critères sont très bien reconnus : *Confi-*

dentialité et *Date*, avec 0,92/0,90 et 0,90/0,98 de contingence respectivement. Cela veut dire, qu'au niveau lexical, où les algorithmes puisent les données, ces critères sont formulés avec des lexiques spécifiques. Par exemple, entre les attributs les plus fréquents de *Confidentialité*, nous avons *identity*, *personal*, *respected*, *individual*, *confidentiality*, *privacy*, qui restent spécifiques à ce critère. Le critère qui reste le plus difficile à reconnaître est *Justification* : il montre une contingence de 0,45/0,33. De plus, il est ambigu avec le critère *Complémentarité*. D'autres couples de critères ambigus sont *Référence* / *Autorité* et *Publicité* / *Sponsor*. En ce qui concerne les critères *Justification* et *Référence*, ils ne cumulent pas une base d'apprentissage suffisante, ce qui peut expliquer les confusions du système d'apprentissage. Par contre, pour le couple *Publicité* / *Sponsor*, la raison principale de la confusion semble provenir du niveau lexical. Ainsi, pour ces deux critères, il est question de *funding*, *maintenance* et *acceptance*.

Comme attendu, à cause du faible nombre de documents catégorisés manuellement sous deux critères, *Justification Référence*, ces critères ne peuvent pas être correctement traités par le système. De ce point de vue, d'autres méthodes peuvent être testées. Par exemple, les mesures de similarité entre les documents (Salton, 1991) comme suggéré par le domaine de recherche d'information. Dans ce cas, la similarité peut être calculée directement avec la définition du critère : la petite taille des données de référence ne constituerait alors pas une limitation. Par ailleurs, il serait intéressant d'effectuer un apprentissage sur les adresses `url` : elles peuvent indiquer quel type d'informations se trouve dans la page en question. Par exemple, si une page porte le nom *privacy.html* ou *policy.html*, cela indique qu'elle peut contenir des informations sur la politique de confidentialité. En outre, la sélection des traits est basée sur la fréquence de documents, tandis qu'il serait intéressant d'essayer pour ceci l'information mutuelle ou Chi2. La combinaison de ces diverses approches et réglages constitue également une perspective de notre travail.

Actuellement, nous effectuons la catégorisation de phrases tandis qu'en réalité les documents entiers devraient être traités. En effet, l'information sur les critères peut être distribuée entre plusieurs phrases dans un document, et même entre plusieurs pages d'un même site. Notons aussi que cette information est noyée dans le contenu de la page. La prise en compte des documents et sites entiers augmente la difficulté de la catégorisation et va sans doute diminuer les performances du système. Mais les premières évaluations montrent que notre système acquiert des données nécessaires pour la catégorisation de page et sites.

Dans sa version actuelle, le système catégorise les phrases selon les critères, mais il n'est pas sensible à la détection du type de contexte, positif ou négatif, dans lequel cette information apparaît. Par exemple, une page peut indiquer que *privacy policy is not respected on the site*, tandis que le système détectera « juste » que la phrase parle du critère *Confidentialité*. Une telle catégorisation est correcte, mais le fait de pouvoir détecter les *nuances* est encore plus important. La détection de tels détails, ou du moins de la négation (Chapman *et al.*, 2001), reste une perspective de notre travail.

Une autre limitation du système provient du fait que les informations données sur les critères peuvent être difficilement vérifiées. Cela veut dire qu'un webmaster peut déclarer qu'il respecte le critère *Confidentialité* alors qu'il vend en réalité les informations sur les utilisateurs. Notons juste que HON a mis en oeuvre un système de plaintes où les trois parties (utilisateur, webmaster et HON) peuvent communiquer, sous anonymat, et résoudre ce type de conflits.

Pour la détection de pages qui parlent des critères du HONcode, nous avons utilisé une base d'exemples positifs. Mais nous pouvons aussi utiliser les exemples négatifs pour détecter les pages qui ne satisfont pas ces critères. Chaque page pourra ainsi être pondérée en fonction de ses poids positif et négatif, et un jugement global pourra ensuite être émis.

Les définitions des critères du HONcode sont actuellement traduites en 32 langues et la certification est adoptée dans le monde entier. Le système appliqué à quatre langues (anglais, français, espagnol et italien) peut être adapté à d'autres langues, si les données d'entraînement suffisantes sont disponibles. De la même manière, la problématique, liée à la qualité et transparence de l'information sur l'Internet, n'est pas propre au domaine de santé mais apparaît aussi dans d'autres domaines. Notre système peut ainsi être entraîné sur des données provenant de ces autres domaines. Quant au domaine médical, le système peut être testé avec d'autres critères de qualité, différents des critères.

7 Conclusion

Nous avons présenté un travail original et ambitieux sur la conception d'un système automatique pour la détection de la qualité et transparence des documents de santé sur l'Internet selon les critères de qualité du HONcode. Le système est basé sur les algorithmes d'apprentissage supervisés. Une première évaluation montre des résultats prometteurs pour la suite : l'algorithme *SVM* appliqué à des mots uniques sans pondération génère 0,78 de micro précision et 0,73 de F-mesure. Le taux d'erreurs reste un des plus faibles (0,06). Ces résultats semblent confirmer la pertinence de notre approche pour la catégorisation des pages de santé sur l'Internet selon les critères du HONcode. Nous avons proposé plusieurs perspectives qui permettront d'apporter des améliorations au fonctionnement, déjà satisfaisant, du système. Dans l'état actuel, le système répond aux attentes réelles ressenties dans le domaine de certification des sites de santé. Les évaluations supplémentaires, par exemple la comparaison de la catégorisation automatique des pages avec la catégorisation manuelle, donnerait des indications plus claires sur la confiance que l'on peut accorder au système.

Remerciements

Ce travail est réalisé dans le cadre du projet PIPS (*Personalised Information Platform for Life & Health*) financé par le programme de la Communauté Européenne *Information society technology* sous le numéro de contrat 507019.

Références

BOYER C., BAUJARD O., BAUJARD V., AUREL S., SELBY M. & APPEL R. (1997). Health on the net automated database of health and medical information. *Int J Med Inform*, **47**(1-2), 27–9.

BOYER C., BAUJARD V. & SCHERRER J. (1999). HONcode : a standard to improve the quality of medical/health information on the internet and HON's 5th survey on the use of internet for medical and health purposes. In *6th Internet World Congress for Biomedical Sciences (INABIS 2000)*.

CARRERAS X. & MÁRQUEZ L. (2001). Boosting trees for anti-spam email filtering. In *Proceedings of RANLP-01, 4th International Conference on Recent Advances in Natural Language Processing*, Tzigov Chark, BG.

CHAPMAN W., BRIDEWELL W., HANBURY P., COOPER G. & BUCHANAN B. (2001). Evaluation of negation phrases in narrative clinical reports. In *Annual Symposium of the American Medical Informatics Association (AMIA)*, Washington.

GROBELNIK M. & MLADENIĆ D. (1998). Efficient text categorization. In *Mining workshop on the 10th European Conference on Machine Learning ECML98*. citeseer.ist.psu.edu/grobelnik98efficient.html.

KOLLER D. & SAHAMI M. (1996). Toward optimal feature selection. In *International Conference on Machine Learning*, p. 284–292.

MANEVITZ L. M. & YOUSEF M. (2001). One-class SVMs for document classification. *Journal of Machine Learning Research*, **2**, 139–154.

NLM (2001). *Medical Subject Headings*. National Library of Medicine, Bethesda, Maryland. http://www.nlm.nih.gov/mesh/meshhome.html.

PAGE L., BRIN S., MOTWANI R. & WINOGRAD T. (1998). *The PageRank Citation Ranking : Bringing Order to the Web*. Rapport interne, Stanford Digital Library Technologies Project.

PANDEY S., ROY S., OLSTON C., CHO J. & CHAKRABARTI S. (2005). Shuffling a stacked deck : The case for partially randomized ranking of search engine results. In K. BRATBERGSENGEN, Ed., *VLDB (Very Large DataBases)*, p. 781–792, Trondheim, Norway.

PORTER M. (1980). An algorithm for suffix stripping. *Program*, **14**(3), 130–137.

SALTON G. (1991). Developments in automatic text retrieval. *Science*, **253**, 974–979.

SELBY M., BOYER C., JENEFSKI D. & APPEL R. (1996). Health on the net foundation code of conduct for medical and health websites. In *MedNet 96 - European Congress on the Internet in Medicine*, Brighton, UK.

SINGHAL A., SALTON G., MITRA M. & BUCKLEY C. (1996). Document length normalization. *Information Processing & Management*, **32**(5), 619–633.

SPERTUS E. (1997). *Smokey : automatic recognition of hostile messages*, In *American Association for Artificial Intelligence*.

VINOT R., GRABAR N. & VALETTE M. (2003). Application d'algorithmes de classification automatique pour la détection des contenus racistes sur l'internet. In *TALN*.

WANG Y. & LIU Z. (2006). Automatic detecting indicators for quality of health information on the web. *International Journal of Medical Informatics*.

WILLIAMS K. & CALVO R. A. (2002). A framework for text categorization. In *7th Australian document computing symposium*.

YANG Y. & LIU X. (1999). Re-examination of text categorisation methods. In *Proc of 22nd Annual International SIGIR*, p. 42–49, Berkley.

Adresse de correspondance

Arnaud Gaudinat
Fondation Health on the Net
Hôpitaux Universitaires de Genève
24, rue Micheli-du-Chrest
CH-1211 Genève 14
Suisse
arnaud.gaudinat@healthonnet.org