Quel est l'impact de la qualité des documents dans un système de Question / Réponses?

Sarah Cruchet, Arnaud Gaudinat, Thomas Rindflesch, Célia Boyer

Sarah Cruchet
Fondation Health On the Net
81 Boulevard de la Cluse
1206 GENEVE

ABSTRACT

Objectives: Search queries make use of keywords to produce results, where usually, the user has to search each page for the required answer and frequently, it is difficult to judge the quality of a document on the Internet.
This study demonstrates a question answering system (QA system) which can address the above mentioned issues.
Design: The QA system was evaluated in two ways during this study; the first studied the effectiveness of the system while the second focused on the quality selection of both HONcode and other websites using the QA system.
Measurements: The results were classed by a medical expert as follows: A = relevant, B = lead to an answer, C = not relevant following which, Trec eval was used for the evaluation of our system.
Results: Systematic (1st) evaluation: 70 out of 100 questions were answered well. Qualitative (2nd) evaluation: A MAP of 59% and a MRR of 76% was obtained for QAHON_honcode with a set of 100 questions.
Conclusion: The results obtained showed that the quality and reliability of the answers obtained by the QA system, is impacted by the trustworthiness of the database used.

Introduction

Le nombre de documents, pour une requête donnée, est en constante augmentation mais la recherche d'information pour l'individu lambda, elle, n'a pas changée! L'internaute a toujours du mal à évaluer la pertinence et la qualité des résultats et il doit parcourir l'ensemble du document avant de pouvoir espérer accéder à une réponse satisfaisante. De plus, poser une question par mots clés n'est pas naturel [1, 2] et ne permet pas d'obtenir des résultats précis. Si certains misent sur l'avènement du Web sémantique pour produire de nouveaux moteurs de recherche intelligents, d'autres espèrent faciliter l'interaction homme / moteur de recherche en proposant des systèmes de Question / Réponses (QR) basés principalement sur des heuristiques et l'utilisation de la redondance de l'information. Le but de tels systèmes est de fournir des réponses pertinentes à une question posée en langage naturel. Ainsi l'utilisateur accède directement à l'information souhaitée et n'a pas à parcourir l'ensemble des documents proposés par le moteur de recherche. A l'heure de la mobilité, des réponses courtes et précises avec par exemple une interface vocale, semblent aller dans le sens de l'intérêt de ce genre de système.
De plus, des campagnes d'évaluation comme le TREC [3] ont montré la faisabilité et l'intérêt des QR. Les meilleurs d'entre eux atteignent des précisions sur les réponses de l'ordre de 71% en 2005 [4]. En ce qui concerne le domaine médical, l'intérêt est grandissant si l'on considère le nombre de papiers dans ce domaine [5-8]. Actuellement un système de QR appliqué au domaine médical est en ligne. Il s'agit de MedQA [9], système qui traite les questions de type définition et qui est destiné aux biologistes pour faciliter leur accès à la littérature biomédicale.
Par ailleurs, la création d'un site Internet est à la portée du tout un chacun, d'où la variabilité de l'information trouvée en ligne. Dans le cas des informations de santé il y a un impact direct sur le bien être des individus [10]. Ainsi des initiatives telles que le HONcode [11-13] (code de bonne pratique des sites Internet de santé composé de 8 principes éthiques) et des outils de recherche verticaux

tentent de proposer des services complémentaires et alternatifs aux moteurs de recherche généralistes. Le HONcode a été développé avec le consensus des responsables de sites Web, par la Fondation Health On the Net (HON), fondation à but non lucratif ayant pour objectif de promouvoir une information médicale de qualité et de confiance. Les études actuelles effectuées sur les QR ne s'intéressent qu'à la pertinence des réponses proposées. La Fondation HON se positionne sur la qualité de l'information médicale. Ainsi elle développe un QR appliqué au domaine médical où la recherche de la réponse s'effectue dans l'ensemble des sites certifiés par la Fondation.

Ce papier présente deux évaluations distinctes du QR développé par la Fondation HON dans le cadre du projet européen PIPS (IST–2002-2.3.1.11) [14]. La première évaluation est systématique et permet de valider l'efficacité du système présenté. La seconde évaluation met en avant l'utilisation du QR sur une sélection de sites de santé de qualité d'une part (ceux du HONcode) et sur tout le Web d'autre part. Pour chaque évaluation le matériel et les méthodes seront présentés. Puis nous présenterons les résultats obtenus, résultats qui seront discutés en indiquant quelques pistes pour mieux intégrer la dimension qualitative dans les systèmes de QR. Enfin nous conclurons et présenterons les perspectives futures.

Matériels et Méthodes

Nous utilisons le système de QR multilingue de HON présenté dans [15]. Nous allons résumer dans ce paragraphe l'architecture de notre QR. Il s'appuie sur un module de détection du type de question par apprentissage et l'utilisation des ressources sémantiques telles que celles présentes dans UMLS pour guider le système, notamment dans le choix des réponses pertinentes. Ce système dispose d'une architecture relativement classique pour un système de QR avec des spécificités propres pour chaque module : 1/ Le QuestionAnalyzer a pour but d'identifier la question de l'utilisateur afin de mieux anticiper la recherche d'information et la sélection de la réponse. 377 questions du domaine médical ont été utilisées pour mieux définir les questions et ainsi prendre en compte la spécificité médicale. 2/ Le QueryGenerator permet de générer une ou plusieurs requêtes en fonction des différents moteurs de recherche et de la question analysée précédemment. 3/ Le module DocumentRetrieval s'occupe d'interroger les différents moteurs de recherche. 4/ Le module PassageExtractor identifie les meilleurs passages en fonction de la question posée et de la réponse attendue. 5/ Le module AnswerExtractor qui s'occupe de sélectionner les meilleures réponses dans les passages en fonction de la question, de la réponse attendue et de la redondance de certaines réponses. 6/ Un module d'affichage qui permet de mettre en valeur les différentes réponses trouvées de manières très épurée avec la possibilité de tracer au mieux l'information dans son contexte. A noter qu'un dernier module original a été ajouté récemment qui a pour but de réaliser une analyse globale des résultats de la recherche (donc intercalé entre le DocumentRetriever et le PassageRetriever), afin d'assister encore mieux la sélection des passages et des réponses pertinentes. Ce module est basé sur une analyse Ngram de mots de la totalité des documents retournés, filtrés par les types sémantiques attendus.

Evaluation systématique

Pour la première évaluation de notre système, nous disposons d'un corpus de 377 questions en anglais ainsi que des 12 sites d'où elles proviennent. Ces questions ont été récoltées manuellement sur Internet dans des forums de discussion et Foire Aux Questions (FAQ) traitant du domaine médical. Ce corpus a été constitué pour une tâche d'apprentissage automatique dans le cadre de l'identification du type de question du module QuestionAnalyzer de notre système QR [15]. Le domaine médical étant très vaste nous nous sommes concentrés sur les questions relatives aux maladies, sujet qui concernent la plupart des requêtes des Internaute à 64% [16]. De plus nous avons choisi une maladie en particulier, afin d'éviter le biais du thème médical dans l'apprentissage du système. C'est le diabète qui a été choisi car ce sujet était utilisé dans le projet PIPS [14] qui intégrait les premières versions de notre système de QR. Cependant, dans cette étude ce n'est pas seule la question qui nous intéresse mais aussi sa réponse. Ainsi un sous-ensemble de 100 questions avec leurs réponses associées a été constitué manuellement pour ces 12 sites, afin de créer un corpus de référence. Les 12 sites ont été téléchargés en 2008, ce qui correspond à 914 pages différentes, et ont été indexés localement afin de palier aux variations de contenus inhérentes au dynamisme du Web.
Voici 3 exemples de questions et d'une réponse associée :
Q1: Who is at greatest risk of developing type 2 diabetes?
Url1:http://www.nutritionaustralia.org/Food_Facts/FAQ/diabetes_detailfaq.html
A1: Until recently, type 2 diabetes was considered as the 'adult onset' form--that is, it was rarely seen other than in middle-aged and older people. However, it is now affecting younger people as well.

Q2: What is the metabolic syndrome?
Url2: http://www.idf.org/home/index0689.html
A2: The metabolic syndrome is a cluster of the most dangerous heart attack risk factors: diabetes and prediabetes, abdominal obesity, changes in cholesterol and high blood pressure.
Q3: Are there different types of diabetes ?
Url3:http://www.health24.com/medical/Condition_centres/777-792-808-1535,28618.html
A3: Yes, this is true. There are two types of diabetes - Type 1 and Type 2

Evaluation qualitative

Notre étude s'appuie sur la méthode d'une étude conduite à la National Library of Medicine (NLM), Bethesda, USA, sur les systèmes de recherche d'information [5]. Elle est ici basée sur la comparaison de 2 versions de QR développés par HON. Le premier système, QAHON_honcode est celui basé sur la recherche exclusive dans la base de données HONcode (soit 6'800 sites). Le second, QAHON_google, est l'utilisation des résultats de Google via notre système de QR (soit la totalité du Web indexée par Google). Seule la partie DocumentRetrieval diffère entre QAHON_honcode et QAHON_google.
Dans le cadre d'un QR, et au contraire de l'étude effectuée par la NLM, nous évaluerons les réponses retournées par le système plutôt que les documents.

Pour la seconde évaluation nous avons utilisé la méthode décrite dans [5] et réutilisé une partie de notre corpus de questions de la première évaluation. Ainsi pour chaque système seules les 10 premières réponses au maximum sont retournées. Ces réponses sont mélangées et « anonymisées » - pas d'information sur la provenance de la base de données - et soumises au jugement d'un expert qui a pour mission de noter les réponses à l'aide de lettres : A+ (très pertinent), A (pertinent), A- (pas toute la réponse), B+ (mène à la réponse), B (peut mener à la réponse), B- (pas clair), C (pas pertinent). La notation correspond à la pertinence de la réponse par rapport à la question ainsi que de la pertinence du document d'où est extraite la réponse (dans le cas où la réponse retournée par le système n'est pas claire l'expert va regarder si le document source comporte la réponse).
Nous avons utilisé le logiciel TREC eval [17] pour évaluer nos résultats. Six mesures ont été prises en compte pour évaluer le système :

1. Mean Average Precision (MAP) : calcule la précision moyenne après chaque document extrait
2. Binary Preference (Bpref): calcule le taux de réponses non pertinentes affichées avant les réponses pertinentes.
3. R precision (R-pre): calcule la précision moyenne après R réponses extraites.
4. Mean Reciprocal Rank (MRR): taux de réponses pertinentes au 1er rang.
5. Precision at five documents (P@5): taux de réponses pertinentes au rang 5.
6. Precision at ten documents (P@10): taux de réponses pertinentes au rang 10.

Ces mesures ont été calculées sous deux conditions : "soft" et "strict" comme dans [5]. Pour la première ("soft") nous avons considéré qu'une réponse est pertinente si elle est notée A ou B. Pour la seconde ("strict"), seulement les réponses ayant obtenues un A ont été considérées comme pertinentes.

Résultats

Evaluation systématique

Nous recherchons la capacité du système à retrouver les réponses récoltées manuellement dans les 12 sites d'où proviennent les questions de la base de test, soit un corpus de 914 pages.
Par un simple test de "pattern matching" (comparaison de chaîne de caractères), nous avons considéré que le système donnait une réponse correcte lorsqu'une réponse du système était contenue dans la sélection manuelle des réponses. Ainsi le système obtient un taux de bonne réponse de 80%. Ce résultat est positif même s'il n'est malheureusement pas un indicateur pour la précision, car les réponses du système qui ne font pas partie du corpus de référence ne sont pas comptabilisées alors qu'elles peuvent être justes. En effet le système trouve des réponses que la personne n'a pas récoltées manuellement. De plus SelectQA trouve en moyenne 3.2 réponses par questions alors que manuellement 2 réponses maximums par questions ont été collectées.

Evaluation qualitative

Système	QAHON_honcode	HONQA_goole
MAP	**0.59**	0.36
Bpref	**0.50**	0.34
R-pre	**0.59**	0.38
MRR	0.76	**0.86**
P@5	**0.54**	0.36
P@10	**0.32**	0.22

Tableau 1. Résultats de l'évaluation soft

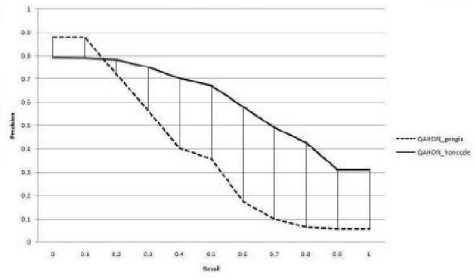

Figure 1. Courbe de précision / rappel

La première observation qui peut être faite est que les résultats sont meilleurs pour QAHON_honcode que pour QAHON_google. En effet, QAHON_honcode a une MAP de 0.59 contre 0.36 pour QAHON_google. Cela signifie que sur l'ensemble des réponses données par QAHON_honcode pour une question, 59% sont pertinentes contre 36% pour QAHON_google. De plus, sur les 5 premières réponses trouvées par QAHON_honcode, plus de la moitié répondent exactement à la question posée. La seule mesure pour laquelle QAHON_google est plus performant que QAHON_honcode est MRR. La première réponse pertinente a un meilleur rang pour QAHON_google que pour QAHON_honcode.

Figure 1 donne le rapport précision / rappel pour les 2 systèmes, selon l'évaluation soft. On constate que QAHON_google est très performant aux 2 premiers rangs puis que QAHON_honcode passe en tête. Cela signifie que sur l'ensemble des réponses données par les systèmes, QAHON_honcode est globalement plus pertinent que QAHON_google.

Discussion sur l'évaluation

Nous avons utilisé 6 mesures pour évaluer notre système : 3 générales (MAP, Bpref et R-prec) et 3 plus précises (MRR, P@5 et P@10). Si l'on considère l'évaluation "soft", avec une MAP et une R-prec de 59%, QAHON_honcode est performant pour obtenir une vue d'ensemble d'un sujet particulier. De plus, sur un échantillon de 100 questions nous constatons que les réponses de QAHON_honcode sont meilleures que celles de QAHON_google. En effet les réponses provenant de sites certifiés ont été mieux notées par l'expert (avec un fort taux de A et de B). Les réponses retournées par QAHON_honcode sont donc, de meilleures qualités avec 59% de réponses pertinentes par question.

Discussion sur la qualité de l'information dans le cadre des systèmes QR

La Figure 1 démontre que les réponses extraites des documents HONcode sont de meilleures qualité que celles extraites uniquement de Google car les résultats de QAHON_honcode dépassent ceux de QAHON_google à partir du rang 3.

Utilisation de ressources fiables

Dans cet article nous avons présenté l'utilisation de ressources dignes de confiance différentes dans le cadre d'un système QR. L'hypothèse est que l'utilisation de documents sélectionnés par rapport à des critères éditoriaux favorise la qualité des réponses du système de QR (à l'instar d'un système de recherche d'information classique sur le Web). De plus, suivant le mode de sélection, politique éditoriale, certification tierce, sélection par des professionnels ou rating collaboratif, se pose la question de la pondération des différentes ressources et de sa combinaison avec le score de pertinence sémantique.

Utilisation de la date comme indice de qualité de l'information

La date est un critère important de la qualité de l'information à prendre en compte dans un système de QR. Un système QR de qualité devrait prendre en compte cette dimension au sein de sa stratégie de classement ou de filtrage des réponses et du moins donner les indications de date de rédaction pour chaque réponse présentée à l'utilisateur final. En effet une déclaration peut être juste à un moment donné et être dépassée quelques années ou mois plus tard.

Utilisation de la redondance de l'information

La redondance de l'information est l'une des stratégies principales des systèmes de QR actuels pour l'identification de réponses pertinentes. Du point de vue de la qualité de l'information, l'hypothèse peut être la même, puisqu'une même information répétée sur plusieurs sites distincts a certainement plus de chances d'être crédible.

Web sémantique et la qualité de l'information

Evoqué dans l'introduction, le prometteur Web sémantique à tout pour séduire et nous offre des possibilités énormes quant à la pertinence de la recherche d'information (aussi dans le cas QR). Cependant là encore il faudra être vigilent, car ceux qui produiront ou collecteront cette information auront des objectifs qui risquent d´être forts différents allant de la philanthropie au mercantile pur et dur. En effet les informations du Web sémantique seront certainement encore moins lisible en terme de traçabilité et seul des garde-fous d'organisation tierce pourront garantir une certaine neutralité ou du moins d'obtenir des détails indépendants concernant cette information.

Limitation de l'étude

L'évaluation a été réalisée sur un corpus de 100 questions et par une seule personne en raison du travail fastidieux que représente la notation manuelle des réponses.
Ensuite, dans la partie du système QR qui récupère les passages pertinents nous effectuons un nettoyage de la page Web. Parfois ce nettoyage est trop fort et enlève de l'information pertinente. Cependant cette étape est nécessaire car elle retire le bruit de la page Web (en-tête et menus).
Le lecteur avisé est en droit de se demander si cette étude met en lumière la distinction ressource médicale (HONcode) vis-à-vis de ressource générale (Google) plutôt que ressource fiables vis-à-vis de ressource non-contrôlée. Cependant les auteurs pensent que les questions utilisées dans cette étude sont suffisamment spécifiques pour ne donner que des réponses de type médical. Et en effet, une lecture aléatoire d'un échantillon de réponses nous a confirmé cette hypothèse.

Conclusion

Les solutions proposées pour prendre en compte la variabilité de la qualité de l'information médicale sur le Web sont justifiées aussi bien dans la recherche d'information classique que dans le cadre d'utilisation de systèmes de QR.
L'utilisation de métas-informations telles que la prise en compte de la date est ici cruciale.
La redondance de l'information, bien qu'étant un indicateur brut, peut être un très bon indice de pertinence qualitative si on s'assure de le retrouver dans des ressources sûres et indépendantes. D'un autre côté un bon système de QR se doit aussi de mettre en exergue la pluralité des réponses.
Dans l'interface, l'utilisateur doit pouvoir à tout moment avoir accès à la réponse dans son contexte pour une meilleure traçabilité de la réponse.

References

1. C. Kwok, & AL, Scaling question answering to the Web, In Proceedings of WWW'10*, 2001*
2. D. ROUSSINOV, How Question Answering Technology Helps to Locate Malevolent Online Content, Intelligence and Security Informatics volume 3495/2005, 2005.
3. http://trec.nist.gov/
4. E. M. Voorhees and Hoa T. Dang. 2006. Overview of the TREC 2005 question answering track. In Proceedings of the Fourteenth Text REtrieval Conference (TREC 2005).
5. C.A. SNEIDERMAN & AL Knowledge-Based Methods to Help Clinicians Find Answers in MEDLINE, JAMIA 2007
6. D. DEMMER-FUSMAN & AL, Answering Clinical Questions with Knowledge-Based and Statistical Techniques, Computational Linguistics 2007
7. D. DEMMER-FUSMAN & AL, Combining resources to find answers to biomedical questions, TREC 2007
8. J. Lin & AL, Semantic Clustering of Answers to Clinical Questions, AMIA 2007
9. Y. HONG & AL, Development, implementation, and a cognitive evaluation of a definitional question answering system for physicians, Journal of Biomedical Informatics, 2007
10. E. HORVITZ, Cyberchondria: Studies of the Escalation of Medical Concerns in Web Search, Microsoft study, 2008
11. M. SELBY & AL, Health On the Net Foundation Code of Conduct for Medical and Health Websites. MedNet 96 - European Congress on the Internet in Medicine, Brighton, U.K., Oct. 14 to 17, 1996
12. C. BOYER & AL, Health On the Net foundation: assessing the quality of health web pages all over the world, MedInfo, 2007
13. http://www.hon.ch/HONcode/Conduct.html
14. http://www.pips.eu.org/
15. S. CRUCHET & Supervised approach to recognize question type in a QA system for health, MIE, 2008
16. S.FOX, http://www.pewinternet.org/pdfs/PIP_Online_Health_2006.pdf, 2006
17. http://trec.nist.gov/trec_eval/index.html