Evaluation d'une interface de restitution de résultats de recherche : Quelles conclusions en tirer ?

Max CHEVALIER[α,β], Gilles HUBERT[α]

[α] Institut de Recherche en Informatique de Toulouse (IRIT),
UPS, 118 route de Narbonne, 31062 Toulouse cedex 04

[β] Laboratoire de Gestion et de Cognition (LGC),
115 avenue de Rangueil, 31077 Toulouse cedex

http://www.irit.fr/~Max.Chevalier
http://www.irit.fr/~Gilles.Hubert

{Max.Chevalier, Gilles.Hubert}@irit.fr

Résumé : La restitution des résultats de recherche aux usagers est un point crucial des outils de recherche d'informations. Bon nombre de métaphores, de visualisations ont été proposées, chacune d'elles ayant été évaluée mais aucune conclusion tangible n'a pu être réellement tirée de ces travaux. Certes l'interface est dite facile d'utilisation, conviviale… Qu'en est-il du réel impact de cette interface sur les résultats de recherche ? A quelle(s) tâche(s) correspond-elle le mieux ? A quel public (débutants ; experts) ? Comment peut-on comparer les résultats obtenus entre deux interfaces ? Le but de cet article n'est pas de définir un cadre d'évaluation strict des interfaces de restitution de résultats de recherche mais de s'interroger sur le travail qu'il reste à accomplir pour atteindre des résultats cohérents et surtout comparables. A partir d'une étude d'une interface de restitution et des nombreux travaux en termes d'évaluation, nous discutons de la nécessité d'un programme d'évaluation expérimental permettant de caractériser plus finement les travaux proposés et surtout de comparer objectivement les interfaces. Cet article ne donne pas de solution clés en main mais seulement quelques pistes éventuellement à suivre.

1 Introduction

Les outils de recherche adhoc génèrent un grand nombre de résultats que l'utilisateur doit parcourir un à un au travers de la classique liste de résultats pour vérifier leur pertinence par rapport à ses besoins. Dans la plupart des cas, l'utilisateur se limite aux premiers documents pertinents occultant ainsi de nombreux documents potentiellement pertinents du fait de la charge cognitive que cela requiert. D'autre part, l'interprétation de cette liste est difficile du fait de sa relative pauvreté en information. Par exemple, cette liste de résultat ne permet pas d'apprécier la similarité des documents par rapport à la requête. Il est donc nécessaire, dans

certains cas, d'avoir une vision globale du résultat de recherche permettant à l'utilisateur d'identifier, dans l'ensemble des résultats, les documents les plus pertinents pour ses besoins. Nous présentons dans la section suivante une interface graphique nommée Easy-DoRView pour laquelle nous décrivons et discutons de l'évaluation qui en a été faite.

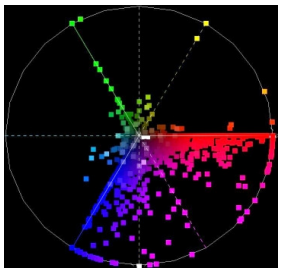
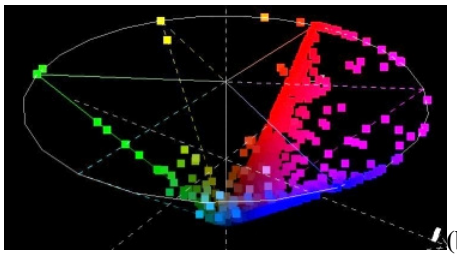
2 Etude de cas : l'interface Easy-DoRView

Cette première partie présente les travaux menés concernant une interface de visualisation en 3D pour les résultats de recherche nommée Easy-DoRView (Chevalier, 2000), (Chevalier, 2001). Cette interface propose de visualiser les documents retrouvés par un moteur de recherche en fonction de l'importance des mots-clés de la requête. Elle repose sur deux aspects : une répartition spatiale 3D spécifique ainsi qu'une utilisation des couleurs (dégradés particulièrement). Elle exploite le modèle de couleur HSV (Smith, 1978) pour représenter les documents dans l'espace. Les différents documents retrouvés sont présentés sous forme de points colorés dans l'espace. Chaque axe de visualisation (au nombre maximum de trois) est associé à une couleur de base (rouge, vert ou bleu). Chaque axe peut représenter un mot-clé de la requête ou une combinaison (par des opérateurs booléens) de mots-clés. La couleur ainsi que la localisation d'un point rend compte d'une combinaison spécifique des axes de visualisation. L'interprétation de ces points se fait en deux temps :

- L'identification de l'importance relative des axes les uns par rapport aux autres. En effet, un point se situe proche du ou des axes prédominants. Ceci repose sur un principe d'attraction pondérée selon l'importance relative de chaque axe dans un document,
- L'identification de l'importance réelle du ou des critères les plus prédominants. Dès lors que l'individu connaît l'importance relative des axes les uns par rapport aux autres, il lui faut connaître l'importance réelle de l'axe prédominant (ou des axes prédominants s'il y en a plusieurs). Cette valeur correspond à la hauteur du point dans le cône. Ainsi, plus un point se situe vers le sommet du cône, plus l'importance du ou des axes prédominants est faible (=0). A l'opposé plus un point se situe vers la base du cône plus l'importance du ou des axes prédominants est forte (=1).

Après une première série de manipulations, nous avons pu nous rendre compte que ce premier modèle souffrait d'un problème d'interprétation concernant les documents se trouvant vers le sommet du cône. En effet, du fait que les documents se situent dans un endroit où l'espace se restreint, il n'est plus possible de distinguer une distance entre les points. En réponse à cela, nous avons modifié notre modèle et proposé un modèle en cylindre nommé HSV* qui repose sur la même interprétation mais qui ne souffre plus de ce problème de zones ayant une haute résolution.

La Figure 1 montre quelques exemples de visualisation de résultats de recherche par le biais de cette interface. Grâce à cette interface de visualisation de résultats de recherche, l'usager peut comprendre l'importance des différents mots-clés de sa requête dans les documents retrouvés. Cette présentation lui permet, en se focalisant par exemple sur une partie de l'espace spécifique, de re-pondérer par exemple les termes de sa requête.

(a   (b

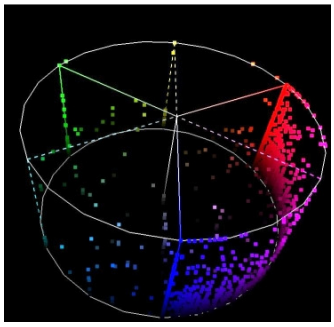
 (c

Figure 1 – Quelques exemples de visualisation (a : cône vu de dessus, b : cône vu de ¾, c : cylindre vu de ¾)

Pour permettre à l'usager de naviguer et manipuler ces résultats, différents outils[1] sont offerts tels que la sélection et la suppression par sélection de zone avec la souris, le zoom, la sélection fine en fonction des critères que l'on veut voir apparaître dans les documents, la visualisation de contenu de documents…

3 Evaluation de l'interface Easy-DoRView

Cette section vise à présenter la phase d'évaluation réalisée pour cette interface et de présenter les résultats obtenus. Le but de cette évaluation était de vérifier la validité des hypothèses sous-jacentes à l'interface proposée c'est-à-dire de deux manières (localisation spatiale et couleur) de visualiser la même information concernant les documents (importance des axes de mots-clés de la requête). Afin d'évaluer l'intérêt de la visualisation que nous proposons, nous avons mis en place une application de test. Les aspects que nous avons souhaité souligner dans cette évaluation sont l'intérêt de la combinaison de deux axes d'interprétation (couleurs, localisation spatiale), ainsi que les aspects cognitifs de la visualisation proposée.

Détail de la tâche d'évaluation

L'application dédiée à l'évaluation a pu soit être exécutée au sein de notre laboratoire de recherche (*évaluation locale*), soit être téléchargée et exécutée à domicile par exemple

[1] Un détail des outils offerts par l'interface est proposé à l'adresse http://www.irit.fr/~Max.Chevalier/easydor.htm

(*évaluation distante*). Dans ce dernier cas, le résultat du test était retourné par courrier électronique.

Durant la tâche d'évaluation, nous avons assisté chaque participant local. Un descriptif de la méthode d'évaluation ainsi que l'ensemble des questionnaires à rendre étaient également fournis comme le questionnaire personnel qui nous permet de mieux connaître les participants (état civil, connaissances informatiques…).

La phase d'évaluation se décompose en trois parties :

- une première partie visant à mettre en évidence les aspects cognitifs de l'interface (en particulier l'aspect intuitif). Cette partie confronte les participants à l'interface sans aucune information préalable. Le résultat de cette phase est un questionnaire ouvert au travers duquel les utilisateurs doivent indiquer le but de l'interface, les outils disponibles etc,
- une deuxième partie visant à mettre en évidence l'intérêt de combiner plusieurs axes de visualisation pour favoriser l'interprétation exacte des documents représentés. Un participant est confronté à une série de 20 mêmes points correspondant à une combinaison des critères. Ces points sont présentés dans un ordre aléatoire au participant selon différents points de vue (avec un point de couleur uniquement, avec un point blanc dans l'espace 3D HSV, avec un point de couleur dans l'espace 3D HSV et les même tests dans l'espace 3D HSV*). Pour chacun des points présentés, le participant doit indiquer la valeur estimée de chacun des critères,
- une troisième partie est destinée à évaluer la satisfaction de l'utilisateur et à vérifier l'utilisation réelle de l'interface grâce à un cas concret de recherche.

Détail des participants à l'évaluation

12 personnes, n'ayant pas participé au développement de l'interface ont accepté de participer à cette évaluation. Ce nombre est relativement faible mais il permet tout de même d'obtenir une première appréciation de l'interface. Le panel de testeurs est mixte bien que majoritairement composé d'hommes (2/3 hommes et 1/3 femmes). L'âge moyen des participants est de 28 ans ce qui correspond globalement à l'âge moyen des internautes (21-35 ans) (GVU, 1998). L'âge des participants varie cependant de 20 à 52 ans (cf tableau 1). Chaque participant est identifié par un numéro. Les numéros des participants à une évaluation distante sont précédés du symbole « X ».

ID	***1***	***2***	***3***	***4***	***5***	***6***	***7***	***8***	***X1***	***X2***	***X3***	***X4***
Sexe	*M*	*F*	*M*	*F*	*F*	*M*	*M*	*M*	*M*	*M*	*F*	*M*
Age	*20*	*25*	*23*	*38*	*25*	*29*	*34*	*23*	*24*	*21*	*52*	*24*

Tableau 1 – Panel des participants

En ce qui concerne les connaissances des participants concernant l'outil informatique et le monde de la recherche d'information, chacun d'entre eux avait déjà utilisé l'outil informatique ainsi que des outils de recherche.

Résultats

Partie 1 : aspects cognitifs

Cette première partie nous a permis de mettre en évidence si les utilisateurs comprenaient l'intérêt de l'interface ainsi que les fonctionnalités offertes après une phase d'utilisation limitée (intuitivité). Le dépouillement du questionnaire souligne que tous les utilisateurs ont globalement compris l'intérêt de l'interface. De plus, ils ont globalement identifié les différentes fonctionnalités offertes. Un détail de ce questionnaire est disponible dans (Chevalier, 2002).

Partie 2 : combinaison des axes d'interprétation

Le dépouillement des résultats de cette phase nous a permis d'apprécier l'impact de la combinaison des deux axes d'interprétation.

Pour chaque échantillon présenté, le système a sauvegardé les valeurs des trois critères saisies par l'utilisateur. A partir de ces informations enregistrées par le système, nous avons évalué si un point dans l'interface était correctement interprété c'est à dire si un utilisateur pouvait, sans aucune aide extérieure, indiquer l'importance des critères correspondant à un échantillon. Pour chaque échantillon évalué, nous avons également mesuré la distance euclidienne entre les valeurs réelles des critères dans l'échantillon et les valeurs saisies par l'utilisateur. Cette distance correspond à l'erreur commise par l'utilisateur.

Les distances euclidiennes calculées au cours des différentes phases, pour l'ensemble des utilisateurs, est présentée dans la figure 2. En plus de ces distances, nous présentons dans la figure 3, le pourcentage moyen d'erreur des participants. Cette erreur correspond à la moyenne des distances entre les échantillons et les valeurs saisies par rapport à la distance maximum.

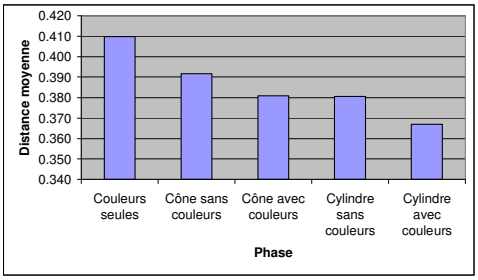


Figure 2 - Distance moyenne entre les échantillons réels et les valeurs saisies par les participants pour les différentes phases

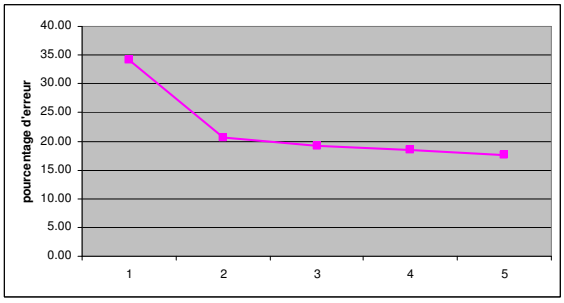


Figure 3 – Pourcentage d'erreur moyen entre les échantillons réels et les valeurs saisies par les participants pour les différentes phases

Cette étude a mis en évidence que la combinaison des deux axes d'interprétation (couleur et localisation spatiale) permet d'obtenir de meilleurs résultats qu'avec seulement un seul axe. En effet, les visualisations en cône ou en cylindre avec couleurs permettent de réaliser une erreur plus faible de 3% en moyenne que les visualisations en cône et en cylindre sans couleur. En utilisant le cylindre avec les couleurs, les participants ont réalisé une erreur inférieure de 10.5% en moyenne par rapport à l'utilisation des couleurs seules. Compte tenu qu'aucune explication ou résultat intermédiaire n'était fourni à l'utilisateur, la baisse de l'erreur résulte donc uniquement des visualisations proposées et non d'une expérience plus grande des utilisateurs.

Cependant, la figure 4, représentant le temps moyen de réponse par échantillon, souligne le fait que le temps augmente selon les phases d'évaluation. L'interprétation de cette augmentation est difficile à dire car nous ne savons pas si celle-ci est due à la fatigue ou à une complexité croissante des visualisations. Au regard de l'augmentation quasi linéaire du temps nous penchons plus vers la fatigue de l'utilisateur (la tâche d'évaluation durait entre 1h30 et 2h pour évaluer 100 échantillons).

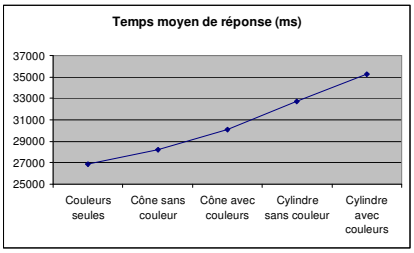


Figure 4 -Temps moyen de réponse

Partie 3 : satisfaction de l'utilisateur

Le résultat du dépouillement des questionnaires de satisfaction subjective est présenté dans la figure 5. Les réponses des différentes questions sont données sur une échelle de valeur de 1 à 9, indiquant le degré de satisfaction de l'utilisateur (9 étant la satisfaction la plus élevée).

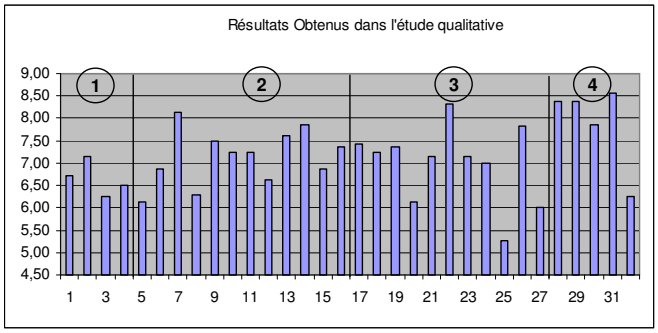


Figure 5 - Résultats de l'étude qualitative

Les réponses sont présentées par rapport au numéro de la question. Les quatre zones correspondent aux différentes parties du questionnaire :

1 – les réactions générales vis-à-vis du système (effet à priori…),

2 – l'écran (présentation, lisibilité des polices de caractères…),

3 – l'apprentissage (nombre d'actions nécessaires…),

4 – le système du point de vue général.

Dans un premier temps, nous pouvons souligner le fait que les réponses sont toutes supérieures à la moyenne ce qui indique que les participants étaient globalement satisfaits de l'interface. Cependant, nous pouvons remarquer que :

- les utilisateurs déplorent le nombre trop important d'étapes à réaliser pour obtenir le résultat escompté (question 25),
- l'achèvement de la tâche n'est pas réellement identifié par l'utilisateur (question 27).

Ces constats doivent être nuancés car le questionnaire a été, nous l'avons remarqué, rempli avant de réaliser l'étude de cas dans la plupart des cas. Les valeurs fournies par les participants se basent donc principalement sur une intuition et non par sur une réelle utilisation de l'interface. Par contre, nous pouvons souligner que l'aspect intuitif de l'interface est assez bien apprécié (questions 12 et 13). De plus, l'utilisation des couleurs sur laquelle repose l'interface est assez bien perçue (questions 14, 15 et 16). Par ailleurs, cette étude souligne que l'appréciation de l'apprentissage est bonne (question 20) même si certains trouvent qu'il est trop long.

Enfin, nous pouvons constater que les performances de l'interface, du point de vue des participants, sont satisfaisantes (questions 28 et 29).

Les remarques libres fournies par les participants nous ont également permis de faire évoluer la visualisation proposée. Par exemple, l'aspect abstraction a été amélioré grâce au clignotement des points correspondant aux documents sélectionnés.

Concernant l'étude de cas, deux exercices de repérages sont proposés. Ils ont pour but de vérifier si l'individu identifie correctement les documents les plus pertinents relatifs à une requête donnée. Dans ces exercices, l'utilisateur a accès aux deux visualisations en couleur (cône et cylindre) proposées. Dans ce cadre, 75% des participants ont réussi à identifier les documents correspondant aux critères du premier exercice. Ces résultats sont encourageants car cela permet de mettre en évidence que les participants ont globalement compris l'interprétation des résultats. Cependant, un deuxième exercice de repérage relatif à des

documents situés dans le sommet du cône confirme la remarque faite dans (Ware, 1985). Elle concerne le problème de la hausse de la résolution lorsque les valeurs sont faibles dans le modèle HSV. En effet, la plupart des utilisateurs ont utilisé la visualisation en cône et seulement 12.5% des participants ont réussi à identifier des documents répondant aux critères. Malgré tout, la plupart des participants savaient où se trouvaient les documents mais n'ont pas réussi à les identifier du fait du grand nombre de documents situés dans cette zone. Ce point nous suggère que l'aspect « affordance » (l'utilisateur doit comprendre quelle interface est la plus adaptée à ses besoins) de l'interface n'est pas optimale car ces mêmes documents apparaissaient de façon plus significative dans la visualisation en cylindre.

Bilan sur l'interface de visualisation

Du fait de la répartition des documents (grâce aux couleurs et à leur localisation spatiale), l'interface permet d'apprécier visuellement l'importance des différents mots-clés de la requête au sein de l'ensemble des documents retrouvés. Cependant, certaines fonctionnalités comme la sélection fine (pour répondre au problème de résolution élevée soulevé par (Ware, 1985)) ont été introduits suite à la phase d'évaluation. Cette évaluation nous a également permis de mettre en évidence que les visualisations proposées permettent d'interpréter les résultats présentés de façon plus précise. En effet, la combinaison de deux axes d'interprétation apporte une meilleure interprétation qu'avec un axe unique. Ces résultats sont encourageants puisque cela nous laisse penser que les performances des utilisateurs (temps, qualité d'interprétation) progresseront avec l'expérience liée à l'utilisation de l'interface. Par ailleurs, la satisfaction des participants nous encourage dans le développement de l'interface proposée. Du point de vue des utilisateurs, la phase d'évaluation nous permet de confirmer, de façon pragmatique, que chaque utilisateur est vraiment unique. En effet, l'allure des courbes d'erreurs individuelles est différente d'un utilisateur à l'autre, ce qui prouve que chaque utilisateur réagit de façon singulière face à la même interface. L'interface respecte les aspects cognitifs dans leur ensemble, même si certains d'entre eux peuvent être améliorés comme l'affordance, pour permettre une manipulation des documents résultant d'une recherche d'information encore plus aisée. Néanmoins, quelle que soit la visualisation proposée, celle-ci ne représente qu'une partie d'une solution générale de visualisation du résultat de recherche. En effet, nous avons pu vérifier que, compte tenu des nombreuses tâches ainsi que de l'expérience variable des utilisateurs, la visualisation doit proposer plusieurs façons de représenter le résultat de recherche. Ceci nous permet de confirmer ce que soulignaient (Vernier, 1997) et (Shneiderman, 1998) à savoir le fait qu'une interface développée pour une communauté d'utilisateurs ou pour une tâche précise pourra ne pas être appropriée à une autre communauté d'utilisateurs ou une autre tâche. Le fait de proposer différentes visualisations combinées doit permettre à chaque utilisateur de trouver l'outil le plus adéquat pour ses besoins et son niveau d'expérience. La combinaison de plusieurs visualisations doit également permettre à l'utilisateur d'apprécier différents types de relations entre les documents retrouvés (Hetzler, 1998) (relation de similarité, liaison par un lien hypertexte…).

L'interface de visualisation proposée montre cependant ses faiblesses du fait de l'interprétation exacte des informations représentées qui n'est pas forcément évidente pour un utilisateur néophyte. Cependant, son apprentissage semble relativement facile et à force d'essais, les utilisateurs semblent pouvoir la maîtriser. Cette visualisation se positionne donc

comme un outil plutôt destiné à des utilisateurs experts en recherche d'information ce qui implique que des visualisations plus « simples » à interpréter doivent être proposées.

4 Initiative pour l'évaluation d'interfaces

Nous visons au travers de cette section à initier une réflexion concernant la mise en place d'une démarche d'évaluation des interfaces de visualisation de résultats de recherche d'information. Dans un premier temps nous dressons un panorama des principales campagnes d'évaluation dans le domaine de la RI. Nous poursuivons ensuite par une discussion concernant le besoin d'une campagne d'évaluation pour les interfaces de visualisation de résultats de recherche. Quelques idées de réalisation sont également présentées. En aucun cas cette section vise à présenter une solution clé en main mais plutôt initier les discussions sur ce thème.

Démarches d'évaluation existantes

TREC ("Text REtrieval Conference") (TREC, 2004) est une campagne d'évaluation visant à proposer différentes tâches ("tracks") permettant d'évaluer différents aspects des SRI. Une tâche intéressante était la tâche interactive qui n'est plus reconduite depuis 2002. Elle a été assimilée à la tâche web qui ne sera pas reconduite en 2005. La tâche interactive s'est intéressée par exemple à l'évaluation, pour chaque système, de la mesure du rappel lors d'une recherche faite par un opérateur humain pour vérifier si la recherche est réellement effective.

CLEF ("Cross-Language Evaluation Forum") (CLEF, 2004) est une plateforme d'évaluation des SRI pour des recherches multilingues. Elle inclut une tâche interactive nommée iCLEF au travers de laquelle les systèmes proposés doivent répondre aux besoins des utilisateurs concernant des documents écrits dans une langue qu'ils ne maîtrisent pas. Elle vise également à mesurer l'utilisabilité de ces systèmes.

INEX ("Initiative for the Evaluation of XML Retrieval") (INEX, 2004) est une campagne d'évaluation à l'initiative européenne et relative aux systèmes de recherche d'information basés sur des documents structurés XML. Il existe une tâche interactive depuis 2004, qui a pour but, dans un premier temps, d'évaluer le comportement de l'utilisateur face à la manipulation de composants XML. Tous les utilisateurs utilisent le même système de recherche afin d'identifier les facteurs à prendre en compte dans le développement d'un SRI basé sur des éléments XML centré utilisateur.

Au regard de ces campagnes d'évaluation, nous pouvons souligner le fait que l'aspect restitution de l'information fait partie intégrante des tâches interactives notamment. Ainsi, l'évaluation de l'interface de restitution n'est pas dissociée du SRI auquel elles sont couplées, ne permettant pas de comparer les caractéristiques propres à ces interfaces.

Des classifications des interfaces de restitution de résultats de recherche ont toutefois été proposées telles que (Zamir, 1998). Ces différentes classifications sont basées principalement sur des critères tels que le type de représentation, les éléments visuels utilisés… La limite de ces travaux est qu'ils ne soulignent pas l'adéquation entre ces critères et les tâches ou les catégories des utilisateurs (débutants, experts…).

Une évaluation de ces interfaces de restitution pourrait être réalisée comme dans la démarche des IHM présentée dans (Shneiderman, 1998). Cependant cette évaluation ne serait toujours pas suffisante car elle ne permet pas de mesurer l'adéquation de l'interface au

contexte et aux tâches spécifiques à la RI. Dans ce cadre, nous initions une réflexion quant à la nécessité et à la forme que pourrait prendre une telle campagne d'évaluation des interfaces de restitution des résultats de recherche. Le but de cette campagne serait de comparer les interfaces entre elles mais surtout identifier les performances de ces interfaces en fonction de différents critères tels que la catégorie d'utilisateurs, la tâche à réaliser, la topologie des résultats, … .

Evaluation des interfaces de restitution des résultats de RI

L'évaluation d'interfaces de visualisation de résultats de recherche peut être considérée selon deux points de vue distincts :

- évaluation de l'interface utilisant son propre système de recherche. Le système de recherche est appliqué sur un jeu de requêtes et le résultat obtenu est visualisé au travers de l'interface,
- évaluation de l'interface indépendamment du système de recherche utilisé. Dans ce cas, l'interface doit fonctionner avec les résultats fournis par un système de recherche générique ou par des utilisateurs experts.

Evaluer l'interface indépendamment du système de recherche offre l'avantage d'assurer des résultats d'évaluation uniquement fondés sur l'utilisation des fonctionnalités de l'interface. La limite majeure réside dans le fait que l'interface doit être développée de manière totalement indépendante du système de recherche c'est-à-dire utilisant le résultat fourni par un système de recherche. Dans le cas contraire, des modifications doivent être apportées pour pouvoir exploiter le résultat d'un système de recherche externe.

Evaluer l'interface avec son système de recherche intrinsèque permet de pouvoir évaluer tout type de système dès l'instant où il est capable de traiter les requêtes fournies. Ce principe réduit fortement les adaptations à apporter pour pouvoir suivre le processus d'évaluation. L'inconvénient majeur de cette approche réside dans le fait qu'il est difficile de garantir une évaluation de l'interface indépendamment du système de recherche sous-jacent. L'utilisateur peut être influencé par l'efficacité du système de recherche utilisé et donc biaiser les résultats d'évaluation des fonctionnalités propres à l'interface. Des critères dévaluation spécifiques et permettant de faire abstraction du système de recherche doivent alors être définis.

Notre opinion s'oriente dans un premier temps vers le dernier type d'évaluation c'est-à-dire l'utilisation des interfaces avec les résultats d'un même système de recherche générique. Ce choix repose sur le fait que le développement d'une interface séparé d'un système de recherche est une « bonne » méthode de développement et par conséquent largement appliquée. Ce type d'évaluation pourrait convenir a priori à une grande majorité de systèmes.

Critères d'évaluation

Une interface de visualisation de résultats de recherche peut être évaluée suivant différents aspects comme :

- l'adéquation à une certaine catégorie d'utilisateur (utilisateur débutant dans l'utilisation de l'outil informatique et dans le domaine de recherche, utilisateur confirmé dans l'utilisation de l'outil informatique mais débutant dans un domaine de recherche, utilisateur expert dans un domaine de recherche mais débutant dans

l'utilisation de l'outil informatique, …) ou à un ensemble de catégories d'utilisateurs (voir toutes),
- l'adéquation à une tâche donnée (recherche sans objectif clairement défini au départ, recherche pour vérification de connaissance, recherche par association d'idées, recherche exhaustive, …),
- l'adéquation à une certaine taille de résultat (taille limite de visualisation des résultats en nombre d'éléments, taille limite de visualisation d'un élément, …),
- l'adéquation à une typologie de résultats (hétérogénéité des éléments, hétérogénéité des tailles des éléments, …),
- la complexité d'utilisation en termes de temps de traitements du résultat pour atteindre un objectif fixé (ex. visualisation d'une partie donnée d'un élément) ou en termes de nombre d'actions à réaliser (ex. clics souris) pour atteindre l'objectif.

Certains critères sont quantifiables comme la complexité d'utilisation, ou détectables comme l'adéquation à une certaine catégorie d'utilisateur (temps nécessaire à la réalisation d'un objectif en fonction de la catégorie d'utilisateur observée). D'autres critères font davantage appel à l'appréciation de l'évaluateur comme par exemple l'adéquation à une certaine forme de recherche.

Mise en oeuvre

La mise en œuvre d'un processus d'évaluation tel que présenté précédemment nécessite :
- la mise en place d'un groupe d'évaluateurs regroupant les différentes catégories d'utilisateurs identifiées (utilisateur débutant dans l'utilisation de l'outil informatique et dans le domaine de recherche, …). Un tel groupe peut être formé à partir de participants à une campagne d'évaluation comme c'est le cas pour INEX,
- la définition d'un ensemble de requêtes permettant d'évaluer chaque critère (cf. 4.2.1.). Une requête décrit un objectif à atteindre non plus en termes de contenu de résultats à l'image de campagnes d'évaluation comme TREC mais plutôt en termes d'accessibilité aux éléments du résultat,
- la définition de règles d'évaluation pour l'ensemble des requêtes. Les critères évaluables pour chaque requête doivent être spécifiés ainsi que la manière de les évaluer (objectif atteint ou pas, délai pour atteindre l'objectif, échelle de valeur pour estimer le degré de réalisation de l'objectif, …),
- un processus d'analyse des résultats offrant la possibilité de comparer les évaluations.

Application des critères à Easy-DoRView

A titre d'exemple, nous présentons l'application des critères introduits précédemment à l'interface d'Easy-DoRView.

La mesure de la corrélation entre les résultats de l'évaluation et la catégorie des utilisateurs par exemple peut être calculée en fin d'évaluation. Ce critère permet, à la lumière des résultats obtenus, de valider si une interface est plus adaptée à une catégorie d'utilisateurs voire à une tâche. Par exemple, nous pensons que l'interface proposée est utile dans le cas

d'une recherche sans objectif clairement défini au départ car les documents retrouvés sont présentés de façon globale en relation avec les mots-clés de la requête. Elle est également intuitivement utile pour une recherche exhaustive car les outils de sélection et de sélection fine permettent une recherche complète par rapport à des critères spécifiques pour l'utilisateur. Par contre, pour rechercher un document spécifique, l'interface est moins adaptée les documents n'étant pas directement accessibles puisqu'un point peut correspondre à plusieurs documents.

En ce qui concerne l'adéquation de l'interface à une certaine taille de résultat, Easy-DoRView permet de représenter visuellement un grand nombre de points (255*255*255 = 16 millions de points affichables dans l'absolu). Cependant, pour les distinguer correctement nous pensons qu'il ne faudrait pas dépasser les 60000 points. En ce qui concerne l'adéquation à la typologie des résultats, l'interface ne semble pas très atteinte par ce critère. Lorsque les documents sont très hétérogènes le nombre de points est égal au nombre de documents ce qui peut, d'après la capacité absolue, être affiché. Dans le cas de documents homogènes, l'interface semble toutefois plus adaptée car le fait qu'un point puisse regrouper plusieurs documents réduit le nombre de points d'où l'obtention d'un espace plus « simple ».

Pour évaluer la complexité d'utilisation en termes de temps de traitements du résultat pour atteindre l'objectif fixé, nous pouvons compter le nombre d'actions appelées par l'interface avec un principe de coût. Par exemple, une annulation peut coûter plus que le simple fait d'effectuer une rotation de l'espace. Ainsi, nous pourrions mesurer le nombre d'actions (nombre de rotations, nombre d'accès à des documents…), le nombre d'annulation d'actions, le temps mis pour obtenir l'objectif depuis la présentation des résultats à l'utilisateur jusqu'à la validation de l'objectif)…

5 Conclusion

A partir de l'évaluation qui a été réalisée sur une interface de restitution des résultats de recherche, nous pouvons nous demander objectivement si de telles évaluations sont utiles dans le contexte de la RI. Certes, cette évaluation, plutôt orientée IHM permet d'identifier si l'interface est compréhensible et utilisable. Elle ne permet pas de souligner l'adéquation entre l'interface et différents critères tels que la tâche de l'utilisateur, le niveau de connaissance de l'utilisateur dans le domaine de recherche ainsi que dans la manipulation d'outils informatique comme les SRI. Ainsi, de par ce constat, les concepteurs d'outils de recherche ne sont pas aidés dans le choix de l'interface à utiliser dans leur contexte spécifique.

Comme solution, nous initions la réflexion sur l'intérêt d'une campagne d'évaluation des interfaces de restitution des outils de recherche. En effet, nous pensons qu'une telle campagne permettrait, indépendamment des SRI auxquels elles sont associées, de comparer les interfaces selon les considérations spécifiques à la recherche d'information. Nous proposons également quelques orientations possibles ainsi que des idées de mise en œuvre d'une telle évaluation. Les solutions proposées ne visent pas l'exhaustivité mais sont quelques réponses aux besoins qui se sont fait sentir.

En conclusion, une réflexion plus profonde est à mener avec les acteurs intéressés par les interfaces de restitutions des résultats de recherche ou tout simplement les créateurs d'interfaces afin de proposer une campagne d'évaluation concrète. Cette réflexion pourrait donner lieu à des initiatives plus globales comme par exemple au niveau européen.

Références

(Chevalier, 2000) Chevalier M., Verlhac M., « ISIDOR: a visualisation interface for advanced information retrieval », 2nd International Conference on Enterprise Information Systems (ICEIS), B. Sharp J. Cordeiro J. Filipe (eds.), ISBN 972-98050-1-6, Stafford (England), pp 414 – 418, July 4-7, 2000.

(Chevalier, 2001) Chevalier M., Julien C., « ISIDORView : une interface de visualisation des résultats de recherche d'informations », Revue Extraction des Connaissances et Apprentissage (ECA), Hermès (ed.) , ISBN 2-7462-0216-6, 1(1-2), pp 135-140, Janvier 2001.

(Chevalier, 2002) Chevalier M., « Interface adaptative d'aide à la Recherche d'Information », Thèse option Science de doctorat de l'Université Paul Sabatier, soutenue le 16/12/2002.

(CLEF, 2004) Cross Language Evaluation Forum, http://clef.iei.pi.cnr.it/

(GVU, 1998) « 10th WWW User Survey », Graphic, visualisation & usability center (GVU), 1998. http://www.gvu.gatech.edu/user_surveys/survey-1998-10/

(Hetzler, 1998) Hetzler B., Harris W. M., Havre S., Whitney P., « Visualizing the full spectrum of document relationships », 5th International Conference of the International Society for Knowledge Organization (ISKO), pp 168-175, Lille, August 25-29, 1998.

(INEX, 2004) INitiative for the Evaluation of XML retrieval, http://inex.is.informatik.uni-duisburg.de:2004/

(Shneiderman, 1998) Shneiderman B., « Designing the user interface », Addison-Wesley Editeur, 3ème édition, ISBN 0-201-69497-2, 1998.

(Smith, 1978) Smith A.R., « Color gamut transform pairs », Computer Graphics, (12), pp 12-19, 1978.

(TREC, 2004) Text REtrieval Conference, http://trec.nist.gov/

(Vernier, 1997) Vernier F., Nigay L., « Représentation multiples d'une grande quantité d'information », 9ème journées Interaction Homme-Machine (IHM), pp 183-190, Futuroscope Poitiers, France, 10-12 Septembre, 1997.

(Ware, 1985) Ware C., Beatty J.C., « Using colour as a tool in discrete data analysis », CS-85-21, Computer Graphics Laboratory, University of Waterloo, August, 1985.

(Zamir, 1998) Zamir, O., « Visualisation of search results in document retrieval systems », General Examination, University of Washington, 1998.