Interface pour l'évaluation de systèmes de recherche sur des documents XML

Benjamin Piwowarski* — **Mounia Lalmas****

* *LIP6, Université Paris 6,*
Paris, France
bpiwowar@poleia.lip6.fr

** *Department of Computer Science,*
Queen Mary University of London, England
mounia@dcs.qmul.ac.uk

RÉSUMÉ. *L'évaluation des systèmes de Recherche d'Information est depuis le début un des piliers de l'évolution de ce domaine. La qualité de l'évaluation est d'une importance capitale puisqu'elle permet de discriminer les différents modèles entre eux. Il est donc primordial de pouvoir constituer des corpus où les questions et leurs jugements de pertinence associés sont de qualité. Alors qu'avec des documents plats les méthodes sont bien établies, ce n'est plus le cas avec des documents structurés de type XML. Il est donc nécessaire de développer de nouvelles façons d'évaluer. Nous présentons dans cet article l'interface utilisée lors de la campagne INEX 2003 qui permet d'évaluer de façon la plus consistante et la plus exhaustive possible les documents XML.*

ABSTRACT. *The evaluation of information retrieval systems is an important topic in information retrieval research. Evaluation must be accurately performed to rightly discriminate between retrieval approaches. To compare retrieval approaches, we require test collections, which consist of documents, queries and relevance assessments, the latter stating which documents are relevant to which queries. Obtaining consistent and exhaustive relevance assessment is crucial for the appropriate comparison of retrieval approaches. Whereas the evaluation methodology for retrieval approaches for flat text is well established, the evaluation of structured document retrieval approaches, such as those based on XML, is a research issue. In fact, new evaluation methodologies need to be developed for this purpose. In this paper, we concentrate on one aspect of the evaluation, that of the development of an interface to allow for the consistent and exhaustive assessment of XML documents. The interface was used at INEX, the evaluation initiative for XML retrieval.*

MOTS-CLÉS : *XML, évaluation, INEX, accès à l'information, recherche d'information*

KEYWORDS: *XML, evaluation, relevance assessment process, INEX*

1. Recherche d'Information et XML

Ce n'est que récemment que l'utilisation de l'information apportée par la structure a commencé à être utilisée dans le domaine de l'Accès à l'Information. Cet intérêt émane de deux communautés : la communauté "recherche d'information" (RI) et la communauté "base de données" (BDD). La communauté RI voit dans la structure un moyen d'améliorer la représentation des documents et d'introduire de nouvelles problématiques telles que les questions qui portent à la fois sur la structure et le contenu. Cette évolution vient également bouleverser la notion d'unité d'information qu'était jusqu'alors le document.

Un des principaux objectifs sur lequel se focalisent les travaux des communautés RI et BDD est le document XML[1] (eXtended Markup Language) qui est de plus en plus utilisé comme format de données pour les librairies digitales (DocBook, TEI), pour le web (XHTML) et de plus en plus supportés par les éditeurs de texte. Des systèmes de Recherche d'Information Structurée (RIS) adaptés à ce nouveau type de document voient actuellement le jour et il est nécessaire de pouvoir évaluer leurs performances [CAR 00, BAE 02, LUK 02, BLA 03, GöV 02].

L'évaluation des systèmes de RIS nécessite la création d'une collection de documents, d'un ensemble de questions et de leurs réponses associées (les *jugements de pertinence*). Il n'est plus possible d'utiliser les interfaces classiques d'évaluation puisque l'unité minimale d'information peut être n'importe quel élément contenu dans le document.

La première initiative internationale (INEX[2]) pour constituer un corpus de documents XML permettant d'évaluer les systèmes de RIS a débuté en avril 2002 avec une première rencontre en décembre 2002 [FUH 03] et une reconduction du projet en 2003. Ces rencontres initient la construction du domaine. De nombreux points aussi fondamentaux que la définition d'un besoin d'information ou d'une requête, l'évaluation de ces systèmes, sans même parler des principes de base sous-jacents aux moteurs de recherche sur ces corpus sont l'objet de discussions encore ouvertes.

L'utilisation dans INEX 2003 d'une échelle spécifique pour évaluer la qualité des réponses renvoyées par les systèmes de RIS nécessite donc une interface spécifique que nous décrivons dans cet article. Nous présentons tout d'abord l'échelle utilisée pour indiquer la pertinence d'un élément d'un document XML (section 3). Puis nous présentons l'interface qui a été utilisée lors de la phase d'évaluation d'INEX 2003 (section 3). Cette interface a des effets sur le comportement des juges que nous analysons dans la section 4. Enfin, nous nous intéressons dans la section 5 aux effets de l'interface sur les jugements portés.

---

1. `http://www.w3.org/XML/`
2. Initiative for the Evaluation of XML Retrieval, `http://inex.is.informatik.uni-duisburg.de:2003/`

2. Pertinence et RIS

Nous nous placerons dans le cadre d'INEX 2003 où les documents sont au format XML : à tout document correspond une structure arborescente unique. À chaque nœud (balise dans le document XML) de cet arbre est associé un label (étiquette). Tout nœud peut également contenir du texte. Nous appellerons *doxel* (pour *Document Element*) l'unité d'information qui peut être renvoyée par un système de Recherche d'Information Structurée, c'est-à-dire tout nœud de l'arbre. La collection INEX est composée d'un ensemble de 16000 articles – soit environ 8 million de doxels – publiés par l'IEEE. Le doxel qui contient tous les autres est le volume. Il correspond à une année de publications d'une revue. Chaque volume est composé d'un ensemble de publications qui à leur tour sont composées d'un certain nombre d'articles. Les articles sont eux-mêmes fortement structurés (en-tête, section, paragraphe, ...). Lors d'INEX 2003, il a été décidé que seuls les doxels qui appartiennent à un article pouvaient être jugés.

En RI classique l'unité d'information est le document et les jugements ont une valeur qui peut varier entre 0 (le document est pertinent) et 1 (le document est non pertinent)[3]. L'utilisation d'une telle échelle pour la RIS est problématique : lorsqu'une section d'un document contient un paragraphe pertinent, quelle valeur de pertinence donner à la section ? Il n'est pas souhaitable de lui donner la valeur 1 puisque la section contient beaucoup d'information non pertinente ; la valeur 0 n'est pas non plus envisageable puisque la section contient de l'information pertinente. Si une valeur intermédiaire est choisie, il sera alors impossible de distinguer une section qui contient un paragraphe pertinent avec une section moyennement pertinente (mais qui ne contient pas de paragraphe pertinent).

Un travail précurseur sur les documents multimédia [CHI 97] montre que la pertinence d'un document peut être vue sous l'angle de deux implications logiques. La première, $d \rightarrow q$ (le document implique la question), est l'exhaustivité du document d pour la question q : est-ce que le document traite de tous les différents sujets soulevés par la question q ? La seconde, $q \rightarrow d$ (la question implique le document), est la spécificité du document d pour la question q : la question implique le document lorsque les seuls sujets abordés par le document sont ceux de la question.

Cette façon de décrire la pertinence d'un document peut s'appliquer telle quelle à la pertinence d'un doxel et permet d'éviter les écueils d'une mesure unidimensionnelle. En effet, en reprenant l'exemple précédent, une section qui contient un paragraphe pertinent sera moyennement spécifique puisqu'elle traite de sujets autres que ceux de la question. Et la section moyennement pertinente sera par contre très spécifique (mais moyennement exhaustive puisqu'elle ne traite pas complètement des sujets de la question).

Pour permettre l'évaluation des systèmes de RIS, une échelle à deux dimensions a ainsi été proposée lors de la campagne d'évaluation INEX 2002 [GöV 02] et a été

3. La plupart du temps, l'échelle utilisée est en fait binaire.

depuis remaniée pour INEX 2003[4]. Nous présentons ici cette dernière, composée également de deux dimensions : la première mesure l'*exhaustivité* et la seconde la *spécificité* du doxel pour une question donnée. Ces deux dimensions sont multivaluées, ce qui permet une plus grande finesse dans les jugements [KEK 02]. Dans ce qui va suivre, nous ne nous intéressons qu'à une question donnée.

Exhaustivité

L'exhaustivité ne tient compte que de *la présence ou de l'absence de l'information recherchée* dans un doxel, même si cette information n'apparaît que dans une toute petite partie du doxel. Par exemple, le doxel représentant le document entier sera considéré comme fortement exhaustif même si un seul paragraphe dans tout le document est très pertinent pour la question et que le reste du document ne l'est pas. Nous distinguerons quatre niveaux de d'exhaustivité :

Non exhaustif (0) Le doxel ne traite pas du sujet de la question ;

Faiblement exhaustif (1) Le doxel traite marginalement le sujet de la question ;

Moyennement exhaustif (2) Le sujet de la question est en grande partie traité dans le doxel ;

Totalement exhaustif (3) Le sujet de la question est traité exhaustivement dans le doxel.

Spécificité

La spécificité est totalement liée à l'évaluation de documents structurés. Cette mesure s'intéresse au *degré avec lequel le doxel traite de toute l'information recherchée* si le doxel contient l'information recherchée ou *d'une partie de cette information* si le doxel en contient une partie. Nous distinguerons à nouveau quatre niveaux :

Pas spécifique (N) Le doxel renvoyé ne contient pas de passage pertinents ;

Faiblement spécifique (F) Une petite partie seulement de l'information contenue dans le doxel est de l'information pertinente ;

Moyennement spécifique (M) La plus grande partie de l'information contenue dans le doxel est de l'information pertinente.

Totalement spécifique (T) Le doxel renvoyé ne contient (presque) que de l'information pertinente.

Ces jugements ne sont pas tout à fait orthogonaux, car lorsqu'un élément n'est pas exhaustif, il n'est pas possible de définir sa spécificité et inversement. Il n'y a donc que 10 valeurs possibles et non 16. Un jugement de pertinence sera représenté par deux lettres, la première étant l'exhaustivité, la seconde la spécificité. Par exemple, $2T$ correspond à "moyennement exhaustif" et "totalement spécifique" ; $0N$ correspond à "non pertinent". Nous dirons dans la suite qu'un doxel est pertinent si son jugement de

4. En particulier, la terminologie a été complètement changée afin d'éviter des confusions

pertinence prend n'importe quelle valeur à l'exception de $0N$. Inversement, un doxel sera non pertinent s'il est jugé $0N$. Nous utiliserons le symbole ? pour noter une valeur inconnue pour l'une de ces deux dimensions.

3. L'interface

Pour permettre de juger de façon efficace des documents XML, il était nécessaire de développer une interface simple d'utilisation qui soit capable d'effectuer certains contrôles sur l'évaluation en cours et de guider l'utilisateur dans ses jugements. L'interface utilisée lors d'INEX'02 souffrait en effet de certains défauts :

Accès au document Deux vues étaient proposées pour tout document XML : la première permettait une lecture facile du document ; la seconde affichait le document XML directement et permettait de donner les jugements de pertinence. Le découplage de ces deux vues rendait complexe la tâche de localisation des informations pertinentes ;

Consistance des jugements Les contrôles sur l'évaluation en cours n'étaient effectuées qu'*a posteriori* ;

Exhaustivité des jugements La liste des doxels à juger était figée : lorsqu'un utilisateur jugeait un doxel comme étant faiblement (F) ou moyennement (M) spécifique, il aurait été intéressant de lui demander de localiser le(s) doxel(s) totalement (T) spécifique(s) contenus dans celui-ci.

Ces trois points sont cruciaux si l'on considère que les juges sont les participants eux-mêmes – contrairement à d'autres campagnes d'évaluations comme TREC [VOO 02] – mais que nous voulons tout de même obtenir une évaluation aussi parfaite que possible. Pour palier à ces défauts, nous avons développé une interface où le document n'a qu'une seule vue permettant la lecture ainsi que le jugement (en faisant apparaître les balises XML). Cette interface contrôle les jugements et modifie la liste des doxels à juger en fonction des actions de l'utilisateur. Cette interface est accessible sur un serveur web grâce à un navigateur.

3.1. *Description de l'interface*

Cette interface utilisateur s'articule autour des composants suivants :

– La vue globale qui récapitule l'état courant de l'évaluation et qui donne accès aux différents volumes de la collection INEX ;

– Une vue d'un volume qui donne accès aux articles qui le composent ;

– Une vue d'un article (figure 1) qui permet à l'utilisateur de juger les doxels du document ;

– La formulation originale de la question pour permettre au juge de s'y référer aussi souvent que possible ;

– L'édition du mode de coloration des mots-clé : l'utilisateur peut à tout moment associer des séquences (mot, phrase) à une couleur donnée. Lors de l'affichage du document, ces séquences sont présentées dans un cadre de la couleur désirée.

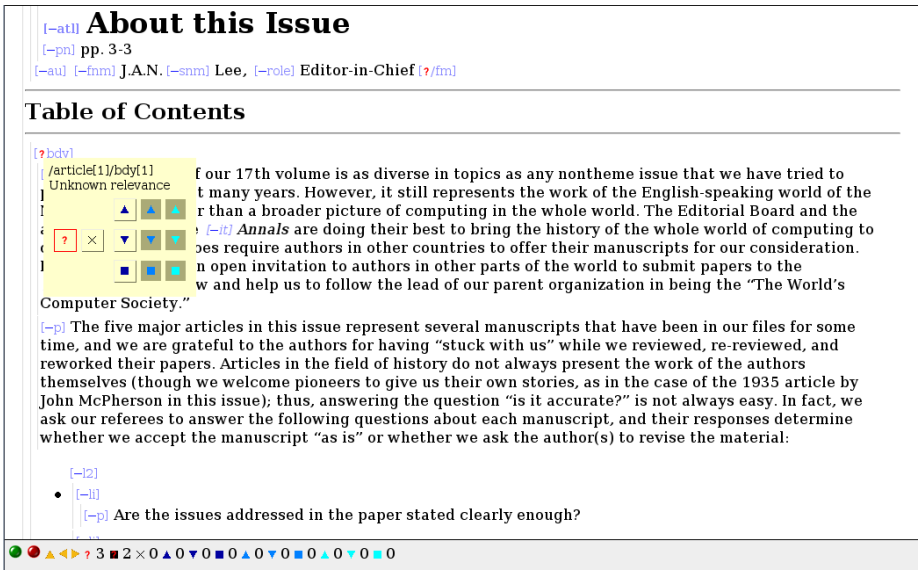

Figure 1. *Fenêtre d'évaluation principale d'INEX 2003 : l'utilisateur est en train de juger le corps du document (`/article[1]/bdy[1]`). Le jugement actuel est inconnu (?), et seules les valeurs 0N (non pertinent) et 1 ? (exhaustivité faible) sont possibles.*

La vue d'un article (figure 1) est composée de deux parties : la première donne une vue du document où apparaissent (en gris clair) les balises XML avec leurs jugements de pertinence associés. Lorsqu'un utilisateur survole une balise, le contenu du doxel est encadré afin d'en marquer les limites. Lorsque l'utilisateur clique sur la balise, un panneau d'évaluation apparaît. Ce panneau donne la localisation précise du doxel sous la forme d'un XPath[5] ; le jugement actuel en toutes lettres ; une série de symboles correspondant aux jugements possibles. Ces jugements comprennent les 10 valeurs de l'échelle d'INEX 2003 et également une valeur "inconnu" qui permet d'effacer un jugement[6]. Certaines de ces valeurs peuvent être jugées impossibles par des règles d'inférence (voir section 3.2). Dans ce cas, elles apparaissent grisées dans le panneau et ne sont pas actives. Lorsque l'utilisateur clique sur une valeur "active", le panneau se ferme, le jugement de pertinence est transmis au serveur. C'est à ce moment que les règles d'inférence entrent en jeu, d'une part pour vérifier la validité de jugement, d'autre part pour mettre à jour la vue du document.

5. Un XPath permet entre autres de repérer de manière unique un doxel dans un document XML, `http://www.w3.org/TR/xpath`
6. Cela peut être utile lorsque le juge veut s'accorder plus de réflexion ou annuler certaines contraintes

En bas, apparaît une ligne où apparaissent des informations sur le document en cours d'évaluation ; de gauche à droite : le premier bouton rond permet de juger plusieurs doxels en même temps, le second efface la sélection courante. Les trois flèches permettent respectivement de passer à la vue qui contient ce document (un volume, *i.e.* une année de publication de la revue), de centrer la fenêtre sur l'élément précédent (ou suivant) à juger. Le reste de la ligne contient des informations sur la vue courante : nombre de doxels à juger, nombre de doxels inconsistants [7] puis nombre de doxels pour chacune des valeurs possibles dans l'échelle d'évaluation d'INEX 2003. Toutes ces informations sont importantes car elles permettent au juge de s'assurer de l'avancement de sa tâche et d'accélérer le processus de jugement en permettant des sauts directs vers les doxels à juger.

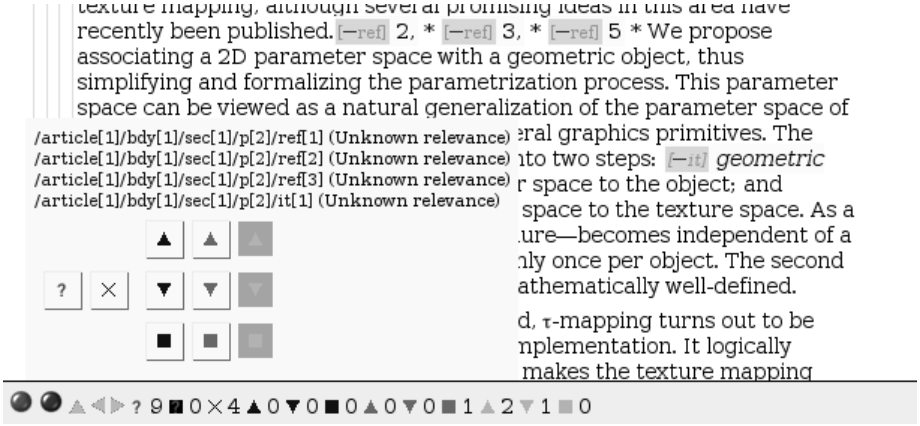

Figure 2. *Évaluation groupée. Dans cette copie d'écran, 4 doxels (dont les balises ont un fond grisé) peuvent être jugées en même temps. Des règles d'inférence partielles sont calculées pour restreindre les valeurs de pertinence communes que peuvent prendre ces quatre doxels.*

Pour porter un jugement de pertinence, il est possible d'utiliser l'un des modes suivants :

un à un Un seul doxel est jugé, comme décrit plus haut ;

frères Un ensemble de frères est jugé. Ceci a pour but d'accélérer l'évaluation lorsque seul un petit nombre de doxels sont pertinents parmi un ensemble de frères (par exemple, un paragraphe parmi n). Dans ce cas, l'interface permet de juger ce doxel puis d'assigner la valeur "non pertinent" à l'ensemble de ses frères ;

groupé (figure 2) Enfin, il est possible de juger un ensemble présélectionné de doxels. La sélection d'un doxel s'effectue en cliquant sur la balise correspondante.

Les deux derniers modes de jugement n'étaient pas initialement contenus dans l'interface ; c'est suite à la demande de nombreux juges que de telles fonctionnalités ont été

7. la notion de consistance est définie dans la section 3.2

développées. Elles ont été immédiatement adoptées. Les choix effectués lors du développement de cette interface ont en général été influencées par les critiques formulées lors d'INEX'02 et les commentaires des premiers utilisateurs du système actuel.

3.2. *Consistance des jugements*

En RI classique, chaque document étant indépendant, les jugements de pertinence sont également indépendants. Ce n'est plus le cas avec des documents structurés. Par exemple, si une section composée exclusivement de paragraphe est pertinente, elle contient forcément un paragraphe qui est pertinent. Il est ainsi possible et souhaitable de définir un certain nombre de règles d'inférence sur les jugements portés à l'intérieur d'un même document.

Les règles utilisées étaient appliqués après chaque jugement afin de refléter les possibilités valides et actuelles de jugements pour les autres doxels du document. Lors de l'évaluation des questions d'INEX'03, nous avons utilisé les trois règles suivantes :

1) Si tous les enfants d'un doxel sont non pertinents, alors ce doxel est non pertinent. Cette règle est évidente lorsqu'on considère la définition des deux dimensions.

2) L'exhaustivité d'un doxel est toujours supérieure ou égale à l'exhaustivité d'un de ses fils. Il est en effet impossible de trouver plus d'informations pertinentes au sein d'un élément que dans la totalité de l'élément.

3) La spécificité d'un doxel est inférieure ou égale au maximum de la spécificité de ses fils : un doxel qui est par exemple totalement spécifique ne peut pas contenir que des doxels qui ne sont que grandement, moyennement ou pas du tout spécifiques.

Les règles présentées ici sont un sous-ensemble de celles qui ont été envisagées afin de s'assurer de la consistance des jugements : l'outil étant nouveau, il n'était pas possible de changer continuellement les règles utilisées. Toutefois, durant la période d'évaluation, nous avons été amenés à ajouter une règle (la troisième) ainsi qu'à changer l'interprétation des règles : au départ, les seuls enfants considérés correspondaient aux balises XML présentes dans le document. Or, dans un paragraphe qui contient un passage en italique (balise `<it>`), une application directe de la règle 1 amène a inférer la non pertinence du paragraphe à partir du moment où le passage en italique est non pertinent. Pour éviter ce type de travers, nous avons par la suite considéré qu'un enfant peut aussi correspondre à un nœud de type texte. Le paragraphe de l'exemple précédent contient dans ce cas au moins trois fils (texte, `<it>`, texte) ce qui rend caduque la règle (1). L'ajout de la troisième règle nous a amené à ajouter un état "(non-)consistant" pour tout doxel jugé : après un changement (ou ajout) de règle, la consistance de tous les jugements était recalculée et l'ensemble des doxels non consistants étaient alors visibles au niveau de l'interface.

Les règles de consistance qui pourront être ajoutées lors INEX'04 sont les suivantes. L'exhaustivité d'un doxel est inférieure ou égale à la somme des exhaustivités des enfants. Ainsi, un doxel fortement exhaustif (3) doit avoir un enfant fortement

exhaustif, ou bien un enfant moyennement exhaustif (2) et un enfant faiblement exhaustif (1), etc. Avec la seconde règle, la spécificité d'un doxel est supérieure ou égale au minimum des spécificités de ses enfants. Ces deux nouvelles règles permettront de s'assurer d'une consistance forte des jugements, ce qui a une importance capitale pour la qualité du corpus. Ces règles seront activement discutées lors d'INEX'03 car elles sont déterminantes pour la qualité de jugements de pertinence.

3.3. *Exhaustivité des jugements*

Le nombre important de doxels (8 million) rend impossible une évaluation exhaustive de la totalité du corpus d'INEX. Une présélection (*pooling*) similaire à celle employée pour les campagnes d'évaluation TREC [SPA 75] a tout d'abord été appliquée. Cette technique qui consiste à sélectionner un sous-ensemble de doxels à juger parmi l'ensemble des doxels renvoyés par les systèmes de RIS n'est toutefois pas totalement satisfaisante pour l'évaluation de documents structurés.

En effet, un système de RIS parfait doit renvoyer uniquement des doxels qui ont une spécificité totale (T). Lors de l'évaluation, il peut donc être intéressant de contraindre le juge à trouver de manière exhaustive ces éléments dans les documents jugés. Toutefois, afin de ne pas demander un jugement complet de tous les doxels présents dans un document contenant des doxels à juger, ce qui serait encore une fois trop coûteux, nous avons choisi la démarche itérative suivante :

1) l'utilisateur juge un doxel ;

2) ce jugement peut éventuellement modifier la liste des doxels à juger.

La procédure s'arrête lorsqu'il ne reste plus qu'un doxel à juger et que le jugement de ce doxel n'ajoute rien à la liste. Lors d'INEX 2003, nous avons choisi les règles suivantes.

– *Lorsque l'utilisateur juge un doxel non pertinent*, aucun doxel n'est ajouté à la liste. Ce choix permet de ne pas trop alourdir la tâche de jugement, car il permet de traiter très vite les documents totalement non pertinent.

– *Lorsque l'utilisateur juge qu'un doxel est totalement spécifique*, seuls ses ancêtres sont ajoutés à la liste des doxels. Ce choix a été fait car il est intéressant de regarder s'il n'y a pas d'autres doxels frères qui pourraient être pertinents. Ne pas ajouter ses enfants permet d'éviter la recherche d'éléments de pertinence inférieure qui alourdirait trop la tâche.

– *lorsqu'un utilisateur juge qu'un doxel n'est pas faiblement ou grandement spécifique*, ses enfants et ses ancêtres sont ajoutés automatiquement à la liste des doxel à évaluer. Cette dernière règle permet de forcer la recherche d'un doxel totalement spécifique.

Pratiquement, il s'est avéré que l'utilisation cumulée de ces deux dernières ajoutait déjà beaucoup de nouveaux doxels à la liste : il sera donc nécessaire de les assouplir lors des éditions suivantes d'INEX.

4. Analyse des sessions

Dans cette section, nous nous intéressons à la façon dont l'interface a été utilisée, et en particulier à la durée d'une session de jugements, à l'utilisation du mode "groupé" ou du mode "frères" et à la façon dont les jugements ont été portés à l'intérieur du document (l'utilisateur suit-il un chemin particulier ou juge-t'il de façon aléatoire ?). Nous aborderons l'influence de l'interface sur les jugements proprement dits dans la section suivante.

Le nombre de jugements analysés est de 203384[8]. La période analysée va du 10 septembre 2003 au 25 novembre 2003. Les traces du serveur web d'INEX ont été utilisées comme seule source d'information : l'analyse de l'URL demandée nous permet de savoir précisément l'heure à laquelle un utilisateur demande à voir un document (et lequel) et l'heure à laquelle il juge un (groupe de) doxel(s) (et quel est son ou leur jugement de pertinence associé).

Nous nous sommes tout d'abord intéressés à dégager les sessions de jugement : une session est l'ensemble des actions (vue d'un document, jugement d'un doxel) effectuées sans interruption par l'utilisateur. Chaque couple d'actions doit être espacé d'au plus $T_{\max}$ secondes pour faire partie de la même session. Nous avons fixé $T_{\max}$ à 18 minutes, ce qui correspond environ à l'intervalle de temps moyen qui sépare la vue d'un document du premier jugement porté – augmenté de sa déviation standard. En utilisant cette valeur, nous avons observé une durée moyenne de 52 minutes, avec une moyenne de 111 doxels et de 20 documents jugés par session.

Pour ce qui est de la façon de juger les doxels, la grande majorité (80 %) a été effectuée en ne jugeant qu'un seul doxel. Dans 17 % des cas, les jugements ont été effectués en groupant les doxels et dans 2 % des cas en jugeant un ensemble de frères. L'introduction de méthodes pour juger un ensemble de doxels de manière simultanée a donc été utilisée. Aucune différence notable dans la façon de juger n'est perceptible : l'utilisation d'un *mode de jugement* ne privilégie pas spécialement un certain *type de jugement*.

Le temps passé est très dépendant de la présence de doxels pertinents dans le document. Ainsi, les juges passent en moyenne 8 minutes (pour 28,2 doxels jugés) pour un document qui contient au moins un doxel pertinent contre 1 minute (pour 1,3 doxels) par document qui ne contient que des doxels non pertinents. Ceci est en partie du à l'interface qui n'allonge jamais la liste des doxels à juger lorsqu'un doxel est jugé "non pertinent".

L'interface a également une grande influence dans les chemins choisis. Dans 38 % des cas, le juge passe d'un doxel à son frère, dans 10 % des cas il passe à son père et dans 12 % des cas à un de ses fils. En moyenne, 90 % des jugements portés le sont à proximité du précédent (distance en nombre de relation parent-enfant ou frère à emprunter inférieure ou égale à 3). Le comportement moyen d'un utilisateur est donc

8. Les jugements "groupé" ou "frères" comptent pour un

de juger de proche en proche, et ce comportement est renforcé par l'interface qui ajoute systématiquement des doxels proches de celui que l'utilisateur vient de juger.

En conclusion, les modes de jugement rapide, et en particulier le mode "groupé", ont été très utilisés lors de la campagne INEX'03. La vue et l'ajout de doxels font du jugement du document une tâche locale, l'utilisateur jugeant tout d'abord les doxels proches de ceux qu'il a déjà jugés.

5. Analyse des jugements

L'interface utilisateur a des effets directs sur le comportement des juges comme nous l'avons vu dans la section précédente. Nous allons maintenant nous intéresser aux effets sur les jugements de pertinence eux-mêmes. Nous nous référerons souvent à ce qui a été observé lors d'INEX'02, puisqu'il s'agit de la seule source d'information comparable. Le but de l'interface était triple. Tout d'abord, l'interface devait être simple et agréable à utiliser. Ensuite, l'interface devait contraindre les jugements portés en fonction de certaines règles d'inférence décrites plus haut. Enfin, l'interface modifie la liste des doxels à juger. Nous analysons dans la suite ces deux derniers points.

Au niveau de la consistance, nous avons l'assurance que les jugements suivent les trois règles énoncées plus haut. Une étude des jugements d'INEX'02 montre que sans vérification de la consistance 23 jugements par question étaient invalides pour un équivalent de la règle 2 (le parent est plus exhaustif que l'enfant). L'introduction de règles plus complexes lors d'INEX'03, et la prévision d'en ajouter de nouvelles lors d'INEX'04, rend indispensable l'évaluation et la prévention automatique de ces inconsistances.

Dans INEX'02, 66 % des jugements portés l'étaient sur des doxels présélectionnés. Avec cette interface, ce nombre est tombé à 26 % : le nombre de doxels ajoutés est donc très important, ce qui est un résultat positif. De plus, nous avons calculé que 68 % des jugements de type "totalement spécifique" (*i.e.* les doxels qui sont les plus pertinents) sont des doxels qui n'étaient pas présent à l'origine dans la liste. Les règles d'inférence permettent de juger *automatiquement* un nombre important de doxels (7 %) ce qui est une aide non négligeable lors de l'évaluation.

6. Conclusion

Globalement, l'interface a été appréciée par les différents participants d'INEX'03 ; en regard de celle utilisée l'année précédente, un grand progrès a été effectué. Les principales critiques qui ont été formulées sur l'interface d'évaluation d'INEX'03 ont porté sur l'ajout automatique de doxels à juger. En effet, dans bien des cas, l'utilisateur était obligé de regarder un ensemble très importants de doxels : par exemple, lorsque l'on juge un élement dans une bibliographie, il fallait juger son parent "grandement" ou "partiellement" spécifique ce qui amenait alors à juger l'ensemble de ses enfants - c'est à dire l'ensemble des autres éléments de bibliographie. Une manière d'éviter ce

problème serait de ne demander à l'utilisateur de juger l'ensemble des enfants à partir du moment où aucun descendant "totalement" spécifique n'a été décelé. Il pourrait être également souhaitable de prendre en compte la nature et la taille du doxel considéré – aussi bien pour les règles d'inférence que pour les règles d'ajout.

Le point positif de cette interface est qu'il permet un contrôle très précis des actions effectuées par le juge et des actions possibles à tout instant. Une études plus précise des règles d'inférence et d'ajout automatique sont nécessaires afin d'aboutir à un outil d'évaluation qui soit simple d'utilisation, qui permette un jugement le plus exhaustif possible et qui minimise le travail à effectuer par le juge.

Cette interface sera très probablement utilisée lors d'INEX 2004 ; les différentes remarques et critiques formulées ici seront bien évidemment prises en compte, ainsi que ce qui résultera des débats autour de l'interface lors d'INEX'03.

7. Bibliographie

[BAE 02] BAEZA-YATES R., FUHR N., MAAREK Y. S., Eds., *ACM SIGIR 2002 Workshop on XML*, août 2002.

[BLA 03] BLANKEN H. M., GRABS T., ANDRALF SCHENKEL H.-J. S., WEIKUM G., Eds., *Intelligent Search on XML Data, Applications, Languages, Models, Implementations, and Benchmarks*, vol. 2818 de *Lecture Notes in Computer Science*, Springer, 2003.

[CAR 00] CARMEL D., MAAREK Y., SOFFER A., Eds., *ACM SIGIR 2000 Workshop on XML*, juillet 2000.

[CHI 97] CHIARAMELLA Y., « Browsing and Querying : two complementary approaches for Multimedia Information Retrieval », *HIM'97 International Conference*, Dortmund, Germany, 1997.

[FUH 03] FUHR N., GOEVERT N., KAZAI G., LALMAS M., Eds., *Proceedings of the First INEX Workshop*, Sophia Antipolis, France, 2003, ERCIM.

[GöV 02] GÖVERT N., « Assessments and evaluation measures for XML document retrieval », *Proceedings of the First Annual Workshop of the Initiative for the Evaluation of XML retrieval (INEX)*, DELOS workshop, Dagstuhl, Germany, décembre 2002, ERCIM.

[KEK 02] KEKÄLÄINEN J., JÄRVELIN K., « Using graded relevance assessments in IR evaluation », *Journal of the American Society for Information Science (JASIS)*, vol. 53, n° 13, 2002, p. 1120–1129.

[LUK 02] LUK R., LEONG H., DILLON T., CHAN A., CROFT W. B., ALLAN J., « A Survey in Indexing and Searching XML Documents », *JASIS*, vol. 6, n° 53, 2002, p. 415–437.

[SPA 75] SPARCK JONES K., VAN RIJSBERGEN C. J., « Report on the need for and provision of an ideal information retrieval test collection », rapport n° 5266, 1975, Computer Laboratory, University of Cambridge, Cambridge, England.

[VOO 02] VOORHEES E. M., HARMAN D. K., Eds., *The Tenth Text Retrieval Conference (TREC 2001)*, Gaithersburg, MD, USA, 2002, NIST.