Recherche de documents dans my.epfl

Predrag.Viceic@epfl.ch, EPFL – Domaine IT - KIS, concepteur de my.epfl

Fulltext search in my.epfl

La recherche full-text dans my.epfl

Avec la migration effectuée en fin de l'année passée, la partie **documents** du projet my.epfl s'est vue offrir la fonctionnalité de recherche **full-text**.
En effet, à moins de connaître l'emplacement exact d'un document, il était impossible de le retrouver. Ceci peut être handicapant dans un système contenant bientôt un demi-million de documents, chose désormais corrigée.
Comment ça fonctionne? C'est simple: cliquez sur l'icône et rentrez la clé de recherche dans le champ texte en n'oubliant pas de cliquer sur **Chercher**. My.epfl vous retourne la liste de documents, en précisant la date de la dernière modification, l'auteur du document, ainsi que le lien vers le dossier où le document se trouve.

Recherche

My.epfl indexe le contenu texte de la grande majorité des types de fichiers, les PDF, les documents MS Office ainsi que les documents OASIS (OpenOffice, etc..). Vos termes de recherche sont pris en compte en considérant les espaces comme des **OU** logiques. Vous pouvez forcer la recherche des expressions en entourant vos termes de recherche par des guillemets.
En plus de ces recherches *intuitives*, vous pouvez également utiliser la syntaxe **Lucene** (**OR**, **AND**, **NOT**, ...) en veillant à écrire les opérateurs en caractères majuscules. Ainsi, la recherche **my AND NOT epfl** retournera les documents qui contiennent le mot **my**, mais pas le mot **epfl**.
Les documents qui vous sont retournés sont bien entendu uniquement ceux pour lesquels vous avez au moins les droits d'accès en lecture.

Indexation

L'indexation est le processus qui commence par transformer un document, par exemple PDF, en texte *simple* en supprimant toutes les spécificités dues à la manière dont les données du texte sont sauvegardées. L'indexeur utilise ensuite ce qu'on appelle un tokeniseur pour extraire les mots significatifs (les termes), c'est-à-dire ceux qui contiennent le plus d'information. Ceci exclut typiquement les déterminants ainsi que les verbes et les noms très courants. Ces termes, devenant ainsi les clés de recherche, sont ensuite stockés de manière optimale dans un fichier qu'on appelle l'index.

Sur la plupart des systèmes, la réindexation se fait périodiquement, en ne prenant en compte que les fichiers modifiés, ajoutés ou supprimés. Ainsi, search.epfl.ch réindexe toutes les nuits les pages Web du domaine epfl.ch. À la différence des autres systèmes, my.epfl réindexe les documents à la volée. Dès qu'un fichier est déposé, supprimé ou modifié, l'indexeur est notifié de l'opération et procède à la réindexation. Ainsi, l'index est en permanence à jour et de plus, un document peut être recherché dès qu'il est déposé sur my.epfl.

Droits d'accès

Comme dit précédemment, les documents retournés respectent les droits d'accès. Ainsi, les résultats de recherche diffèrent en fonction de l'utilisateur. Il va de soi que, lors de la modification de ces droits, le document est réindexé. Ainsi, lors des recherches subséquentes les nouveaux droits d'accès sont pris en compte immédiatement.
Pour les curieux, ce tour de passe-passe est réalisé grâce à une astuce: lors de la réindexation, les droits d'accès (utilisateurs et groupes ayant droit) sont appondus aux mots-clés extraits par le tokeniseur et sont ajoutés à l'index. Ces droits deviennent ainsi les mots clés supplémentaires ajoutés au document.
Lors de la recherche, le système append à vos clés de requête votre identifiant de l'utilisateur, ainsi que les identifiants des groupes dont vous faites partie en les combinant à l'aide des opérateurs logiques. Ces informations deviennent donc partie intégrante de la requête, ce qui permet de ne rechercher que les documents auxquels vous, ou les groupes dont vous faites partie, avez droit. Bien entendu, les droits sont vérifiés une dernière fois avant de fournir le document à l'utilisateur. Cette astuce permet ainsi non seulement d'accélérer les recherches en les limitant aux documents auxquels vous avez droit, mais permet aussi l'utilisation d'un moteur de recherche standard et bien supporté, **Lucene**, sans trop modifier celui-ci.

Liens

- lucene.apache.org/java/2_3_2/queryparsersyntax.html

GLOSSAIRE

tokeniseur: Il s'agit du processus permettant de démarquer les différentes sections d'une chaîne de caractères. En effet, un ordinateur n'est pas capable seul de déterminer quels sont les mots d'une phrase; il n'y voit qu'une chaîne de caractères. Un processus de tokenization consisterait donc à séparer ces mots, selon les espaces. *W*

W = tiré de Wikipédia

ISSN 1420-7192