Combo - Development #20376

revoir l'indexation intégrale

30 novembre 2017 20:06 - Frédéric Péters

Statut:	Fermé	**Début:**	30 novembre 2017
Priorité:	Normal	**Echéance:**	26 janvier 2018
Assigné à:	Frédéric Péters	**% réalisé:**	0%
Catégorie:		**Temps estimé:**	0:00 heure
Version cible:			
Patch proposed:	Oui	**Planning:**	

Description

Aujourd'hui on utilise toute la mécanique django-haystack pour l'indexation des pages; c'est bien sauf que souvent en premier ce sont les démarches qui sont vraiment importantes, pas les quelques pages sur le portail qui listent celles-ci. Là-dessus je pense qu'on doit continuer à utiliser haystack parce que ça fournit une cible unique vers whoosh pour une indexation locale facile et solr et elasticsearch pour des trucs plus avancés, qui peuvent déjà se trouver utilisés par le site de la collectivité; mais zapper la couche modèle pour la gérer nous-mêmes, avec la possibilité pour une cellule de fournir plusieurs "hits" (une cellule "démarches d'une catégorie" fournira un hit par démarche, genre).

Révisions associées

Révision 273f1b00 - 03 janvier 2018 11:43 - Frédéric Péters

general: add external links to search results (#20376)

Historique

#3 - 26 décembre 2017 16:36 - Frédéric Péters

- Fichier 0001-general-add-external-links-to-search-results-20376.patch ajouté

- Statut changé de Nouveau à En cours

- Patch proposed changé de Non à Oui

Ce n'est pas trop possible de zapper la couche modèle et de quand même profiter d'haystack, j'ai pris l'option d'un modèle supplémentaire, ExternalLinkSearchItem, qui est géré de manière totalement automatique : avant l'indexation du site il y a parcours de toutes les pages et les cellules de celles-ci peuvent déclarer des liens externes (méthode get_external_links_data).

#4 - 02 janvier 2018 11:03 - Thomas Noël

Dans combo/data/search_indexes.py tu déclares une classe PageIndex qui existe déjà juste au dessus.

Je ne vois pas trop l'usage de l'option « --skip-external-links-collection » dans la commande update_index, tu as une idée en tête ?

Je n'arrive pas trop à voir comment va être gérée la visibilité des résultat de recherche : si une démarche n'est accessible qu'à certains utilisateurs, comment s'assurer qu'elle ne pas pas être visible à d'autres dans les résultats de recherche ? (peut-être que c'est déjà un "soucis" sur les pages, en fait, j'ai jamais bien fait attention)

Concernant ce passage dans combo/apps/wcs/models.py :

```
            text = ' '.join([self.cached_json.get('description', ''),
                             ' '.join(self.cached_json.get('keywords', []))]).strip()
```

Je me disais que les keywords pourraient en fait avoir un boost=2, ce qui voudrait dire d'ajouter un index "keywords" dans ExternalLinkSearchIndex ? L'idée derrière étant que si on repère qu'un mot est particulièrement recherché sur un site de démarches, on le met dans le/les formulaires qui y sont liés et ça booste le résultat.

Mais ça peut être une évolution plus tard, quand on aura aussi un mécanisme de mot clé sur les pages internes aussi.

#5 - 02 janvier 2018 12:14 - Frédéric Péters

- Fichier 0001-general-add-external-links-to-search-results-20376.patch ajouté

Dans combo/data/search_indexes.py tu déclares une classe PageIndex qui existe déjà juste au dessus.

Oops copié/collé.

> Je ne vois pas trop l'usage de l'option « --skip-external-links-collection » dans la commande update_index, tu as une idée en tête ?

Pas particulièrement, j'ai du sur le moment trouver utile de pouvoir zapper cette partie du code pour arriver directement à la partie indexation.

> Je n'arrive pas trop à voir comment va être gérée la visibilité des résultat de recherche : si une démarche n'est accessible qu'à certains utilisateurs, comment s'assurer qu'elle ne pas pas être visible à d'autres dans les résultats de recherche ? (peut-être que c'est déjà un "soucis" sur les pages, en fait, j'ai jamais bien fait attention)

Aujourd'hui l'indexation des pages se fait uniquement sur les pages anonymes; pour les liens externes ça devrait être le cas aussi pour les liens externes (même si là pour les interrogations à w.c.s. ça fait déjà uniquement récupérer les infos accessibles à tout le monde). (et tout ça devrait évoluer à un moment)

> Je me disais que les keywords pourraient en fait avoir un boost=2 (...)

En fait je me suis rendu compte que le boost était un peu plus compliqué que ça (et le boost=1.5 sur le titre est là pour faire joli), ça demande aussi des interventions au moment de la requête et j'ai préféré laisser ça de côté.

~~

Patch mis à jour pour retirer la partie dupliquée par erreur et zapper les pages/cellules non visibles à l'usager lors de la collecte des liens.

#6 - 02 janvier 2018 18:14 - Thomas Noël

En fait, l'usage de "keywords" n'a pas, selon moi, été prévu dans un optique SEO, c'est plutôt un outil technique pour un jour construire une navigation transverse ("toutes les démarches pour les enfants"). Donc je ne suis pas trop chaud pour les envoyer dans le texte à indexer, la description doit suffire. Mais cette histoire de keywords techniques c'est peut-être dans ma tête seulement...

Et sinon, on est d'accord qu'avec ce patch, seuls les formulaires accessibles en mode anonyme seront indexés ?

#7 - 02 janvier 2018 18:25 - Frédéric Péters

> En fait, l'usage de "keywords" n'a pas, selon moi, été prévu dans un optique SEO, c'est plutôt un outil technique pour un jour construire une navigation transverse ("toutes les démarches pour les enfants"). Donc je ne suis pas trop chaud pour les envoyer dans le texte à indexer, la description doit suffire. Mais cette histoire de keywords techniques c'est peut-être dans ma tête seulement...

À lire #6194 les mots-clés ont été ajoutés en douce, sans en réfléchir l'usage… Je dirais que de mon côté l'intention était l'indexation mais que c'est rapidement devenu un moyen pas cher pour distinguer des démarches (influencé par spip qui a dévié comme ça). Même si aujourd'hui ça sert de distinction technique, c'est quand même rare, et je serais pour les considérer pour l'indexation. Aussi dans l'idée que les pages gagneront aussi des mots-clés et que ça sera avant tout pour l'indexation (#7427).

> Et sinon, on est d'accord qu'avec ce patch, seuls les formulaires accessibles en more anonyme seront indexés ?

Oui, uniquement les contenus accessibles aux usagers non identifiés sont indexés (pages ou démarches). (noté comme un truc à faire évoluer)

#8 - 02 janvier 2018 19:03 - Thomas Noël

> Même si aujourd'hui ça sert de distinction technique, c'est quand même rare

Allez, ça me va aussi de ne plus en faire un machin technique, je crois bien qu'on l'a en fait jamais utilisé nulle part (je pense que les mots clés ne sont utilisés nulle part, du moins à ma connaissance)

#9 - 02 janvier 2018 19:03 - Thomas Noël

Donc ack.

#10 - 02 janvier 2018 19:09 - Frédéric Péters

> je pense que les mots clés ne sont utilisés nulle part

On a utilisé "mobile" comme mot-clé à un moment, pour faire la "vue mobile" de w.c.s.; mais c'est bien vieux et je ne pense pas qu'on utilise ça

ailleurs.

#11 - 03 janvier 2018 11:34 - Brice Mallet

Frédéric Péters a écrit :

> On a utilisé "mobile" comme mot-clé à un moment, pour faire la "vue mobile" de w.c.s.; mais c'est bien vieux et je ne pense pas qu'on utilise ça ailleurs.

Je confirme, dans le cas de Meaux : https://meaux.test.au-quotidien.com/backoffice/forms/8/

#12 - 03 janvier 2018 11:43 - Frédéric Péters

- Statut changé de En cours à Résolu (à déployer)

```
commit 273f1b003293f90813b2c99611a8c8051c7f48d7
Author: Frédéric Péters <fpeters@entrouvert.com>
Date:   Tue Dec 26 16:22:29 2017 +0100

    general: add external links to search results (#20376)
```

#13 - 23 décembre 2018 15:14 - Frédéric Péters

- Statut changé de Résolu (à déployer) à Solution déployée

Fichiers

0001-general-add-external-links-to-search-results-20376.patch	13,3 ko	26 décembre 2017	Frédéric Péters
0001-general-add-external-links-to-search-results-20376.patch	12,9 ko	02 janvier 2018	Frédéric Péters