Module Moteur de recherche ASE

Fonctionnalités

Le moteur de recherche ASE est un moteur dit Full Text. Il indexera non seulement le contenu textuel que vous insèrerez directement dans Automne (au niveau des pages ou des modules) mais il indexera aussi tout le contenu présent dans les fichiers insérés dans Automne.

Types de fichiers supportés :

- Fichiers PDF (.pdf),
- Fichiers Microsoft Word et Microsoft Excel (.doc, .dot, .rtf, .docx, .xls, .xlsx),
- Fichiers Open Office (.sxw, .odt),
- Fichiers Microsoft PowerPoint (.ppt, .pps),
- Fichiers HTML (.htm, .html, .xhtml),
- Fichiers de texte brut (.txt, .csv).

De plus son interface modulaire permet d'ajouter assez simplement le support de types de fichiers supplémentaires.

Fonctions annexes :

Le moteur propose aussi un certain nombre de fonctions annexes parmi lesquelles :

- Gestion du pourcentage de pertinence.
- Support de la lemmatisation (exemple : une recherche sur cheval recherchera aussi chevaux).
- Support des "Mots vide" (ou stop words).
- Vérification orthographique des termes recherchés.
- Proposition de termes en relation avec la recherche pour affiner les résultats.
- Prise en compte d'un résultat pour affiner la recherche (cette fonction permet par exemple, pour une recherche sur "cheval", de préciser au moteur que nous sommes plus intéressé par les résultats parlant de "chevaux vapeur" que de "chevaux de course").
- Recherches complexes à l'aide d'opérateurs booléens (ET, OU, etc.).
- Filtres par types de documents ou par langue.
- Intégration dans le moteur de recherche du navigateur des internautes (protocole "OpenSearch".
- Support complet des droits d'Automne.
- Indexation et réindexation en temps réel.
- Exclusion d'arborescences de pages.
- Support des textes Japonais, Chinois, Coréens.
- etc.

Vous pouvez tester ces fonctionnalités sur la page de recherche du site Automne-cms.org.

Fonctions avancées :

En complément des fonctions ci-dessus qui sont disponibles nativement, il est possible en exploitant l'API PHP du moteur de compléter les fonctionnalités offertes avec les fonctions suivantes :

- Restriction dynamique de la zone d'arborescence recherchée.
- Restriction dynamique des résultats à certains modules ou à certaines catégories ou langues de modules.

Performances :

Le moteur ASE ainsi que la librairie Xapian ont été pensés pour pouvoir gérer de très gros volumes de données indexés tout en gardant un temps de réponse optimal. A l'heure actuelle, nos tests les plus poussés portant sur plusieurs dizaines de milliers de documents indexés (plusieurs centaines de Giga Octets de données) ont toujours donnés des temps de réponse inférieur à une seconde pour une recherche sur le corpus entier de documents avec un simple serveur dédié.

Installation

Pré requis :

En plus des pré requis d'Automne, ce module nécessite l'installation d'un certain nombre de composants supplémentaires sur le serveur d'hébergement :

- **Système d'exploitation :** Unix (Linux, Solaris, Mac OS X). **Ce module est incompatible avec Windows**.
- **Librairie Xapian Core** version 1.0.2 minimum et son **binding** (extension) PHP : Consultez le site Xapian.org pour plus de détail sur l'installation de ces deux éléments.
  Attention, la branche 1.2.x de Xapian n'est supportée qu'à partir de la version 0.73 du module.
- Pour convertir les documents indexés par le moteur de recherche, la précense d'un certain nombre de binaires est nécessaire. Ces binaires doivent tous être installés sur le serveur et être disponible dans le PATH du serveur.
  - **unzip**, **sed**, **iconv** : ces binaires sont normalement disponible sur les serveurs unix.
  - **pdftotext** : Disponible dans la suite XPDF ou dans le paquet Debian "*xpdf-utils*".
  - **catdoc**, **xls2csv** : Disponibles sur ce site ou dans le paquet Debian "*catdoc*".
  - **ppthtml** : Disponible sur ce site ou dans le paquet Debian "*ppthtml*".

Du fait du nombre important de binaires nécessaires à l'usage de ce module, l'usage d'un serveur dédié est pratiquement obligatoire. En effet, les fournisseurs d'hébergement mutualisé ne proposent pas ou très rarement ce type de configuration.

Si vous employez un serveur mutualisé, **rapprochez vous de votre hébergeur avant d'installer ce module pour savoir si il est en mesure de répondre à ces pré-requis.**

Installation et mise en route du module :

Le module se présente sous la forme d'un patch Automne. Son installation est donc simple, il vous suffit d'uploader le fichier disponible en téléchargement sur ce site directement dans l'interface d'administration d'Automne pour déclencher son installation automatique.

Pour cela, allez dans l'interface d'administration d'Automne avec un compte administrateur puis, dans le panneau latéral, allez dans "Administration" > "Paramètres Serveur" > "Mises à jour". Fournissez le fichier du module dans le champ "Fichier de mise à jour" puis cliquez sur valider.

Une fois le module installé, il est nécessaire de configurer les différents modules qui doivent être indexés par le moteur de recherche. Voir le paragraphe "fonctionnement" ci-dessous pour plus d'informations.

Pour savoir si l'installation de tous les composants indiqués dans les pré-requis est correctement effectué, une fois le module installé, allez dans le panneau latéral > Moteur de recherche > Configuration du moteur. Vous pourrez y voir l'état de fonctionnement de chacun des composants utilisés par le module :



Mise à jour du module

Pour mettre à jour le module, il vous suffit, comme pour son installation d'uploader le fichier de la dernière version du module dans l'interface d'administration d'Automne. Cela déclenchera sa mise à jour automatique.

Fonctionnement

Rangée :

Lors de son installation, le module créera une nouvelle rangée "*[ASE] Moteur de Recherche*".

Insérer cette rangée dans une page vous permet de créer la partie cliente du moteur de recherche qui sera employée par les internautes.

Cette rangée contient le code XML suivant :

<row>
<block module="ase" type="search"></block>
</row>

Vous pouvez ajouter à ce code XML un attribut optionnel "*language*" pour contrôler la langue des résultats fournis par le moteur (exemple : language="fr"). Si cet attribut n'est pas présent, la langue de la page dans laquelle se trouve la rangée sera utilisée.

Paramètres :

Le module contient un certain nombre de paramètres qu'il est nécessaire de correctement configurer pour assurer un fonctionnement optimal. Ces paramètres sont accessibles dans l'administration d'Automne via la page "Paramètres" du module ou directement en éditant le fichier /automne/classes/modules/ase_rc.xml sur le serveur.

Voici le descriptif de ces paramètres :

- **DOCUMENT MAX WORDS TO INDEX** : Nombre de mots maximum à indexer par objet. Par défaut cette valeur est fixée à 20 000 mots. Au delà de ce nombre, les mots d'un document ne sont plus indexés. Augmenter cette valeur peut ralentir sensiblement le temps d'indexation des documents.
- **DOCUMENT MAX INDEXABLE DOCUMENT LENGTH** : Nombre de caractères maximum à indexer par objet. Par défaut cette valeur est fixée à 300 000 caractères. Au delà de ce nombre, les caractères d'un document ne sont plus indexés. Augmenter cette valeur peut ralentir sensiblement le temps d'indexation

des documents.

- **DOCUMENT MIN INDEXABLE WORD LENGTH** : Nombre minimum de caractères pour indexer un mot. Par défaut cette valeur est fixée à 0 : tous les mots sont indexés quelque soit leur longueur.

- **XAPIAN RESULTS EXCLUDED ROOTS** : Identifiant des pages racines des sections d'arborescences exclues de l'indexation. Ce paramètre permet d'exclure des zones entières de l'arborescence de pages d'Automne de l'indexation (pour, par exemple, ne pas indexer une zone de test). Vous pouvez spécifier plusieurs pages en séparant leurs identifiants avec des virgules.

- **XAPIAN SEARCH OPENSEARCH PAGES** : Ce paramètre permet de spécifier quelle sera la ou les pages utilisées comme pages de résultats pour le protocole OpenSearch qui permet d'intégrer le moteur de recherche dans la barre de recherche du navigateur de l'internaute.
Si vous n'avez qu'un seul site dans Automne, il vous suffit de spécifier ici l'identifiant de la page. Si vous gérez plusieurs sites dans un même Automne, vous pouvez spécifier une page de recherche par site à l'aide de la syntaxe suivante :
"*id-site,id-page;id-site,id-page;...*"

- **XAPIAN SEARCH MAX RESULTS PER PAGES** : Ce paramètre permet de spécifier le nombre maximum de résultats par page à afficher dans les résultats d'une recherche. Valeur par défaut : 50 résultats par page. Augmenter ce paramètre peut avoir un impact négatif sur le temps d'affichage des résultats de recherche.

- **XAPIAN SEARCH DEFAULT RESULTS PER PAGES** : Ce paramètre permet de spécifier le nombre de résultats par page à afficher dans les résultats d'une recherche. Valeur par défaut : 20 résultats par page. Augmenter ce paramètre peut avoir un impact négatif sur le temps d'affichage des résultats de recherche.

- **XAPIAN SEARCH MIN MATCH RESULTS CHECK** : Ce paramètre permet de spécifier jusqu'où le moteur de recherche doit se montrer précis dans le décompte du nombre total de résultats qu'il propose pour une recherche donnée. Compter précisément le nombre de résultats pour une recherche prends du temps. Pour limiter cette perte de temps, à partir du moment ou une recherche retourne plus de résultats que le nombre indiqué dans ce paramètre, le nombre total de résultats indiqués sera une approximation. Valeur par défaut : 100 résultats. Augmenter ce paramètre peut avoir un impact négatif sur le temps d'affichage des résultats de recherche.

- **XAPIAN SEARCH EXPAND SET MAX NUMBER** : Ce paramètre permet de spécifier le nombre de mots qui seront proposés pour affiner une recherche donnée. Valeur par défaut : 10 mots. Augmenter ce paramètre peut avoir un impact négatif sur le temps d'affichage des résultats de recherche.

- **DOCUMENT TITLE WDF** : Ce paramètre permet de spécifier le poids des mots du titre d'un résultat par rapport aux autres mots du résultat. Valeur par défaut : 2. Un mot, si il est contenu dans le titre du résultat aura ainsi un poids double par rapport aux autres mots du résultats. Augmenter ce paramètre peut avoir un impact négatif sur la pertinence des résultats proposés par le moteur.

- **USER AGENT REJECTED** : Ce paramètre permet de bannir du moteur de recherche les internautes possédant ces valeurs dans leur User Agent. Concrètement, ce paramètre permet d'éviter que les robots des moteurs de recherches (Google, Yahoo, Bing, etc.) ne puisse utiliser le moteur de recherche car de part sa conception, il se comporte alors comme un "puit d'indexation" dont les robots des moteurs ne ressortent pas ce qui est très négatif pour les performances du site. Par ailleurs, l'indexation par les robots du moteur de recherche de votre site n'apporte rien puisque le contenu qui y est référencé se trouve déjà dans le reste du site ou il sera naturellement référencé par les robots.

Indexation des pages :

Pour indexer le contenu des pages d'Automne, il faut suivre les étapes suivantes :

1. Vérifier que le paramètre Automne "*Activer l'utilisation des pages imprimables*" est actif Dans le panneau latéral > Administration > Paramètres Automne.
2. Vérifier que les modèles de pages employés dans le site possèdent bien des espaces clients imprimables. Dans le panneau latéral > Modèles > Modèles de pages > Modifiez les modèles > Onglet "*Impression*", vérifiez qu'au moins un espace client est sélectionné.

En effet, le moteur de recherche indexe les pages imprimables du site, ce qui permet de ne n'indexer que les éléments de contenu pertinents des pages. Si vous avez fait des modifications sur les deux points ci-dessus, pensez à régénérer l'ensemble de votre site (dans le panneau latéral > Administration > Gestion des scripts > Tout Régénérer).

Eventuellement, si vous souhaitez exclure de l'indexation certaines sections d'arborescence de votre site, employez le paramètre XAPIAN RESULTS EXCLUDED ROOTS (voir l'explication sur les paramètres ci-dessus).

Puis, dans le panneau de gestion des modules indexés du moteur de recherche (dans le panneau latéral > Moteur de recherche > Modules indexés), Sur la section "Gestion des pages", cliquez sur le bouton "Réindexer".



Cette opération permettra de créer l'index initial de votre ou de vos sites. Une fois les scripts d'indexation terminés (vous pouvez contrôlez leur progression dans le panneau latéral > Administration > Gestion des scripts), le moteur est prêt à être employé coté client du site.

Il n'est pas nécessaire de recommencer cette opération d'indexation du site à chaque modification de votre contenu. En effet, le moteur tiendra à jour son index automatiquement au fur et à mesure des changements que vous ferez sur le contenu des pages d'Automne. Vous ne devez faire une réindexation complète que dans le cas de fortes modifications sur l'arborescence des pages d'Automne (par exemple lors d'une duplication de branche d'arborescence).

Indexation des modules Polymod :

Pour indexer le contenu d'un module Polymod d'Automne, il faut suivre les étapes suivantes :

1. Editez la configuration de l'objet Polymod du module à indexer. De nouvelles options liées au moteur de recherche seront disponibles permettant de contrôler l'indexation de ce type d'objets :



   Cochez la case "*Indexé dans le moteur de recherche*" et dans le champ texte qui suit, indiquez quelle doit être l'adresse utilisée par le moteur de recherche pour pouvoir renvoyer l'internaute du résultat fourni par le moteur vers l'objet dans le site.
   Sur la capture ci-dessus, vous pouvez voir en exemple comment est géré l'adresse des éléments du blog du site Automne-cms.org : Les billets français du blog sont renvoyés vers la page 5 du site alors que les billets anglais du blog sont eux renvoyés vers la page 316 qui correspond à la page anglaise du blog d'Automne-cms.org. Vous pouvez mettre dans ce champ toute combinaison nécessaire pour pouvoir

générer l'adresse de l'objet affiché dans les résultats du moteur de recherche.

2. Editez chaque champ qui composent l'objet à indexer. Vous trouverez dans les propriétés des champs de l'objet à indexer une option supplémentaire "*Indexé dans le moteur de recherche*". Cela vous permettra de contrôler champ par champ si oui ou non le contenu doit être indexé par le moteur. Vous pourrez ainsi filtrer l'information indexé pour ne conserver que les informations pertinentes pour le moteur de recherche.

Puis, dans le panneau de gestion des modules indexés du moteur de recherche (dans le panneau latéral > Moteur de recherche > Modules indexés), Sur la section correspondant au module Polymod à indexer, cliquez sur le bouton "Réindexer".



Cette opération permettra de créer l'index initial de votre module. Une fois les scripts d'indexation terminés (vous pouvez contrôlez leur progression dans le panneau latéral > Administration > Gestion des scripts), le moteur est prêt à être employé coté client du site.

Il n'est pas nécessaire de recommencer cette opération d'indexation du module à chaque modification de votre contenu. En effet, le moteur tiendra à jour son index automatiquement au fur et à mesure des changements que vous ferez sur le contenu du module Polymod. Vous ne devez faire une réindexation complète que dans le cas de modifications sur la structure des objets indexés du module (par exemple lors de l'ajout ou la suppression d'un nouveau champ indexé).

Interface d'administration :

L'interface d'administration du module est minimale et se contente de deux panneaux qui permettent de voir l'état de fonctionnement du moteur pour chaque modules ainsi que l'état des différents composants sur le serveur. Voici une vue de ces deux panneau lors du fonctionnement optimal du moteur :

Panneau de gestion des modules indexés :



Panneau de gestion de la configuration du moteur :



Droits d'accès :

Le moteur respecte parfaitement tous les droits d'accès des utilisateurs du site. Un utilisateur possédant des droits d'accès restreint à une partie du site ou des modules n'aura comme résultat pour ces recherches que les éléments qu'il est sensé pouvoir voir.

Fonctions avancées :

Pour modifier le comportement par défaut de la rangée proposant les résultats de recherche, vous pouvez étudier et modifier le fichier /automne/templates/mod_ase_search.php qui est utilisé pour créer l'affichage coté client de la rangée fournie avec le moteur.

N'hésitez pas à poser des questions sur son fonctionnement dans le forum ou à compléter cette documentation via les contributions ci-dessous.