GSA | Remarques en situation

Google Search Appliance

Sujets supplémentaires et questions-réponses

Mars 2014

Sujets supplémentaires et questions-réponses

Ce document traite des sujets et des questions qui n'ont pas été abordés dans les autres documents "GSA en pratique".

À propos de ce document

Les recommandations et informations rassemblées dans ce document sont le résultat de notre travail sur le terrain avec de nombreux clients et partenaires dans des environnements variés. Nous tenons à les remercier chaleureusement d'avoir partagé avec nous leurs expériences et leurs observations.

Thèmes abordés	Ce document inclut un certain nombre d'astuces rassemblées par l'équipe de déploiement GSA de Google, ainsi que des conseils sur l'utilisation de la boîte à outils d'administration de GSA.
Lecteurs cibles	Administrateurs et développeurs GSA
Environnement informatique	Système GSA configuré pour la recherche publique sécurisée
Phases de déploiement	Configuration du système GSA et post-déploiement
Autres ressources	● Le site Web Learngsa.com fournit des ressources pédagogiques sur le système GSA. ● La documentation produit de GSA fournit des informations complètes sur le système. ● Le Portail d'assistance Google for Work permet d'accéder à l'assistance Google.

Sommaire

Chapitre 1 : Utilisation d'Apache en tant que proxy de filtrage

Présentation

Lorsque vous configurez Google Search Appliance (GSA), vous avez un contrôle limité sur la manière dont le contenu est exploré et transmis au GSA pour un traitement approfondi. Toutefois, l'ajout d'un serveur Apache en tant que proxy au sein de l'environnement de déploiement vous offre la possibilité de modifier le contenu au moment de l'exploration pour servir un certain nombre d'objectifs.

Le type de modification de contenu le plus fréquemment rencontré est le filtrage. Il permet d'ajouter du contenu sur les pages ou d'en supprimer au cours de l'exploration. La modification du contenu lors de l'exploration permet également de changer la manière dont le robot d'exploration consulte vos sources de contenu.

Vous pouvez utiliser Apache en tant que proxy de filtrage en suivant les deux étapes suivantes :

1. Configuration d'Apache en tant que proxy
2. Configuration de votre système GSA pour l'utilisation du serveur proxy

Configuration d'Apache en tant que proxy

Pour configurer Apache en tant que proxy, vous devez ajouter les lignes de code ci-dessous dans le fichier `httpd.conf` :

```
LoadModule proxy_module modules/mod_proxy.so
LoadModule proxy_http_module modules/mod_proxy_http.so
Listen 8080

<VirtualHost *:8080>

ProxyRequests On
<Proxy *>
  Order Deny,Allow
  Deny from all
  Allow from 192.168.0.20
</Proxy>

### Ajouter les filtres ici ###

</VirtualHost>
```

Les deux premières lignes de cette configuration permettent simplement de charger le module `mod_proxy` et de demander au serveur Apache d'écouter l'activité du port 8080.

La section suivante définit un hôte virtuel sur le port 8080 et le transmet aux requêtes du proxy (plutôt que de traiter les résultats comme un serveur Web lambda).

Verrouillage de la configuration

Si vous effectuez cette configuration sur une machine publique, vous devez absolument la sécuriser davantage pour éviter qu'elle ne soit utilisée à mauvais escient. Pour ce faire, il suffit d'utiliser des règles IP simples n'autorisant que les requêtes du proxy provenant de l'adresse IP 192.168.0.20, c'est-à-dire du système GSA.

Test du serveur proxy

Une fois que le serveur a démarré, testez-le en procédant comme suit :

1. Exécutez le protocole Telnet sur le port 8080.
2. Saisissez la commande "GET http://www.google.com/".

Si le serveur fonctionne, la source provenant de la page d'accueil de Google est renvoyée. Si vous n'êtes pas connecté à partir d'une adresse IP autorisée, une erreur 403 (accès refusé) est renvoyée.

Configuration de votre système GSA pour l'utilisation du serveur proxy

Pour que le système GSA utilise le serveur proxy, configurez-le comme suit :

1. Dans la console d'administration du système GSA, accédez à **Sources de contenu > Exploration du Web > Serveurs proxy** (avant la version 7.2 : **Explorer et indexer > Serveurs proxy**).

2. Saisissez les formats d'URL devant passer par le proxy, l'adresse IP ou le nom de domaine complet, ainsi que le port du serveur proxy que vous avez configuré.

3. Cliquez sur **Enregistrer** (avant la version 7.2 : **Enregistrer la configuration des proxys du robot**).

Il se peut que vous souhaitiez utiliser le proxy pour explorer la totalité du contenu. Il suffit alors de saisir "/" dans le format d'URL. Le proxy peut également vous permettre d'explorer un type de contenu spécifique (images ou vidéos, par exemple). Dans ce cas, saisissez le format d'URL approprié.

Utilisation de plusieurs configurations de proxy

Si vous avez besoin de plusieurs configurations de proxy pour votre application, vous pouvez exécuter plusieurs instances d'Apache sur différents ports. Une autre solution consiste à utiliser une seule configuration Apache et à définir des filtres pour gérer le contenu en fonction de formats d'URL ou d'autres paramètres.

Création de filtres

Apache est compatible avec deux types de filtres :

- Entrée
- Sortie

Pour les proxys, l'entrée est la requête que le système GSA envoie au serveur Web de destination. La sortie est la réponse envoyée par le serveur Web au système GSA. Pour la plupart des applications, un filtre de sortie doit être créé.

Apache dispose de quelques directives pour la création de filtres de sortie, notamment :

- SetOutputFilter
- AddOutputFilterByType

Les filtres sont simplement définis dans le bloc `Proxy Virtual Host`.

Directive SetOutputFilter

La directive SetOutputFilter permet d'appliquer un filtre sur TOUT le contenu transitant par le proxy :

```
# Filtrer les balises Meta robots
ExtFilterDefine fixrobots mode=output intype=text/html \
cmd="/bin/sed -r 's/(noarchive|noindex|nofollow)>//g'"
SetOutputFilter fixrobots
```

Dans cet exemple, nous avons défini un filtre externe intitulé "fixrobots" qui transmet simplement l'entrée standard (le document demandé) via sed, en supprimant les chaînes "noarchive", "noindex" et "nofollow".

Le système GSA peut ainsi ignorer les balises Meta robots intégrées. "sed -r" permet la manipulation simple et rapide des formats d'expression régulière et des chaînes simples. Mais, il s'avère aussi simple d'utiliser un script Perl, PHP ou shell. Apache transmet simplement le fichier en tant qu'entrée standard et renvoie au système GSA la sortie du filtre.

Directive AddOutputFilterByType

Avec la directive AddOutputFilterByType, vous pouvez aller un peu plus loin et appliquer un filtre basé sur un type MIME. Vous pouvez ainsi explorer le contenu qui n'est pas compatible avec GSA en mode natif (images, vidéos, etc.).

```
# Filtrer des fichiers vidéo
ExtFilterDefine filtrevideo mode=output outtype=text/html \
cmd="/home/ericl/mediaFilter.php"
AddOutputFilterByType filtrevideo video/x-msvideo video/mp4 video/ audio/mpeg
audio/ video/quicktime
```

Dans cet exemple, nous avons créé un filtre externe intitulé "filtrevideo" qui appelle le script externe "mediaFilter.php". Le script accepte les fichiers vidéo binaires en entrée et envoie en sortie un fichier HTML (métadonnées intégrées) et des vignettes.

Nous utilisons la directive **AddOutputFilterByType** pour spécifier les formats multimédias qui correspondent aux types de contenu pour lesquels nous souhaitons appliquer ce script.

Vous pouvez également modifier les en-têtes HTTP. Par exemple, vous pouvez remplacer la chaîne GSA "User-Agent" par une autre :

```
# Définir le nom de l'agent utilisateur du proxy
RequestHeader set User-Agent "Mozilla/4.0 (compatible; MSIE 7.0; Windows NT
6.0; SLCC1; .NET CLR 2.0.50727;)"
```

Ce code ne modifie pas les en-têtes du système GSA, car ces derniers ne transitent pas par le proxy. Il permet simplement de définir l'en-tête utilisé par Apache lors de la récupération d'une page.

Cela peut s'avérer utile si, vous avez besoin de définir un cookie, un agent utilisateur ou tout autre en-tête pour explorer votre contenu.

Une fois que le proxy et les filtres sont configurés, vous pouvez les tester en envoyant vos propres requêtes GET ou en :

1. configurant votre navigateur pour qu'il utilise le proxy ;
2. demandant des URL ;
3. affichant la source.

Lorsque tous les paramètres sont validés, ajoutez simplement les formats de proxy concernés au système GSA et lancez l'exploration. Vos documents mis en cache devraient afficher la sortie filtrée.

Autres ressources

- Page Apache mod_proxy module
- Guide de mise en cache Apache
- Page Apache mod_cache module

Chapitre 2 : Utilisation de la boîte à outils d'administration de GSA

Présentation

La boîte à outils d'administration de GSA est une bibliothèque d'outils Open Source dédiée aux administrateurs GSA. Vous pouvez télécharger chaque outil individuellement à partir de la boîte à outils d'administration de GSA.

Les outils disponibles figurent dans le tableau suivant et sont accompagnés d'une courte description.

Outil	Description
monitor.sh	Script de surveillance permettant de vérifier le traitement sur le système GSA.
load.py	Exécute des tests de charge du GSA.
authn.py	Serveur Web permettant de tester la SPI d'authentification.
authz.py	Serveur Web permettant de tester la SPI d'authentification.
Connect.java	Classe Java permettant de tester la connexion JDBC à la base de données.
cached_copy_checker.py	Script de surveillance permettant de vérifier le bon fonctionnement de l'exploration, de l'indexation et du traitement sur un système GSA.
sso.py	Serveur Web permettant de tester les sites de cookies. Peut être configuré pour offrir des fonctionnalités similaires à Oblix.
searchstats.py	Analyse du fichier de journalisation de recherche (taux d'erreur, nombre de requêtes par seconde, temps de réponse moyen).
smbcrawler.py	Script reproduisant la manière dont le système GSA explore SMB. Cet outil est utile pour résoudre les problèmes d'exploration du SMB quand le message d'erreur sur le système GSA ne permet pas de le faire.
reverse_proxy.py	Proxy inverse pouvant être utilisé pour placer les requêtes envoyées au système GSA en file d'attente afin de limiter le nombre de connexions simultanées. Cet outil a été développé en tant que démonstration de faisabilité et n'a pas été testé dans un environnement de production.
gsa_admin.py	Script Python permettant d'automatiser des tâches de la Console d'administration. Il est utilisé si l'API GData de la Console d'administration ne fonctionne pas (pour les versions du logiciel antérieures à 6.0, par exemple) ou si la fonctionnalité n'est pas disponible dans l'API).
remove-or-recrawl-urls.html	Page de l'outil d'aide HTML+JavaScript qui conçoit les flux permettant de supprimer des URL ou de les explorer une nouvelle fois.

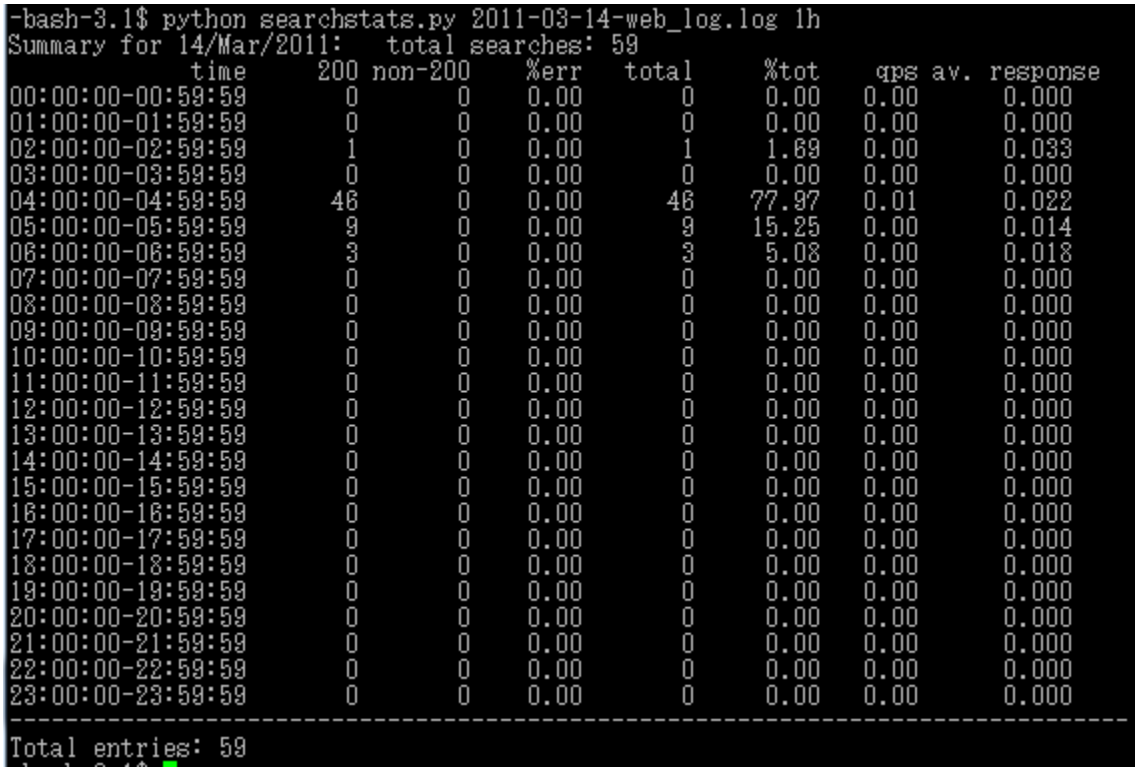
urlstats.py	Script Python permettant de générer des rapports relatifs aux URL figurant dans le système GSA.
ssoproxy.py	Script Python configurable qui transmet au proxy des règles de connexion aux systèmes SSO pour fournir au système GSA des cookies SSO d'exploration et de traitement.
CrawlReport.java	Classe Java permettant de récupérer, via la nouvelle API Admin (version du logiciel 6.0.0), le nombre d'URL explorées depuis la veille.
connectormanager.py	Gestionnaire de connecteurs simple et exemples de connecteurs accompagnés d'instructions sur l'écriture d'un nouveau connecteur.
Kerberos Validation Tool	Application HTML permettant de valider la configuration Kerberos/IWA (Keytab/Active Directory, etc.)
interactive-feed-client.html	Page HTML/JavaScript permettant de générer des flux XML à partir d'entrées telles qu'une URL ou un nom d'hôte, puis de les soumettre au système GSA.
search_report_xhtml.xsl	Feuille de style XSLT permettant de transformer des codes XML de rapports de recherche exportés en fichier XHTML intelligible.
search_results_analyzer.py	Outil permettant d'analyser les résultats de recherche et de comparer ceux renvoyés par deux systèmes GSA.
convert_cached_copy_to_feed.py	Outil permettant de convertir des versions mises en cache en flux de contenu en vue d'une migration.
fetch_secure.py	Permet de récupérer les résultats d'une recherche sécurisée envoyés par le système GSA en suivant toutes les redirections de connexion universelle.
GSA-GA.zip	Ressources d'intégration Google Analytics.
SHT	Self Help Tool (outil d'autoassistance)

Les sections suivantes passent en revue les quatre outils les plus utilisés de la liste ci-dessus :

- Comment analyser les journaux de recherche avec searchstats.py ?
- Comment automatiser des tâches de la Console d'administration ?
- Comment supprimer ou explorer une nouvelle fois les documents figurant dans l'index ?
- Vérification de la configuration Kerberos

Comment analyser les journaux de recherche avec searchstats.py ?

searchstats.py est un outil d'analyse du journal de recherche. La capture d'écran suivante illustre l'analyse d'un fichier de journal de recherche (2011-03-14-web_log.log) toutes les heures. Le journal de recherche est téléchargé depuis la Console d'administration GSA, sous **Rapports > Journaux de recherche** (avant la version 7.2 : **États et rapports > Journaux de recherche**).



```
-bash-3.1$ python searchstats.py 2011-03-14-web_log.log 1h
Summary for 14/Mar/2011:   total searches: 59
             time     200 non-200    %err   total     %tot    qps av. response
00:00:00-00:59:59       0       0    0.00       0     0.00   0.00       0.000
01:00:00-01:59:59       0       0    0.00       0     0.00   0.00       0.000
02:00:00-02:59:59       1       0    0.00       1     1.69   0.00       0.033
03:00:00-03:59:59       0       0    0.00       0     0.00   0.00       0.000
04:00:00-04:59:59      46       0    0.00      46    77.97   0.01       0.022
05:00:00-05:59:59       9       0    0.00       9    15.25   0.00       0.014
06:00:00-06:59:59       3       0    0.00       3     5.08   0.00       0.018
07:00:00-07:59:59       0       0    0.00       0     0.00   0.00       0.000
08:00:00-08:59:59       0       0    0.00       0     0.00   0.00       0.000
09:00:00-09:59:59       0       0    0.00       0     0.00   0.00       0.000
10:00:00-10:59:59       0       0    0.00       0     0.00   0.00       0.000
11:00:00-11:59:59       0       0    0.00       0     0.00   0.00       0.000
12:00:00-12:59:59       0       0    0.00       0     0.00   0.00       0.000
13:00:00-13:59:59       0       0    0.00       0     0.00   0.00       0.000
14:00:00-14:59:59       0       0    0.00       0     0.00   0.00       0.000
15:00:00-15:59:59       0       0    0.00       0     0.00   0.00       0.000
16:00:00-16:59:59       0       0    0.00       0     0.00   0.00       0.000
17:00:00-17:59:59       0       0    0.00       0     0.00   0.00       0.000
18:00:00-18:59:59       0       0    0.00       0     0.00   0.00       0.000
19:00:00-19:59:59       0       0    0.00       0     0.00   0.00       0.000
20:00:00-20:59:59       0       0    0.00       0     0.00   0.00       0.000
21:00:00-21:59:59       0       0    0.00       0     0.00   0.00       0.000
22:00:00-22:59:59       0       0    0.00       0     0.00   0.00       0.000
23:00:00-23:59:59       0       0    0.00       0     0.00   0.00       0.000
--------------------------------------------------------------------------------
Total entries: 59
```

Veuillez noter que le fichier "searchstats.py" est codé avec le langage de programmation de script Python. Pour exécuter ce script, l'environnement d'exécution Python est requis. Vous pouvez télécharger ce dernier sur la page Python prévue cet effet.

Après avoir installé l'environnement d'exécution, vous pouvez exécuter le script en utilisant une syntaxe similaire à celle affichée dans la capture d'écran.

Comment automatiser des tâches de la Console d'administration ?

Pour automatiser des tâches de la Console d'administration, il est préférable de passer par l'API d'administration de GSA. Cependant, l'API Admin ne permet pas d'effectuer toutes les tâches disponibles dans la Console d'administration. Par exemple, il n'est pas possible de s'en servir pour synchroniser des bases de données à distance. Pour traiter ces cas d'utilisation non compatibles, utilisez le script gsa_admin.py. Ce dernier peut se connecter à la Console d'administration et sélectionner automatiquement les éléments de menu souhaités.

Vous trouverez ci-dessous deux captures d'écran :

- Syntaxe de gsa_admin.py
- Exemple de commande permettant de synchroniser deux bases de données.

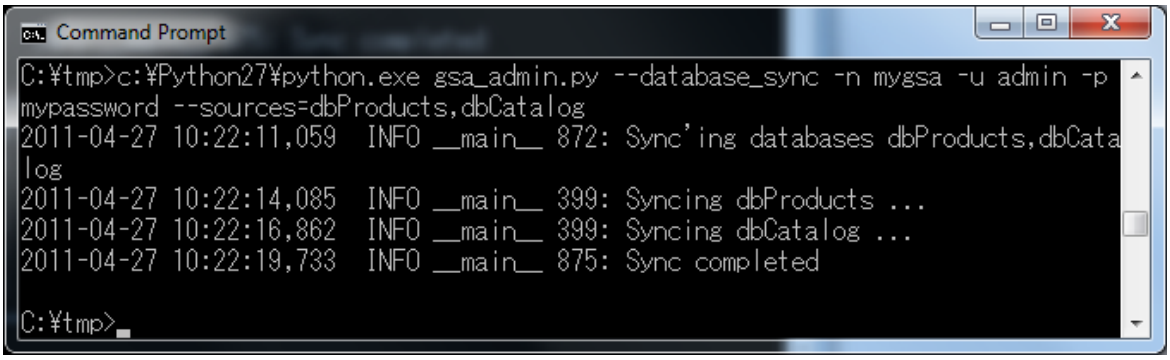
```
C:¥tmp>c:¥Python27¥python.exe gsa_admin.py --help
Usage: gsa_admin.py [options]

Options:
  -h, --help            show this help message and exit
  -f FILE, --input-file=FILE
                        Input XML file
  -o FILE, --output=FILE
                        Output file name
  -g SIGNPASSWORD, --sign-password=SIGNPASSWORD
                        Sign password for signing/import/export
  --cache-timeout=CACHETIMEOUT
                        Value for Authorization Cache Timeout
  --max-hostload=MAXHOSTLOAD
                        Value for max number of concurrent authz requests per
                        server
  --sources=SOURCES     List of databases to sync
                        (database1,database2,database3)
  --frontend=FRONTEND   Frontend used to export keymatches or related queries
  -v, --verbose         Specify multiple times to increase verbosity

  Actions::
    -i, --import        Import config file to GSA
    -e, --export        Export GSA config file from GSA
    -s, --sign          Sign input XML file
    -r, --verify        Verify signature/HMAC in XML Config
    -a, --set           Set Access Control settings
    -l, --all_urls      Export all URLs from GSA
    -d, --database_sync
                        Sync databases
    -k, --keymatches_export
                        Export All Keymatches
    -y, --synonyms_export
                        Export All Related Queries
    -z, --get-status    Get GSA Status

  GSA info:
    -n GSAHOSTNAME, --hostname=GSAHOSTNAME
                        GSA hostname
    --port=PORT         Upload port. Defaults to 8000
    -u GSAUSERNAME, --username=GSAUSERNAME
                        Username to login GSA
    -p GSAPASSWORD, --password=GSAPASSWORD
                        Password for GSA user
```

Syntaxe de commande



Syntaxe de commande permettant de synchroniser deux bases de données (dbProducts et dbCatalog)

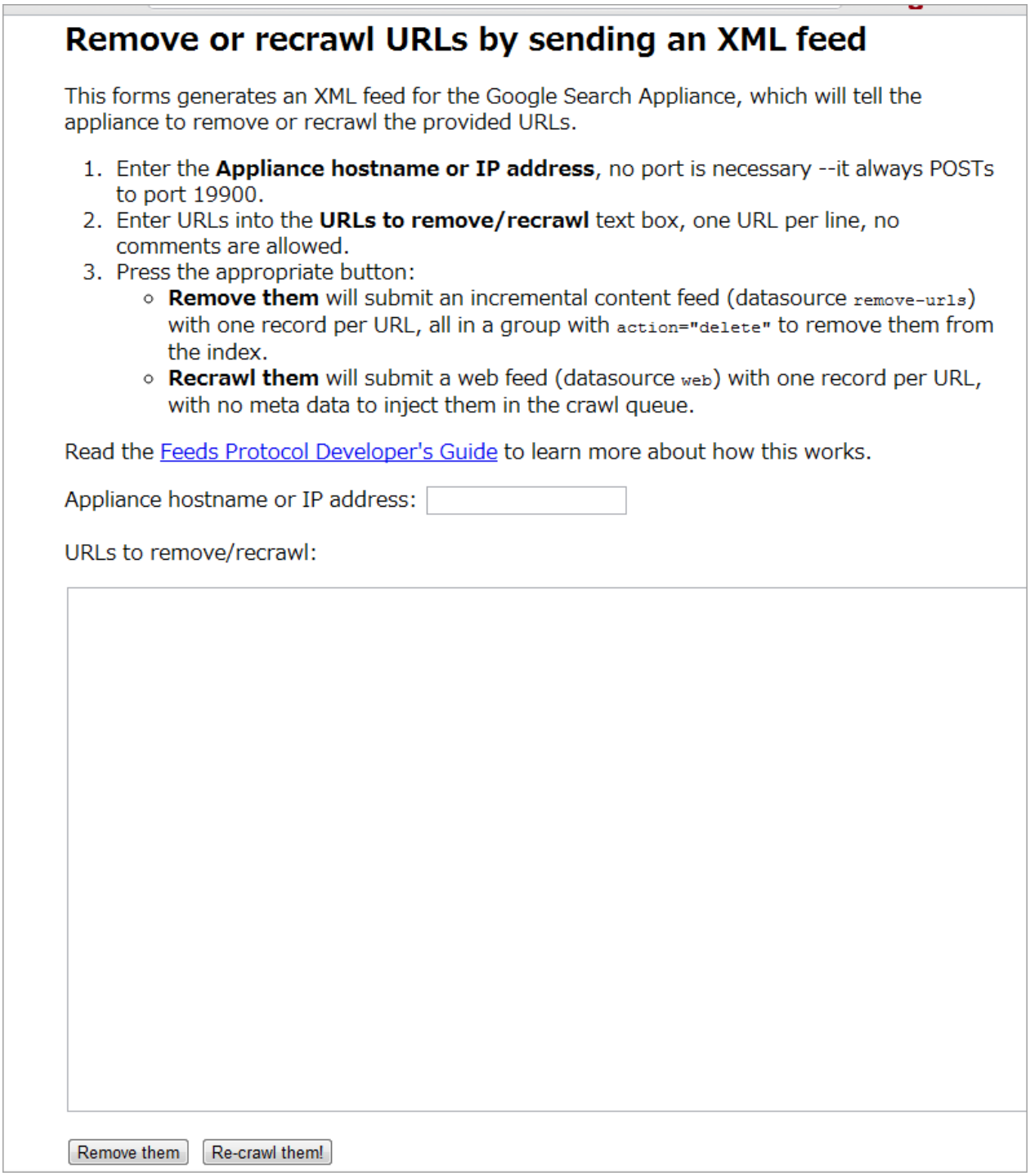
Comment supprimer ou explorer une nouvelle fois les documents figurant dans l'index ?

Si vous devez supprimer des documents spécifiques de l'index ou explorer ces derniers une nouvelle fois, remove-or-recrawl-urls.html s'avère un excellent outil. Il s'agit d'une page HTML intégrant du code JavaScript. Ce code accepte une liste d'URL en entrée, crée des fichiers de flux et envoie ces derniers au système GSA. En fonction des informations disponibles dans les listes d'URL, le système GSA supprime les URL spécifiées ou les explore une nouvelle fois. Ce script peut être ouvert et exécuté par n'importe quel navigateur Web compatible avec JavaScript. Il est actuellement compatible avec les trois paramètres d'entrée suivants :

- Nom d'hôte et adresse IP du système GSA
- Liste des URL à supprimer ou à explorer une nouvelle fois
- Choix entre une suppression et une nouvelle exploration

Cependant, veuillez noter qu'avant d'envoyer un flux au système de recherche, l'adresse IP de votre navigateur Web doit figurer dans la liste blanche d'adresses IP. Vous pouvez afficher ou modifier la liste blanche sur la page **Sources de contenu > Flux** (avant la version 7.2 : **Explorer et indexer > Flux**) de la Console d'administration.

La capture d'écran suivante illustre la page remove-or-recrawl-urls.html.



Remove or recrawl URLs by sending an XML feed

This forms generates an XML feed for the Google Search Appliance, which will tell the appliance to remove or recrawl the provided URLs.

1. Enter the **Appliance hostname or IP address**, no port is necessary --it always POSTs to port 19900.
2. Enter URLs into the **URLs to remove/recrawl** text box, one URL per line, no comments are allowed.
3. Press the appropriate button:
   - **Remove them** will submit an incremental content feed (datasource `remove-urls`) with one record per URL, all in a group with `action="delete"` to remove them from the index.
   - **Recrawl them** will submit a web feed (datasource `web`) with one record per URL, with no meta data to inject them in the crawl queue.

Read the Feeds Protocol Developer's Guide to learn more about how this works.

Appliance hostname or IP address:

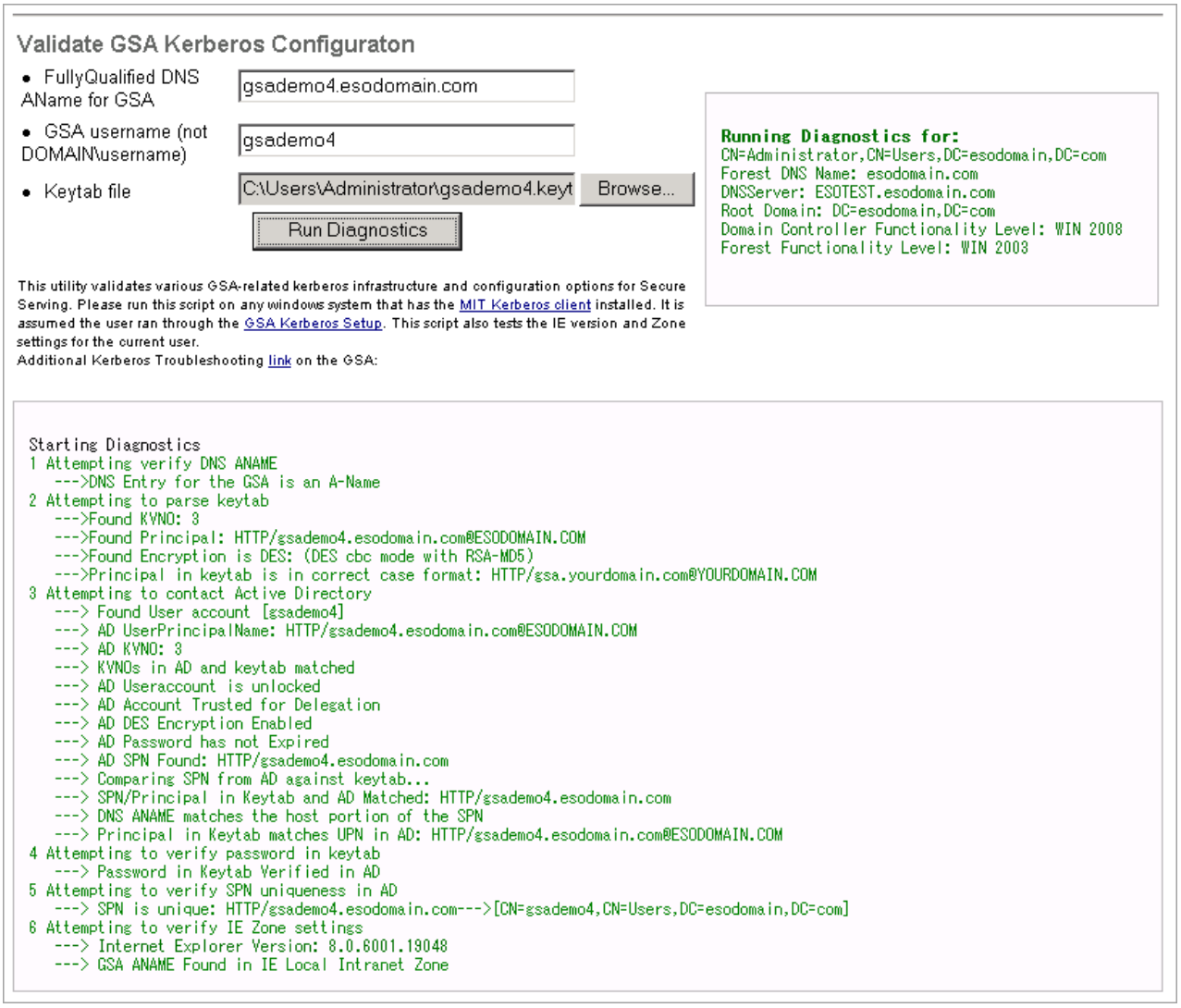
URLs to remove/recrawl:

Remove them | Re-crawl them!

Vérification d'une configuration Kerberos

Le fait de configurer le système de recherche afin d'effectuer une authentification Kerberos silencieuse implique de nombreux éléments hétérogènes. La configuration inclut notamment les éléments suivants : Active Directory, fichiers keytab, DNS, zone de sécurité Internet Explorer et méthodes de chiffrement de données. Si vous rencontrez un problème lors de la configuration de l'authentification Kerberos, vous pouvez utiliser l'outil de validation Kerberos pour effectuer une vérification rapide de chaque composant configurable.

La capture d'écran suivante illustre l'outil de validation Kerberos, ainsi que la liste des éléments configurables qu'il vérifie.



Pour en savoir plus, veuillez vous référer à la documentation en ligne suivante :

- Utilisation de l'utilitaire de validation de la configuration Kerberos. Soyez particulièrement attentif à la configuration système requise, qui est très stricte : Windows XP 32 bits, Vista, Windows 7.
- Dépannage de la configuration Kerberos et des recherches sécurisées

Chapitre 3 : Questions-réponses

Présentation

Cette section contient des questions-réponses sur des sujets d'ordre général relatifs au système GSA.

Pondération des résultats

Question : Pourquoi la page de résultats n'affiche-t-elle pas le document attendu, alors que j'effectue une pondération forte de son type de contenu ?

Réponse : Si le résultat que vous attendez ne s'affiche pas, cela peut être dû à la manière dont le système GSA récupère les résultats les plus pertinents pour votre terme de recherche. Lorsque vous effectuez une recherche (sur le terme "douleur", par exemple), le système GSA procède comme suit :

1. Le système localise d'abord les documents classés parmi les 1000 premiers (par PageRank) dans l'index contenant le terme. Veuillez noter que cette étape ne concerne aucunement la fréquence des termes ou une quelconque règle de pondération configurée pour le frontal.

2. Ces mille résultats passent ensuite par divers algorithmes exécutant des processus tels que le tri en fonction de la fréquence d'apparition du terme dans les documents, le score de pondération et de nombreux autres facteurs.

   Si l'URL que vous cherchez a un classement PageRank plus faible que les mille premiers documents de l'étape 1, elle ne sera pas incluse dans les algorithmes suivants prenant en compte les règles de pondération.

   Lorsqu'une URL est incluse dans les milles premiers documents obtenus à l'étape 1, même la règle de pondération la plus "forte" ne permettra par nécessairement à l'URL d'arriver en tête de liste. La pondération n'est qu'un des nombreux facteurs qui déterminent le classement final lors de l'étape 2 du processus.

Tri des métadonnées

Question : Comment se fait-il qu'un document donné dont je suis sûr de l'existence et qui dispose de métadonnées valides soit écarté lorsque je trie les résultats par métadonnées ?

Réponse : Le tri de métadonnées ne s'applique qu'aux 1000 premiers documents du classement (par PageRank). Consultez également la section Pondération des résultats.

Surveillance du système GSA

Question : Comment pouvez-vous contrôler l'état de fonctionnement du système GSA et des autres composants d'architecture déployés ? Est-il disproportionné d'exécuter une requête contenant un script sur l'index du système GSA ?

Réponse : Deux options permettent de contrôler l'état de fonctionnement. Beaucoup de nos clients vérifient simplement si le port 80 est ouvert. Certains vont plus loin en contrôlant le traitement, l'exploration et l'indexation. Pour obtenir des informations sur les stratégies de surveillance, consultez la section Configuration de la surveillance de l'article *Concevoir une solution de recherche*.

Le temps qu'il vous faut pour contrôler la solution dépend de la complexité et de l'exhaustivité de la stratégie de surveillance que vous souhaitez déployer. Différentes options s'offrent à vous, chacune ayant ses avantages et ses inconvénients.Voici quelques-unes de ces options :

- Utiliser SNMP : voir la section Objets SNMP sur la page d'aide associée à **Administration > Configuration SNMP**.
- Exécuter un script de surveillance.
- Exécuter un script Cached-Copy-Checker.
- Exécuter une requête pour une URL connue figurant dans l'index et s'assurer qu'un résultat est renvoyé à l'aide de XSLT.
- Exécuter une requête et s'assurer que l'état 200 est renvoyé.
- Appeler le frontal sans exécuter de requête pour vérifier que l'état 200 est renvoyé.

L'exécution d'une requête pour détecter une pulsation ne monopolise pas de ressources excessives. Si cette solution est actuellement mise en oeuvre et qu'elle fonctionne sur votre système, elle suffit certainement à détecter un problème et à effectuer un basculement. Vous pouvez améliorer légèrement la solution en créant un frontal exploitant le moins de ressources possibles (c'est-à-dire, sans correspondance, sans suggestion de requêtes, etc.) ou en envoyant une requête pour une URL connue à l'aide de la clause "info:".

Les autres outils disponibles incluent :

- monitor.sh : script de surveillance qui vérifie le traitement sur le système GSA.
- cached_copy_checker.py : script de surveillance qui vérifie que l'exploration, l'indexation et le traitement fonctionnent sur un système de recherche.
- searchstats.py : analyse du fichier journal de recherche (taux d'erreur, requêtes par seconde, temps de réponse moyen).
- urlstats.py : script Python générant des rapports relatifs aux URL sur le système GSA.
- search_report_xhtml.xsl : feuille de style XSLT permettant de transformer des scripts XML de rapport de recherche exportés en fichiers XHTML intelligibles.

Surveillez un connecteur en vérifiant le servlet Test Connectivity : http://[adresse-et-port-du-gestionnaire-de-connecteurs]/connector-manager/testConnectivity

En récupérant l'URL ci-dessus, vous devriez obtenir une réponse XML avec un état HTTP 200. La valeur <StatusId>0</StatusId> dans le fichier XML indique que tout fonctionne correctement.

Le format du code XML entrant est le suivant :

```
<CmResponse>
  <Info>Google Search Appliance Connector Manager 3.0.8 (build 3222 3.0.8-RC3
May 24 2013); Sun Microsystems Inc. Java HotSpot(TM) 64-Bit Server VM 1.6.0_33;
Windows Server 2008 R2 6.1 (amd64)</Info>
  <StatusCode>5501</StatusCode>
  <StatusId>0</StatusId>
</CmResponse>
```

N'oubliez pas d'autoriser l'accès de la machine depuis laquelle vous effectuez la surveillance.

Il est également recommandé de surveiller le processus Java qui exécute l'instance Tomcat du connecteur afin de vérifier le niveau d'utilisation de la mémoire et du processeur.

Suggestions de requêtes

Question : La fonctionnalité de suggestions de requêtes renvoie des termes choquants ou inappropriés. Comment puis-je empêcher ces termes d'être renvoyés dans les suggestions de requêtes ?

Réponse : Il est possible que des utilisateurs effectuent, sur le système GSA, des recherches portant sur des termes jugés inappropriés par votre organisation. Il y a deux manières d'empêcher ces termes d'être renvoyés par la fonctionnalité de suggestions de requête.

La méthode la plus efficace consiste à importer une liste noire de termes à l'aide de l'API GDATA. La fonctionnalité est compatible avec les expressions régulières. Vous pouvez donc créer une expression qui effectue une correspondance exacte ou partielle. Veuillez noter que, si votre système GSA est antérieur à la version 6.14, vous ne pouvez pas ajouter ou modifier de liste noire à partir de l'interface utilisateur de la Console d'administration. Vous devez le faire via les API. Pour en savoir plus sur la procédure à suivre, consultez la rubrique "Liste noire de suggestions de requêtes" dans la documentation de l'API GSA appropriée :

- Guide des API d'administration pour les développeurs : protocole
- Guide des API d'administration pour les développeurs : Java
- Guide des API d'administration pour les développeurs : .NET

Il existe un utilitaire JAVA Open Source permettant d'importer des listes noires à l'aide de l'API GDATA : http://code.google.com/p/gsa-admin-reports/.

Pour obtenir des détails sur son utilisation, consultez le wiki suivant : http://code.google.com/p/gsa-admin-reports/wiki/BlacklistEditor

Une autre méthode pour mettre en liste noire des termes de suggestion de requêtes consiste à modifier le code du frontal qui affiche les suggestions. Par exemple, employez l'extrait de code suivant :

```
<!--
**********************************************************************
Suggestions de variantes orthographiques sur la page des résultats (ne pas personnaliser)
**********************************************************************
-->
<xsl:template name="spelling">
  <xsl:if test="/GSP/Spelling/Suggestion">
    <p>
      <span class="p">
        <font color="{$spelling_text_color}">
          <xsl:value-of select="$spelling_text"/>
          <xsl:call-template name="nbsp"/>
        </font>
      </span>
      <xsl:variable name="apps_param">
        <xsl:choose>
          <xsl:when test="/GSP/PARAM[@name='exclude_apps']">
            <xsl:text disable-output-escaping='yes'>&amp;exclude_apps=</xsl:text>
            <xsl:value-of select="/GSP/PARAM[@name='exclude_apps']/@original_value" />
          </xsl:when>
          <xsl:when test="/GSP/PARAM[@name='only_apps']">
            <xsl:text disable-output-escaping='yes'>&amp;only_apps=</xsl:text>
            <xsl:value-of select="/GSP/PARAM[@name='only_apps']/@original_value" />
          </xsl:when>
        </xsl:choose>
      </xsl:variable>
      <a ctype="spell"
href="search?q={/GSP/Spelling/Suggestion[1]/@qe}&amp;spell=1&amp;{$base_url}{$apps_param}">
        <xsl:value-of disable-output-escaping="yes" select="/GSP/Spelling/Suggestion[1]"/>
      </a>
    </p>
  </xsl:if>
</xsl:template>
```

Vous devrez insérer une liste noire de termes dans la condition `if` afin que les suggestions ne s'affichent pas sur le frontal.

Par exemple, pour une correspondance exacte :

```
<xsl:if test="/GSP/Spelling/Suggestion and
GSP/Spelling/Suggestion[1]/@q!='terme_offensant'">
```

Pour une correspondance partielle :

```
<xsl:if test="/GSP/ Spelling/Suggestion and
not(contains(/GSP/Spelling/Suggestion[1]/@q,'terme_offensant'))">
```

Question : Comment puis-je réinitialiser complètement le contenu des suggestions de requêtes ?

Réponse : à partir de la version 7.2, vous pouvez supprimer toutes les suggestions de requêtes sur la page **Rechercher > Fonctionnalités de recherche > Suggestions** en cliquant sur **Réinitialiser** pour l'option **Réinitialiser les suggestions** (avant la version 7.2 : vous devez déposer une demande d'aide auprès de l'assistance Google).

Correspondances sur des requêtes partielles

Question : J'utilise le système GSA pour créer un outil de recherche de contacts. Mes utilisateurs se plaignent que leurs recherches sur des noms partiels ne renvoient aucun résultat. Que puis-je faire pour améliorer le fonctionnement de l'outil de recherche de contacts et renvoyer des correspondances en cas de requêtes partielles sur le nom ou le prénom ?

Réponse : pour l'instant, le système GSA ne permet pas d'effectuer des correspondances sur les termes partiels comportant des caractères génériques. Les utilisateurs ont l'habitude de rechercher des informations sur leurs collègues en saisissant uniquement un prénom ou nom de famille partiel. Il existe donc des solutions permettant d'optimiser le système GSA de façon à ce qu'il permette un tel fonctionnement.

Une première solution consiste à ajouter des métadonnées aux enregistrements relatifs aux contacts qui figurent dans le système GSA et d'y inclure une balise pour chaque combinaison des six premières lettres du prénom et du nom d'un contact. Par exemple, les douze balises Meta suivantes doivent être associées à l'enregistrement du contact "Jennifer Johnson" en tant que métadonnées supplémentaires :

j	j
je	jo
jen	joh
jenn	john
jenni	johns
jennif	johnso

Selon le mode d'acquisition du contenu relatif aux contacts de votre environnement, le script peut afficher ce contenu sur une page HTML explorée par le système GSA ou par l'intermédiaire d'un flux de contenu dans le système GSA.

Une autre solution consiste à inclure une liste personnalisée de synonymes de type nom dans la fonctionnalité d'extension de requête du système GSA afin de faire passer certains surnoms en noms complets.

L'utilisation de surnoms sur le lieu de travail pour se référer aux membres de l'organisation est une pratique courante. L'extension de requête permet d'intégrer les surnoms qui ne sont pas directement associés aux enregistrements des contacts en tant que métadonnées ou contenu. Il est possible de veiller à ce que les surnoms locaux soient associés aux surnoms internationaux. Par exemple, "Mathieu" peut être définit pour s'étendre à "Mat", "Matt" ou "Matthew".

La fonctionnalité de suggestion de requêtes peut également permettre d'optimiser la recherche de contacts. La fonctionnalité expérimentale d'affichage instantané des résultats peut être mise en oeuvre afin de fournir une base de données de suggestions de noms personnalisés et aider les utilisateurs à créer leurs requêtes de recherche d'informations sur leurs collègues.

Filtres de requêtes et modules OneBox

Question : Le module OneBox de recherche de contacts disparaît lorsque j'ajoute un filtre ("période", par exemple) à ma requête. Comment faire en sorte que le module OneBox de recherche de contacts reste affiché ?

Réponse : Les résultats du module OneBox de recherche de contacts sont renvoyés en fonction de correspondances des termes de la requête dans l'index. Lorsque vous essayez d'appliquer un filtre tel que "période" sur les résultats de la recherche organique, il se peut que vous ne souhaitiez pas qu'il s'applique sur les résultats de recherche de contacts. Les résultats de recherche de contacts proviennent également de correspondances dans l'index d'une collection de contacts. Ce résultat s'inscrit donc le fonctionnement normal du système GSA.

Par exemple, si les documents sur lesquels repose la collection ne contiennent pas de valeur de métadonnées de date, les résultats de recherche de contacts ne sont pas renvoyés par le requête de recherche incluant ce type de filtre.

La méthode consiste alors à s'assurer que les métadonnées requises par le filtre sont également incluses dans les enregistrements sur lesquels repose le module OneBox de recherche de contacts. Dans le cas de "période", si une date ne peut pas être définie dans les enregistrements de contacts, pensez à saisir une valeur de date statique avec une condition "OR" dans la requête de recherche sur le serveur.

Par exemple, joignez la date "01/01/1901" en tant que date modifiée pour tous les enregistrements de contacts. Ensuite, au moment d'appliquer le filtre sur le frontal, ajoutez "OR daterange:..02/01/1900" à la requête pour capturer les données statiques qui ont été ajoutées aux enregistrements.