Moteur de recherche Intranet IS

Table des matières

1. Introduction

Le but de ce TP est de réaliser un moteur de recherche de type IIS (Intranet Information Server) à l'aide de Index Server. Index Server est un outil de recherche pour IIS. Les clients peuvent formuler des requêtes en utilisant n'importe quel Web browser pour remplir les champs d'un formulaire HTML. Le serveur Internet envoie la requête au moteur de recherche qui trouve les documents pertinents et qui retourne les résultats sous forme de page HTML. A l'inverse de la plupart des outils de recherche, Index Server peut inclure des documents Word ou Excel, ce qui permet de ne pas avoir à formater ce type de documents au format HTML.

2. Formulaire de requête

Les utilisateurs soumettent leurs requêtes en remplissant des champs d'un formulaire. Avec Index Server, l'administrateur peut modifier le formulaire pour que l'utilisateur puisse effectuer des recherches par contenu, ou par propriété de document, comme l'auteur ou le sujet. L'administrateur crée un formulaire de recherche avec du HTML standard et la page peut être créée en quelques minutes.

2.1. Restrictions

La recherche peut s'effectuer sur le contenu d'une page Web et d'autres documents servis par IIS et Index Server. Le type de documents possibles sont HTML, Word, Excel, Powerpoint et documents de type texte. Avec Index Server, il est possible de chercher des mots multiples ou des phrases ou des mots et des phrases s'en approchant. Il est aussi possible de donner du texte libre, mot ou phrase complète que Index Server interprètera, identifiera les noms et les groupes nominaux et les utilisera pour effectuer la recherche.

Par exemple, si l'on entre la phrase suivante : *le grand amphithéâtre sera fermé demain pour cause de conférence académique*. Index Server identifiera les noms suivants : *grand, amphithéâtre, demain, cause, conférence, académique* ; et les groupes suivants : *grand amphithéâtre, cause de conférence académique, conférence académique*. Ces mots et groupes sont combinés à une restriction, pondérés pour le classement des résultats et envoyés au moteur. La recherche par texte libre est précédée de **$contents**.

2.2. Restrictions sur les propriétés

En addition aux recherches par contenu, les utilisateurs peuvent rechercher les propriétés enregistrées dans les objets. Ces propriétés incluent la taille des fichiers, les dates de création et de modification, les noms de fichiers, l'auteur, etc. Le client peut combiner recherche numérique (date, taille…) et textuelle (nom de fichier, auteur…). On peut aussi utiliser les comparaisons standard dans les requêtes : =, >, <, >=, <= et != pour les propriétés textuelles et numériques. De plus, pour les recherches textuelles, les possibilités identiques aux recherches par contenu sont intégrées.

2.3. La feuille de résultats

Index Server regroupe les résultats dans une page qui est retournée au client. L'administrateur peut limiter le nombre de résultats pour les clients. On peut par exemple envoyer 150 résultats par tranches de 10 par page. Index Server utilise le système de protection des fichiers de NTFS (NT File System). Les résultats ne portent que sur les fichiers pour lesquels l'utilisateur a un accès en lecture. En plus de retourner les propriétés de documents, Index Server peut générer un sommaire qui introduit brièvement le contenu du document.

2.4. Fonctionnement interne

Effectuer des recherches par l'intermédiaire de Index Server est un processus complexe qui interagit avec IIS. A cause de cette interaction, le processus de recherche emprunte le même modèle que IIS utilise pour effectuer des requêtes sur une base de données via ODBC (qui doit être préalablement installé). Le connecteur de base de données (IDC : « Internet Database Connector »), composant de IIS, convertit un formulaire d'une page HTML en une requête qui fonctionne avec ODBC. Cette particularité permet d'effectuer une requête sur n'importe quelle base de données ayant un pilote ODBC. Lorsqu'une base de données reçoit une requête, elle retourne le résultat, que le IDC convertit en page HTML et affiche la page sur l'écran de l'utilisateur.

Dans le modèle IIS, les fichiers « *.idc » aident le IDC à convertir les requêtes venant d'un formulaire HTML. Fonctionnant en tandem avec les fichiers « *.idc », les fichiers de type « *.htx » spécifient comment le résultat est formaté et affiché à l'utilisateur. Les mêmes types de fichiers sont utilisés de la même façon avec Index Server. Les fichiers « *.htx » contiennent cependant des particularités spécifiques aux fonctionnalités de Index Server.

Dans notre cas, le fichier formulaire est « default.htm » :

```
<HTML>
<HEAD>
<TITLE>Formulaire de recherche</TITLE>
</HEAD>
<BODY>
  <CENTER><IMG SRC="logo1.bmp">
  <FORM ACTION="search.idq?" methode="POST"><B><I>Entrez votre recherche ci-
dessous :</B></I><BR>
    <INPUT TYPE="TEXT" NAME="CiRestriction" SIZE="60" MAXLENGTH="100"
VALUE=" "><BR>
    <INPUT TYPE="SUBMIT" VALUE="Rechercher">
    <INPUT TYPE="RESET" VALUE="Effacer"><BR>
  </FORM>
</BODY>
</HTML>
```

Comme on peut l'observer, c'est un simple formulaire qui fait appel au fichier « search.idq ». L'utilisateur entre un nom à rechercher et le formulaire envoie la requête vers le fichier « search.idq ». Typiquement, les plages de recherche dont les répertoires virtuels de IIS.

3. Le fichier IDQ

Ce type de fichier (Internet Data Query) définit les paramètres de recherche comme l'étendue de pages de recherche, les restrictions et les résultats. Voici en détail le fichier « search.idq » utilisé dans le cas d'une recherche à partir de « default.htm » :

```
[Query]
CiColumns=filename,size,rank,characterization,vpath,DocTitle,write
CiFlags=DEEP
CiRestriction=%CiRestriction%
CiMaxRecordsInResultSet=150
CiMaxRecordsPerPage=10
CiScope=/
CiTemplate=/iissamples/ISSamples/search.htx
CiSort=rank[d]
```

[Query] identifie l'information qui suit comme une recherche.

CiColumns indique le type d'information à retourner dans la page des résultats.

CiFlags demande de chercher dans tous les sous-répertoires de la plage des pages.

CiRestrictions indique le terme à rechercher.

CiMaxRecordsInResultSet indique le nombre maximum de résultats à retourner (ici 150).

CiMaxRecordsPerPage détermine combien de résultats sont indiqués dans chaque page (ici 10).

CiScope indique où commencer la recherche ; ici la recherche commence à la racine de l'espace de stockage.

CiTemplate indique quel fichier utiliser pour mettre en forme le résultat (ici **search.htx**).

CiSort définit comment classer les résultats (ici par ordre décroissant).

4. Le fichier HTX

Ce type de fichier (HTml eXtension) est un fichier HTML qui contient des variables qui réfèrent aux données d'un résultat de recherche. Le fichier « search.htx » suivant est écrit pour fonctionner en tandem avec le fichier « search.idq » décrit précédemment :

```
<HTML>
<HEAD>
  <%if CiMatchedRecordCount eq 0%>
    <TITLE><%CiRestriction%> - Aucun document correspondant.</TITLE>
  <%else%>
    <TITLE><%CiRestriction%> - documents <%CiFirstRecordNumber%> &agrave;
<%CiLastRecordNumber%></TITLE>
  <%endif%>
</HEAD>
<BODY background="is2bkgnd.gif">
<CENTER><IMG SRC ="logo1.bmp" border=0><h4> a trouvé
<%CiMatchedRecordCount%> documents correspondant à votre
requête</h4></CENTER>
<H5>
   <%if CiMatchedRecordCount eq 0%>
     Aucun document correspondant &agrave; la requ&ecirc;te "<%CiRestriction%>".
   <%else%>
     Documents <%CiFirstRecordNumber%> &agrave; <%CiLastRecordNumber%> des
     <%if CiMatchedRecordCount eq CiMaxRecordsInResultSet%>
       meilleurs
     <%endif%>
     <%CiMatchedRecordCount%> correspondant &agrave; la requ&ecirc;te
"<%CiRestriction%>".
   <%endif%>
</H5>
<TABLE WIDTH=80%>
   <TD> <A HREF="default.htm">Nouvelle requ&ecirc;te</A> </TD>
   <%if CiContainsFirstRecord eq 0%>
     <TD ALIGN=LEFT>
       <FORM ACTION="search.idq" METHOD="GET">
         <INPUT TYPE="HIDDEN"
           NAME="CiBookMark" VALUE="<%CiBookMark%>" >
         <INPUT TYPE="HIDDEN"
           NAME="CiBookmarkSkipCount" VALUE="-<%EscapeRAW
CiMaxRecordsPerPage%>" >
         <INPUT TYPE="HIDDEN"
           NAME="CiMaxRecordsInResultSet" VALUE="<%EscapeRAW
CiMaxRecordsInResultSet%>" >
         <INPUT TYPE="HIDDEN"
           NAME="CiRestriction" VALUE="<%CiRestriction%>" >
         <INPUT TYPE="HIDDEN"
           NAME="CiMaxRecordsPerPage" VALUE="<%EscapeRAW
CiMaxRecordsPerPage%>" >
         <INPUT TYPE="HIDDEN"
           NAME="CiScope" VALUE="<%CiScope%>" >
         <INPUT TYPE="HIDDEN"
           NAME="TemplateName" VALUE="<%TemplateName%>" >
         <INPUT TYPE="HIDDEN"
           NAME="CiSort" VALUE="<%CiSort%>" >
         <INPUT TYPE="HIDDEN"
           NAME="HTMLQueryForm" VALUE="<%HTMLQueryForm%>" >
         <INPUT TYPE="SUBMIT"
```

```
            VALUE="Voir les <%CiMaxRecordsPerPage%> résultats
pr&eacute;c&eacute;dents">
        </FORM>
      </TD>
    <%endif%>
    <%if CiContainsLastRecord eq 0%>
      <TD ALIGN=RIGHT>
        <FORM ACTION="search.idq" METHOD="GET">
          <INPUT TYPE="HIDDEN"
            NAME="CiBookMark" VALUE="<%CiBookMark%>" >
          <INPUT TYPE="HIDDEN"
            NAME="CiBookmarkSkipCount" VALUE="<%EscapeRAW
CiMaxRecordsPerPage%>" >
          <INPUT TYPE="HIDDEN"
            NAME="CiMaxRecordsInResultSet" VALUE="<%EscapeRAW
CiMaxRecordsInResultSet%>" >
          <INPUT TYPE="HIDDEN"
            NAME="CiRestriction" VALUE="<%CiRestriction%>" >
          <INPUT TYPE="HIDDEN"
            NAME="CiMaxRecordsPerPage" VALUE="<%EscapeRAW
CiMaxRecordsPerPage%>" >
          <INPUT TYPE="HIDDEN"
            NAME="CiScope" VALUE="<%CiScope%>" >
          <INPUT TYPE="HIDDEN"
            NAME="TemplateName" VALUE="<%TemplateName%>" >
          <INPUT TYPE="HIDDEN"
            NAME="CiSort" VALUE="<%CiSort%>" >
          <INPUT TYPE="HIDDEN"
            NAME="HTMLQueryForm" VALUE="<%HTMLQueryForm%>" >
          <INPUT TYPE="SUBMIT"
            VALUE="Voir les <%CiRecordsNextPage%> résultats suivants">
        </FORM>
      </TD>
    <%endif%>
</TABLE>
<HR>
<%begindetail%>
<table border=0>
  <tr class="RecordTitle">
    <td align="right" valign="top" class="RecordTitle" style="background-color:blue;">
      <B><%CiCurrentRecordNumber%></B>.
    </td>
    <td><b class="RecordTitle">
      <%if DocTitle isempty%>
        <a href="<%EscapeURL vpath%>" class="RecordTitle"><%filename%></a>
      <%else%>
        <a href="<%EscapeURL vpath%>" class="RecordTitle"><%DocTitle%></a>
      <%endif%></b>
    </td>
  </tr>
  <tr>
    <td>
    </td>
    <td><b><i>R&eacute;sum&eacute; :  </i></b><%characterization%>
    </td>
  </tr>
  <tr>
    <td>
    </td>
    <td>
```

```
        <i class="RecordStats">
          <a href="<%EscapeURL vpath%>" class="RecordStats"
style="color:green;">http://<%server_name%><%vpath%></a>
          <br><%if size eq ""%>(taille et date inconnues)<%else%>taille  <%size%>
octets - <%write%> GMT<%endif%>
        </i>
      </td>
    </tr>
</table><br>
<%enddetail%>
</dl>
<P>
<%if CiMatchedRecordCount ne 0%>
  <HR>
<%endif%>

<TABLE WIDTH=80%>
  <TD> <A HREF="<%HTMLQueryForm%>">Nouvelle requ&ecirc;te</A> </TD>
    <%if CiContainsFirstRecord eq 0%>
    <TD ALIGN=LEFT>
      <FORM ACTION="search.idq" METHOD="GET">
        <INPUT TYPE="HIDDEN"
          NAME="CiBookMark" VALUE="<%CiBookMark%>" >
        <INPUT TYPE="HIDDEN"
          NAME="CiBookmarkSkipCount" VALUE="-<%EscapeRAW
CiMaxRecordsPerPage%>" >
        <INPUT TYPE="HIDDEN"
          NAME="CiMaxRecordsInResultSet" VALUE="<%EscapeRAW
CiMaxRecordsInResultSet%>" >
        <INPUT TYPE="HIDDEN"
          NAME="CiRestriction" VALUE="<%CiRestriction%>" >
        <INPUT TYPE="HIDDEN"
          NAME="CiMaxRecordsPerPage" VALUE="<%EscapeRAW
CiMaxRecordsPerPage%>" >
        <INPUT TYPE="HIDDEN"
          NAME="CiScope" VALUE="<%CiScope%>" >
        <INPUT TYPE="HIDDEN"
          NAME="TemplateName" VALUE="<%TemplateName%>" >
        <INPUT TYPE="HIDDEN"
          NAME="CiSort" VALUE="<%CiSort%>" >
        <INPUT TYPE="HIDDEN"
          NAME="HTMLQueryForm" VALUE="<%HTMLQueryForm%>" >
        <INPUT TYPE="SUBMIT"
          VALUE="Les <%CiMaxRecordsPerPage%> documents
pr&eacute;c&eacute;dents">
      </FORM>
    </TD>
    <%endif%>
    <%if CiContainsLastRecord eq 0%>
    <TD ALIGN=RIGHT>
      <FORM ACTION="search.idq" METHOD="GET">
        <INPUT TYPE="HIDDEN"
          NAME="CiBookMark" VALUE="<%CiBookMark%>" >
        <INPUT TYPE="HIDDEN"
          NAME="CiBookmarkSkipCount" VALUE="<%EscapeRAW
CiMaxRecordsPerPage%>" >
        <INPUT TYPE="HIDDEN"
          NAME="CiMaxRecordsInResultSet" VALUE="<%EscapeRAW
CiMaxRecordsInResultSet%>" >
        <INPUT TYPE="HIDDEN"
```

```
            NAME="CiRestriction" VALUE="<%CiRestriction%>" >
          <INPUT TYPE="HIDDEN"
            NAME="CiMaxRecordsPerPage" VALUE="<%EscapeRAW
CiMaxRecordsPerPage%>" >
          <INPUT TYPE="HIDDEN"
            NAME="CiScope" VALUE="<%CiScope%>" >
          <INPUT TYPE="HIDDEN"
            NAME="TemplateName" VALUE="<%TemplateName%>" >
          <INPUT TYPE="HIDDEN"
            NAME="CiSort" VALUE="<%CiSort%>" >
          <INPUT TYPE="HIDDEN"
            NAME="HTMLQueryForm" VALUE="<%HTMLQueryForm%>" >
          <INPUT TYPE="SUBMIT"
            VALUE="Les <%CiRecordsNextPage%> documents suivants">
        </FORM>
      </TD>
      <%endif%>
</TABLE>
<P><BR>
<%if CiTotalNumberPages gt 0%>
<CENTER>
  <P><B>
    Page <%CiCurrentPageNumber%> sur <%CiTotalNumberPages%></B>
  <P>
<%endif%>
</BODY>
</HTML>
```

On remarque ici l'utilisation de nombreuses commandes telles que :

- **CiRestrictions** : désigne les paramètres de recherche
- **CiMatchedRecordCount** : désigne le nombre de résultats en accord avec la recherche
- **CiLastRecordNumber** : désigne le numéro du dernier résultat
- **CiBookMark** : désigne les informations de chaque résultat telles que le nom, la taille
- **CiMaxRecordsPerPage** : désigne le nombre maximum de résultats sur une page
- **CiRecordsNextPage** : désigne le nombre des résultats suivants
- **CiCurrentRecordNumber** : désigne le nombre de résultats courant
- **CiCurrentPageNumber** : désigne le numéro de la page affichée
- **CiTotalNumberPages** : désigne le nombre total de pages avec des résultats

Mis à part ces quelques commandes propres à l'architecture Index Server, le reste du code est écrit en langage HTML, ce qui le rend plus facile et plus rapide à réaliser par l'administrateur.

5. Conclusion

Les trois fichiers utilisés pour effectuer la recherche (default.htm ; search.idq ; search.htx) sont situés dans le répertoire /iissamples/ISSamples. Lors d'une recherche, Index Server se constitue lui-même un cache pour y mettre les résultats et ainsi accélérer le traitement de la requête. Bien que ce soit un avantage certain, ceci nécessite de vider le cache et relancer le service IIS à chaque modification de l'un des trois fichiers concernés par la phase de recherche.

Grâce à ce TP, j'ai pu apprendre comment fonctionne un moteur de recherche sous IIS, savoir quelles opérations étaient effectuées derrière une recherche de quelques secondes. Cela sera je pense très utile pour la suite, les autres moteurs de recherche utilisant a priori les mêmes principes.