UNIVERSIDAD FASTA

Ingeniería en Informática – FIM 42 – Proyecto Final

Indexación de archivos adjuntos

CommonJobs

(Sistema de Recursos Humanos CommonSense)

Alumnos:

* Andrés Moschini.
* Matías José.
* Juan Diego Raimondi.

Director Funcional: Gabriel Buyatti.

Director Técnico: Ing. Alejandro Fantini.

Auditora: Ing. Ana Haydee Di Iorio.

Cátedra:

* Profesor titular: AS. Hilario Fernando Schechtel
* Profesor asociado: Ing. Roberto Giordano Lerena
* Profesor asociado: Lic. Alejandro Nikolic

Fecha de presentación: 19/04/2012

Tabla de Contenidos

Indexación de archivos adjuntos

Dado que *CommonJobs* debe permitir buscar en el contenido de los archivos adjuntos, por ejemplo curriculums, y estos pueden tener formatos muy diferentes se decidió realizar el siguiente diseño.

Interface para extracción de contenido

Cuando se realiza el upload de un archivo, este es analizado por una lista de extractores que satisfacen la interface IContentExtractor:

```
public interface IContentExtractor
{
    bool TryExtract(string fullPath, Stream stream, string fileName, out ExtractionResult result);
}

public class ExtractionResult
{
    public string ContentType { get; set; }
    public string PlainContent { get; set; }
}
```

El sistema almacena el resultado del primer extractor exitoso de manera de que pueda ser indexado por nuestra base de datos.

De esta manera será posible crear distintos extractores de forma modular. Ahora mismo, la configuración se está realizando al inicio de la aplicación, pero en un futuro podrían utilizarse como plugins.

IFilter

Microsoft utiliza en sus productos *Windows Indexing Service*, *Windows Desktop Search* y *SQL Server* la interface COM IFilter para la extracción de textos.

Decidimos no utilizarla como interface principal de extracción de textos ya que nos limitaría demasiado al momento de intentar migrar el sistema a otras plataformas y además dificulta el desarrollo de filtros propios. Pero dada la gran disponibilidad de plugins IFilter, decidimos crear una implementación de IContentExtractor que lo soporte.

Para utilizar un nuevo plugin IFilter con *CommonJobs*, es suficiente con instalarlo en el equipo que hace las veces de servidor web. Para verificar que filtros están disponibles en los equipos estamos utilizando Citeknet IFilterExplorer.

Pueden encontrarse gran cantidad de filtros gratutios en la Web, por ejemplo en Citeknet o IFilterShop.

Filtros IFilter utilizados actualmente

Además de los filtros que Windows Server 2008 trae de serie, instalamos los siguientes:

- Adobe PDF IFilter v6.0
- Office 2010 FilterPack
- Citeknet ZIP IFilter