Search Engine API

Client Name: Alagad Inc.

Current Business Process Description

This is a quick description of the current existing business process.

Currently, the search system indexes a set of HTML pages each time it is executed. Execution is trigged by a scheduled process. The search system is not very flexible and is tied to a database and to Mach-II.

When indexing, the system pulls a set of 10 URLs from a table in the database based on the last time indexed. Items which have not been indexed or which have least recently been indexed are given preference. Content is pulled from a URL. Content is extracted from the HTML and passed to Lucene which indexes it. Links are extracted from the content (if it's a local document) and are added into the table of URLs to index.

The system indexes content one link into external sites.

When searching, the request is passed to Lucene which returns matching items. Matched are summarized on the fly and presented to the user.

This system is limited in the following ways:

1) It is tied to Mach-II for configuration and execution.
2) It relies on a Microsoft SQL database to hold addresses which have been or need to be indexed.
3) The number of pages to index on one pass is limited.
4) It can not easily be integrated into other systems.
5) Only HTML content can be indexed.
6) Summarization is slow, reasonably inaccurate, and done on the fly when searching (so it's inefficient).
7) Not as configurable as I would like.

Things which are good about the system:

1) The system can index pages external to the website (one level out).
2) The system uses Lucene so, in theory, it's cross platform.
3) The system doesn't use Verity.

Business Process Breakdown

This section describes and explains the goals of major portions of the existing business.

Index Content

Content indexing is the process by which the system makes content searchable. The first step in the process is to identify content to index. The system does this by examining a queue of items and identifying one to index. In general this process favors content which has not been indexed or which has least recently been indexed.

Once content has been identified for indexing it is obtained. Textual content and links are extracted from the document. A summary is made from the extracted content based on the first sentences of most paragraphs. All of this information and the address to the document are passed to the indexer.

In the case the content can not be obtained the system will make a note that there were problems. If problems occur too many times for this particular address the address will be permanently removed from the Queue.

The system can interpret a wide range of documents identified by their mime types. However, there may be documents which the system can't interpret. These documents will be discarded and their address will be permanently removed from the queue.

When content is being indexed for the first time a new record will be created in the index. If the content already exists in the index it will be updated. If content already exists in the index but is, for whatever reason, permanently removed from the queue, the content will be deleted from the index.

Search Content

Uses will be able to provide criteria to the system for searching. Based on the criteria, the system will search the index and return a set of data. The data will indicate the address of the document, the summary which was created during the index, the preference or ranking given to the content as well as indicating if the content is internal or external to the current site.

System Users and Roles

This section describes the various system users and roles. These might be a person, system, device or something or someone else who uses the system.

Search User

Type of user: Person or System

Description of the the user or role.

A Search User is a person who uses the search API. The user probably wants to find some specific information.

It is conceivable that a Search User might also be another system. The API could potentially be exposed as a Web Service.

The responsibilities of the user or role.

The Search User is responsible for providing search criteria.

Description of the existing rules, policies or procedures which determine how the user or role does their job.

Rules can be thought of as agreed upon ways of working which are, potentially, changeable.

The Search User must provide search criteria to receive search results. If the Search User does not provide criteria the system will return a message indicating that the Search User must provide search criteria.

Description of performance related concerns for this user or role

There is no defined limit to the number of Search Users who may be using the Search System at any given time. However, realistically, the system must support a maximum one search a second.

Search Users should not need to wait more than a few seconds before the search system displays results. Ideally this will be less than 1 second to retrieve results. Users can wait up to 5 seconds under peak load.

Resources on which this user or role is dependant

The Search User is dependant upon the Search System. Without the Search System the user (obviously) can't perform their searches.

Content Queue

Type of user: System

Description of the the user or role.

The Content Queue is the system which identifies the next content to index. The Content Queue is neither responsible for retrieving the content or indexing the content. All it does is identify which content to index next.

The responsibilities of the user or role.

The Content Queue is the system responsible for identifying the next piece of content to index. The Content Queue does not actually get the content, it simply has a list of content and decides what should be retrieved next.

This "users" role is different from other roles in that it does not retrieve or index the content.

Resources on which this user or role is dependant

The Content Queue is dependant on it's Queue data.

The Content Queue is also depenant on the Content To Index which provides links to to new content which are stored in the queue.

Content Retriever

Type of user: System

Description of the the user or role.

The Content Retriever is the system responsible for retrieving the content to index.

Description of the unchangeable constraints which affect how this user or role does their job.

Unchangeable constraints might be legal, contractual or regulatory in nature.

To be able to retrieve content, the Content Retriever must have access to the content it's retrieving.

Resources on which this user or role is dependant

The Content Retriever is dependant on the Content To Index. Without Content To Index the retriever doesn't have anything to retrieve. The Content Retriever will acquire the Content To Index.

To function correctly the Content Retriever must be told which Content To Index. This information comes from the Content Queue.

Content Inspector

Type of user: System

Description of the the user or role.

The Content Inspector inspects content. Its purpose is to make the document understandable to other systems.

Content Inspectors are a generic type of system role. The role may be broken down into smaller parts in the design phase.

The responsibilities of the user or role.

The Content Inspector will be responsible for the following:
- Extracting text from various content formats
- Extracting links from various content formats
- Summarizing content

Resources on which this user or role is dependant

The Content Inspector is dependant on the Content To Index.

Indexer

Type of user: System

Description of the the user or role.

The Indexer is responsible for indexing content to make it searchable.

The indexer will likely use Lucene. There is a possiblity that the system might also support Verity.

The responsibilities of the user or role.

The Indexer is responsible for creating the and populating the search index.

Description of the unchangeable constraints which affect how this user or role does their job.

Unchangeable constraints might be legal, contractual or regulatory in nature.

The Index can only understand plain text. It can not understand anything else.

Resources on which this user or role is dependant

The Indexer depends on Content To Index and the Index. The Content To Index must be provided so that it can be placed into the Index for searching.

System Resources and Data

This section describes resources, information, or data which the system will rely on.

Content To Index

Description of Resource

The Content to Index is content provided to the system to be indexed and made searchable.

Description of the unchangeable constraints on this resources.

Unchangeable constraints might be legal, contractual or regulatory in nature.

For the system to be useful there must be content to index for searching. This content can come from anywhere, meaning from websites from the file system or elsewhere. The indexer simply must receive data to function correctly.

Describe any existing rules, policies or procedures which affect this resource.

The content to be indexed will need to be in a format which the Content Inspector supports. To begin with, these will include:

- HTML
- PDF
- DOC
- PPT
- RTF

Description of performance related concerns for this resource

The Content To Index must be processed reasonably quickly and should take no more than a second or two.

The API will only receive one piece of content at a time so quantity is not directly a factor. However, the API may be called several times in a row to index multiple pieces of content. For this reason the system does need to be quick in processing documents.

Index

Description of Resource

The Index is a file or set of files populated by the Indexer which hold data indexed.

The Index will be stored and managed by Lucene. There may also be support for Verity.

Description of the unchangeable constraints on this resources.

Unchangeable constraints might be legal, contractual or regulatory in nature.

The index must exist to be populated. This does not exist by default. The system must create an index if one does not exist.

Describe any existing rules, policies or procedures which affect this resource.

The index must be able to be manipulated. The system must be able to delete the index.

The index should be able to be created anywhere. This will allow implementations of the API to specify where the index is stored.

Queue Data

Description of Resource

The Queue Data is the set of data used by the Content Queue to determine the next content to index.

The Queue Data is an XML document which holds URLs indexed and to be indexed as well as status information on the URLs such as times indexed, times attempted, falures, etc.

Functional Portions of System

This section describes functional portions of the system and their requirements.

Search Content

Description of Business Process Requirements For Functionality

This is a broad description of the requirements for the functionality from a business perspective.

The Search API must allow for searching indexed content and returning a set of matching data. This functionality is essential to the API's primary purpose.

When using the search functionality, users expect the system to provide results relevant to their search criteria.

In the case that there are problems and search results can not be returned, the system will throw an error. The implementing system is expected to catch the error and handle it.

The following are possible scenarios:

- The user provides search criteria. The system finds matches in the index. The matching data is returned. In the case that no matching data is found, an empty result set is returned.

- The user provides no search criteria. The system does not know what to find. An error is thrown stating that no search criteria were provided.

- The user provides search criteria. The index is not available for any reason. An error is thrown stating the reason the index is not available.

This functionality will only be used by the Search User. However, the Search User might be a person or system.

Description of the unchangeable constraints on this functionality.

Unchangeable constraints might be legal, contractual or regulatory in nature.

The system must also have an index to search. If the index is not available for whatever reason, the system will throw an error.

Description of the existing rules, policies or procedures which affect how this functionality is preformed.

For this functionality to work the user must provide search criteria. Without this the system does not know what to find and will throw an error.

Description of performance related concerns for this functionality

It is not expected that there will be any higher volume than one search per second. Up to that rate, the system should take no more than a second or two to return results to the user. The speed of the system is constrained by the underlying speed of the indexing engine, Lucene. (Possibly by Verity too.)

Index Content

Description of Business Process Requirements For Functionality

This is a broad description of the requirements for the functionality from a business perspective.

The system must be able to index content to make it available to be searched. This functionality is essential because without this functionality, and without it being used, there will be no content to be searched by Search Users.

This process will probably be mostly automated and initiated by another system or process such as a scheduled task. For information on the business process see the Index Content Business Process section.

The index system must keep track of what content has been indexed as well as the state of the content. In the case that content can not be indexed a variable number of times the content will be marked as unavailable, permanently removed and never retried.

The index process will be initiated primarily by a system. This could be a manual process.

Description of the existing rules, policies or procedures which affect how this functionality is preformed.

The system might not have an index on disk when indexing content. In this case the system must create the index.

The Index Content functionality by default will to the next content identified in the Content Queue. However, the Index Content functionality can also explicity be told what content to index.

Description of performance related concerns for this functionality

The indexing process is not as sensitive to time constraints as the Search Content functionality. However, this process should still take less than a second or two to complete per content indexed. Even though the API only indexes one piece of content at a time, the process can be executed multiple times in a row.

Create Index

Description of Business Process Requirements For Functionality

This is a broad description of the requirements for the functionality from a business perspective.

The system will not have an index by default. The system must allow for the creation of the index.

This process will primarily be called by the system in the case that the system does not have an index and it needs to create one.

In the case that the process fails an error will be thrown.

Delete Content From Index

Description of Business Process Requirements For Functionality

This is a broad description of the requirements for the functionality from a business perspective.

The Delete Content From Index functionality removes content from the index permanently. This is provided so that content which needs to be removed can be. This is important in the case where content needs to be removed for copyright or any other reasons.

This process will be called by an administrative user.