1

Introduction to Enterprise Search

With the explosion of digitally borne information in the workplace, Enterprise Search has become more critical than ever before. Gone are the days when you could remember the location of all the file shares, web sites, and SharePoint sites, where the information you needed was stored. Instead, sites with terabytes of data are normal now, rather than being the anomaly they were just a few years ago. Remembering where you stored something last year, or even last week, has become an exercise in searching for a needle in a haystack. Also, with the growth of Internet Search, companies have begun to question why they do not have as good a search engine inside the firewall as they do outside the firewall. Internal customers are demanding that you provide a robust, scalable infrastructure for them to search against and provide in return relevant and timely results. Not a short order in any way, but reading this book will help!

Why Enterprise Search

Some of you may be scratching your heads, wondering why there is a distinction between Enterprise Search and Internet Search. Aren't the problem sets and technologies the same between the two? Yes and no. Some of the algorithms and protocols are the same, but some are different. While some Internet technologies grew out of Enterprise Search products, the technologies are distinctly different for a number of reasons that we will discuss.

A Tale of Two Content Types

While there is some overlap between Internet content types and Enterprise Search content types, the majority of the corpuses remain distinct. The Internet is made up mostly of web files ranging from Hypertext Markup Language (HTML) to Extensible Markup Language (XML) with not as much Office document content, while the reverse is true for Enterprise Search, where the majority of content is usually Office documents. Of course, this all depends on the types of content you

crawl in the enterprise. With SharePoint and the myriad of content sources it supports, you can get a lot of Office documents, emails, calendars, contacts, people, lines-of-business, or other types of content. This content is very different from web content in that metadata is critical, and the content does not include a lot of linking like web content. Therefore, web-style algorithms will not crack the content effectively, and the results will be less relevant.

Security, Security, Security

As they say, when you are on the Internet, no one knows you are a dog. That is fine for Internet Search, where you do not expect security trimming of results, since the search engine is crawling, indexing, and searching publicly available information. Plus, security is enacted when you try to click on the results in the results list. If you have to log on to access the web site or the file, you will be prompted for your credentials. If you do not have the right credentials, you will be denied access. You will have found the content, but just cannot see it.

Think about the same situation using Enterprise Search. You do not want folks being able to find information they do not have permissions for — not in the search results, not when clicking through to the information, not ever. Imagine that you search your intranet, and you can find the Excel spreadsheet with all the Social Security numbers, salaries, and bonuses for everyone in your organization. The search engine displayed just the first 500 words in the results, but when you click through you get an access denied notification on the spreadsheet. The damage is done. You saw information that was confidential and that should never have been returned to you based on your permissions. This is why security is paramount with Enterprise Search. Information stored inside of a company is usually more restricted and secured than information on the Internet. That is why any good Enterprise Search product has the ability to security trim both at index time and at results time.

In addition, you want your Enterprise Search product to support authentication and authorization against a myriad of authentication and authorization systems, since many environments are heterogeneous. If only one authentication type is supported, the ability to trim based on line of business security, Windows security, or even Lightweight Directory Access Protocol (LDAP) security is compromised and makes the system less useful. It is not uncommon to see three or four authentication and authorization systems inside of companies, while with Internet sites you basically run into forms-based authentication as the primary authentication and authorization method.

We will be covering security, in depth, in this book since it is a broad, diverse, and important topic.

Algorithms to the Rescue

Relevance is the name of the game when it comes to Search, both Internet and Enterprise. To achieve relevance, search algorithms have to scan the corpus of information and rank the results. Different algorithms target different criteria, and then a master algorithm brings together the different rankings to form a single master ranking. The strength of all the different algorithms affects the relevance of the results, and the results affect the value of the search application. We'll be covering the algorithms used by SharePoint later in the book, but it is a good idea to introduce some of them here, since SharePoint uses a number of different algorithms to try to get you the best result.

The main algorithms that SharePoint uses are anchor text, property weighting, property length normalization, URL matching, title extraction, click distance, URL depth, language detection, and file type biasing. The following section quickly describes these algorithms.

- ***Anchor Text*** **—** The anchor text is the text or link label that describes a hyperlink. It's the words you click on for the hyperlink. SharePoint will use the anchor text to establish the rank, but if the query terms are not included in the content of the item that it links to, the content will not be included in the final result list.
- ***Property Weighting*** **—** Different properties are more important than others. The property weighting will affect the ranking of the different properties in the final results. You can affect the property weighting by using the search administrative interface or object model. Be careful when changing property weighting, however, since this can severely affect your search results.
- ***Property Length Normalization*** **—** Different properties have different lengths. If you have a long property with a lot of values, this could affect relevance, since you can get false positives without taking into account the length of the property. This algorithm can only be customized through the object model.
- ***URL Matching*** **—** SharePoint looks at the URL of an item to see if it matches the query terms.
- ***Title Extraction*** **—** For Microsoft Office files, SharePoint attempts to extract the title of items, especially if the documents use the default titles from Office such as Slide 1 or Document 1.
- ***Click Distance*** **—** The distance between an authoritative source and the target content determines the relevance of the item. SharePoint will lower the rank the further away an item is from an authoritative source.
- ***URL Depth*** **—** How far off the main hierarchy content is found affects the relevance ranking of that content. This algorithm calculates the depth and sets the appropriate ranking. Content closer to the root of a site is more relevant than content further down in the hierarchy.
- ***Language Detection*** **—** Most users want to find content in their own language, so SharePoint detects the language of the user and ranks content in that language higher than in other languages.
- ***File Type Biasing*** **—** Certain types of documents are usually more relevant than others. For example, PowerPoint and Word are more relevant than Excel or plain-text documents. SharePoint ranks certain types of documents higher than other types of documents based on their type.

We All Love the Web and HTTP

While different, Enterprise Search and Internet Search share the same love for web standards and protocols. This is where the sharing between the two workloads becomes apparent. There is a significant amount of internal content that is hosted in web standard formats such as HTML, and access to that content is accomplished through the Hypertext Transfer Protocol (HTTP). In addition, the ability to provide the search experience through web-based user interfaces is critical, but Enterprise Search also provides client support for the desktop user, who may want a richer, more interactive experience.

Finally, programming against both types of search engines is similar in that web standards and protocols are used. Many times, both search engines can be customized by just tweaking the URL that you send to the engine and passing different parameters as part of the URL string. In addition, web services and

standards, such as OpenSearch, are critical components for the developer for writing code or customizing the search queries and experience. Sharing these technologies across both environments makes integrating the two technologies easier, which is important since we are seeing more and more convergence between the Internet and intranet experiences.

Conclusion

In this chapter, we took a quick look at Enterprise Search and some of the differences between Enterprise Search and Internet Search. While there are some similarities, there are some glaring differences between the two technologies. Content types, algorithms, and security are the main differentiators. As you read through the rest of the book, keep in mind that these differentiators are critical to your Enterprise Search deployment, relevancy, and development.