

Northern Light® Enterprise Search Engine

Overview White Paper

August 17, 2003



One Broadway, 14th Floor, Cambridge MA 02129

617-242-5960

Copyright 2003, Northern Light Group, LLC, All Rights Reserved

Table of Contents

Overview	3
Performance	3
Number of Documents and Average Size of Documents	4
Hardware Configuration	5
Competition For the Query Server	6
Complexity of the Incoming Queries	6
Scalability	7
Relevance Ranking	8
Automatic Classification	9
Classification of Content to Multiple Domains	9
Northern Light's Taxonomies	9
Custom and Third Party Taxonomies	11
Search and Filter On Any Metadata	12
Query Syntax and Parsing	12
Patent On Clustering Results	13
Security	13
Open API	14
Content Integration	14
Discovery-based Crawler	15
Administration Tools	15
Sparky	15

Overview

The corporate world is being rocked by seismic shifts in the volume and diversity of information available to professionals and Web-site visitors. Some of the trends include research being pushed out of the corporate library to end-user desktops, a demanding search-literate workforce and customer base, proliferation of information sources (internal, licensed, the Web), divisional and geographic intranets being hooked together to form gigantic worldwide networks, and the desire to link all corporate file servers to the network to make their contents accessible. And just to add to the challenge, IT budgets are being slashed while the demands on information systems are skyrocketing.

Enterprise search engines based on pre-Web search engine technology often do not have the performance, relevance ranking, classification and scalability required to make this colossal and diverse content universe accessible to end users. By contrast, the Northern Light Enterprise Search Engine is based on the award-winning Northern Light Web search engine that powered millions of daily queries from millions of end users searching a database of hundreds of millions of documents as vast and varied as the Web. You can bring the power of a Web search engine to your company's search applications.

This white paper provides a high-level overview of the features and architecture of the Northern Light Enterprise Search Engine. More detailed information is available from the company's technical staff. Contact us with your requirements or to arrange a demo or a trial.

Performance

Because the Northern Light Enterprise Search Engine is based on the Northern Light Web search engine, it was built to scale to very large databases and very high query volumes. Below are some estimated query capacities for a single installation of the Northern Light Enterprise Search Engine under different document counts searching a database of journal articles using a single Sun Fire 480 Solaris enterprise class server. (Note, multiple database installations can raise the effective performance above these levels).

1. 13,000 document database: 1190 queries per second
2. 5 million document database: 120 queries per second

3. 25 million document database: 43 queries per second

How many queries per second do you need? There is a universal tendency by organizations that do not have an experience base to analyze to greatly overestimate the requirement for simultaneous users and queries. One rule of thumb from the Web search engine industry is that each regular user averages one query per day. Because of peaking, the peak hour might be 12.5% of daily volume. So if you have 50,000 end users, you might get 50,000 queries in a day, with 6,250 of those in the peak hour, or 1.7 queries per second during the peak hour.

Query performance is a function of many factors:

- Number of documents in the database and the average size of the documents in the database
- Hardware configuration of the query server
- Competition for processing cycles on the query server
- Complexity of the incoming queries

Number of Documents and Average Size of Documents

On every query, the search engine has to scan the index to locate documents with the query terms, retrieve those documents identifiers and summaries, and rank order the list of documents in relevance order. And in the case of Northern Light, the list is further examined to identify relevant subject themes and to organize the results into Custom Search Folders™.

Web pages average around 5K per document, news stories are much less than this, journal articles are longer, and equity analyst's reports can be much larger. Corporate content varies all over the map. As a rule of thumb, the index is often about 60% of the size of the content that is indexed. So if a document database had 1 million Web documents, it would have around 5 gigabytes of content and the index would be 3 gigabytes.

The more documents in the database, and the larger the documents in the database, the lower the capacity of a given software and hardware configuration for searching that content. It simply requires more processing cycles to scan an index with more entries than an index with fewer entries. Longer results lists require longer to sort into relevance order. Lastly, the document identifiers (e.g., URL's) and summaries have to be retrieved by the database and longer results lists take longer for the identifiers and summaries to be retrieved.

The average size of a document is less important than the number of documents, but it can have an effect. In the Northern Light database index, there is a representation for each word and each word pair. So far example, if a document consisted of this text:

Barns dot the countryside. There are many red barns and a few that are not painted.
Red barns are the traditional color in this part of the world, but there is no evidence that the cows themselves prefer red barns to other-colored barns.

When the search engine indexes the above document, it will create an entry in the index for this document and the words barn(s), dot, the, countryside, there, are, many, red, and, a, few, that, are, not, painted, traditional, color, in, this, part, of, world, but, is, no, evidence, cow(s), themselves, prefer, to, other, colored, "barn(s) dot," "dot the," "the countryside," "countryside there," "there are," "are many," "red barn(s)," etc.. As a document gets longer, it will most often repeat words and word pairs, and the index need not expand every time this occurs. That said, a longer document do tend to extend the word list and the word-pair list somewhat, so it does affect query volume performance somewhat as well.

Hardware Configuration

Like all database operations, the speed and configuration of the hardware it is running on affects query performance. Some items to keep in mind are:

1. Number of processors. Modern enterprise class servers have multiple CPU's and each CPU can be executing a query during any given time slot. The effect is not linear however. Northern Light's testing indicates that two CPU's are 80% faster than one, but four CPU's are only 25% faster than two. The reason for this is that there is contention for memory and disk I/O and sometimes a CPU is waiting for another CPU to release a resource before it can continue. For a very demanding application, four CPU's may be worth the investment, but for most applications, two provide high performance.
1. Memory. Modern enterprise class servers can have up to 32 gigabytes of memory. The more memory, the larger the cache of already served queries that can be maintained. The search engine first checks the cache to see if the query can be served from the cache, and when it can be, the query is almost "free" from a processing time viewpoint. In any large application, many queries are repeated often, so large caches can greatly aide performance.
2. Processor speed. Modern enterprise class servers run at speeds over 1 gigahertz, and there is a near linear relationship between processor speed and query performance.

Northern Light maintains a customer-testing lab and will run a client's application on various machine configurations to arrive at a hardware configuration that will be optimal for a particular need.

Competition For the Query Server

For most applications in most organizations, it is perfectly acceptable (and somewhat simpler) to have the crawler and indexing processes running on the machine that functions as the query server. The query server module of the software will always have priority, and the capacity of a single installation so greatly exceeds most applications requirements, that there is no perceivable loss of query performance from operating all processes on one machine. However, in a very demanding application in terms of the queries per second requirement, these functions might be best separated out to other machines so that there is no competition for the CPU's, memory, and I/O of the query server. For these demanding applications, there is almost always a desire to have a powered-up, on-line, content-loaded backup machine ready to kick in to serve queries if the primary query server burns out a board or a power supply. It is often convenient to set up crawling and loading on the backup machine and simply suspend those operations in the event the primary query machine goes offline.

Complexity of the Incoming Queries

Consider three queries:

1. barns
2. red barns
3. red barns and (New England and not (Vermont or Maine))

Each of these queries requires a different number of processing steps to respond to. For the first query, all the query server has to do is find, retrieve, and relevance rank all the documents that have barn or barns in them. In the second case, the query server must cross two lists, documents with red and documents with barn or barns, determine the overlap between the lists, and then retrieve and relevance rank the overlap. In the third case, the query server must do everything it does in the second case, but before proceeding to retrieval and ranking, it must first look for New England, passing only documents that have New England in them, and then look for Vermont or Maine and drop those from the list. This last query takes many times more processor cycles to execute than the first one.

The mix of queries in terms of complexity can affect query performance. Rules of thumb are that young people, consumers, and staff that are not knowledge-worker professionals tend to use simple one or two word queries, while knowledge-worker professionals, information technology professionals, and, especially, librarians really work a query parser. Testing a software and hardware configuration with your content set and a file of your actual queries can address this issue, or more simply, look at the query logs for your existing solutions and see how many complex queries are received as a percentage of the total. (Assuming of course, your existing search engine *can parse complex queries at all.*)

Scalability

The trend toward worldwide corporate intranets with scores and even hundreds of file servers hooked to them has geometrically increased the amount of content that is available to search on a corporate network. Some intranets commonly having millions of documents directly on the intranet, or on the file servers hooked to the network. The issue of scale of the search solution is increasingly important

Many enterprise search engines architected on pre-Web search engine technology never expected to search really large databases with millions of documents or tens of millions of documents. These enterprise search engines, including many industry leaders, struggle scaling above 1 or 2 million documents. Some actually shut down as the document count rises, failing to return a result at all. Others progressively slow down until they are not productive tools for end-users.

A common method of scaling beyond the design capacity of a search engine is to divide the database into smaller parts and search the parts separately, consolidating the results before returning the results list to the end-user. While there is some size beyond which all search engines must pursue this solution, an important consideration is how quickly this threshold is met. For example, one industry-leading maker of enterprise search engines only has the capacity to search 150,000 documents before a second hardware appliance must be added. The highest cost element in most IT departments is server administration, and search engines that require new server hardware for modest increments in database size have important hidden cost elements and operational headaches. .

By contrast, the Northern Light Enterprise Search Engine was derived from the Northern Light Web Search Engine that served millions of end-users daily with millions of queries from a database with hundreds of millions of documents in it. The Northern Light Enterprise Search

Engine can search databases of up to 25 million documents with a single software installation on a single server.

Relevance Ranking

As document databases grow in scale, the single most important factor in making them useful to end-users is the effectiveness of the search solution in relevance ranking. There are material differences between vendors in this crucial dimension.

Northern Light has a unique seventeen-factor approach to relevance ranking that considers statistical text measures, link-popularity analysis, subject classification, and date – and balances all these dynamically to weight the factors based on what will be most useful for a given query.

What, you ask, are statistical text measures? Well, a few examples would be the number of times the query terms are in the document relative to the length of the document, the proximity of the query terms in the document, the word order of the query terms in the document, the presence of the query terms in the document metadata, and the inverse term frequency of the query terms in the database as a whole.

This last point is an important one. When a query contains both common and uncommon terms in the database as a whole, it is much preferred to rank documents higher that are comparatively richer in the more rare terms than to rank documents higher that are richer in the terms that are relatively common. Some leading enterprise search engine vendors do not incorporate this factor and as a result their relevance ranking effectiveness declines rapidly as the database size grows.

Also, hyperlink analysis has received much publicity in the Web search engine community and there is no doubt that considering the “link popularity” of a document on the Web makes an important contribution to relevance ranking it effectively. Indeed, Northern Light considers link popularity when there is good link popularity data in the content set. However, in corporate applications or applications in which there are large bodies of published content (e.g., news or journal articles), there is virtually no useful link popularity data to consider, and vendors that rely too heavily on link popularity will not be as effective in these settings.

Subject classification is one area in which Northern Light’s approach to relevance ranking is completely unique. Northern Light was designed from the ground up to use classification meta data at query time to filter, organize, and relevance rank results. Out of the box, all content indexed by Northern Light is subject classified. Northern Light is able to compare the query terms to the node-descriptors of each node in the taxonomy, and give boosts to documents that are classified to nodes that contain the query terms in the node-descriptor.

An independent academic researcher has studied Web search engines and concluded that Northern Light's relevance ranking effectiveness exceeds that of all of our competitors including the ones that also have enterprise search engines.

Automatic Classification

As document databases grow to tens of thousands or documents to millions of documents, the ability of the search solution to organize the results for end-users is vitally important to making the database useful for end-users.

Classification of Content to Multiple Domains

Northern Light has patented, proprietary technology that classifies every document against four distinct domains: subject, type, source, and language. While the majority of the classification is done automatically via artificial intelligence algorithms, human editors and librarians train, guide, and continually improve the classification process. We also use metadata and structural elements when appropriate to augment or customize classification.

Classification serves several purposes in the Northern Light Enterprise Search Engine. It allows for multiple points of access for finding a document, as one document can be classified to each of the four domains. For example, a document may be about Agriculture (subject), in the format of a statistical report (type), from the journal Agricultural History (source), written in English (language).

Classification is also one of the 17 factors used in relevancy ranking of results. It also powers unique navigation aids like custom Search Folders™, and allows Northern Light to create vertical search solutions with ease.

We provide a ~17,000 node subject taxonomy that is extensible and customizable. In addition, Northern Light can create custom classification schemes for customers and partners, as well as classify Web, third party, and customer's proprietary content to multiple simultaneous classification schemes.

Northern Light's Taxonomies

Northern Light's multiple taxonomies power our Custom Search Folders™ that subcategorize results on the fly for each search. Each domain draws upon a taxonomy consisting of a set of discrete values from which folders are selected.



These taxonomies are hierarchical, and may also contain cross-references, i.e., a given value may have more than one parent (strictly speaking, each taxonomy is an acyclic directed graph).

- The **subject** taxonomy contains approximately 17,000 terms (often referred to as “nodes”). Subjects answer the question, “what is this about?” There are 16 top-level subjects covering broad categories such as Business & Investing, Health & Medicine, and Humanities, and they are highlighted on the Northern Light Power Search form. The hierarchical structures beneath these top levels can be up to 9 levels deep in some areas to provide specific subject classification and description for areas such as Melanoma or Robotics.
- The **type** taxonomy is also unique to Northern Light. Types answer the question, “what is this?” Some examples of possible genres from the type hierarchy include Articles, Reviews, Editorials, Market research reports, etc.
- **Source** refers to the origin of a document. The hierarchy is bifurcated at the highest level into Internet or Special Collection sources. Internet is further broken down by domain name (e.g., Commercial sites, Education sites, Government sites, etc.). For Special Collection documents, source is the title of the work from which an article came. The Source taxonomy is also used in the corporate setting to define sets or folders of documents so they, paired with information about a user’s identity and network permissions, can be used to enforce access control policies.
- **Language** specifies the predominant language of the document – Northern Light currently identifies English, French, German, Spanish, and Italian documents.

Northern Light uses a mixed approach to process and classify documents to our multiple taxonomic structures. Documents that do not have reliable metadata can be loaded with the automatic classifier assigning subject and other classification metadata to the document at indexing time. This “out of the box” solution works extraordinarily well on content sets that are of the type represented by Web pages, news articles, and our journal articles. The classifier works by comparing the documents being classified to the “training documents” that our expert gang of librarians identified as best representing that node in the taxonomy. While we mere humans can imagine this process at the level of one document and two nodes (“Is this document more about barns or about cows?”), Northern Light’s classifier, in effect, compares 17,000 subject nodes simultaneously to a document to pick the best fit, and it does this hundreds of thousands of times a day if a large database is being classified.

Out In the corporate setting we can take advantage of your tagging or document structure conventions to augment or replace our automatic classification. Filters can be easily written using Northern Light's open filter architecture to capture whatever metadata is of interest. For example, "sales call reports," "telecom marketing studies," "hiring procedures," or "cancer drug trials" might be reliably present as document tags and be of interest in particular corporate settings.

Custom and Third Party Taxonomies

In the context of a project, Northern Light will evaluate parts of the taxonomic structures of particular interest to the client, and if needed, Northern Light will work with the client to support new or specialized concepts within one or more of the taxonomies. If the concept does not exist in the NL taxonomy, it is added to the appropriate section of the hierarchy, necessary cross-reference relationships are established, and the training of the document for the auto-classifier is undertaken.

Years of experience with automatic classification have taught us the importance of a rigorous and thorough evaluation of any term that is proposed to be added. We seek to provide clear distinctions between subjects, and constantly evaluate balancing the level of granularity we can support through automatic methods. Two aspects we constantly keep in mind during this process are:

Fragmentation: Fragmentation occurs when two or more very similar concepts exist in a structure. These "near-related" concepts prevent the auto-classifier from distinguishing between subjects to accurately classify a document. It is important to consider whether the term is a distinct subject, with its own associated terms and concepts, so that it will classify cleanly and not draw documents from a near-related concept.

Inheritance: Items classifying to a child node in the structure inherit up to the parent concept. Therefore, the scope of the child must be one aspect of the scope of the parent and cannot be broader than the parent. Moreover, sibling terms should have the same degree of specificity or aspect of the parent's topic.

When new nodes are desirable in the taxonomy, one of the key tasks is to identify a set of ten or more training documents for that new node. The training documents are those that in the judgment of a practitioner or librarian best represent that node in terms of really "being about" that subject and also correctly representing the "breadth" of that node. The training documents are submitted to the automatic classifier and it extracts what is most different about those training documents from the training documents in each of the 17,000 nodes of the taxonomy and, in

effect, learns how to distinguish documents about this subject in the future. If this sounds eerie and magical, well, it is.

This is also the general procedure for building or using a completely custom taxonomy with Northern Light. Design the taxonomy, develop a set of training documents, submit the training documents to the automatic classifier, stand back, and let it rip.

Interfacing to third-party taxonomies is also very easy. All that is required is to write a filter to capture and metadata the third party taxonomy assigns to the document and Northern Light can use that metadata like any other.

Search and Filter On Any Metadata

All metadata is represented in the Northern Light index, which means you can use search forms or syntax to qualify the results. Search on title, sources, documents types, etc. You can add any metadata that makes sense to your organization and search on that tag. Custom search forms can expose the metadata to end-users to prompt use of metadata to qualify results and improve relevance.

Query Syntax and Parsing

Query parsing is one of the most fertile areas for improving the user experience. A good search solution should provide effective results when being used for the first time by an untrained end-user, provide spectacular results for the trained end-user or professional information worker, and should send an organization's librarian into gleeful orbit. Northern Light worked really hard on building a query parser that accommodates this broad range of expectations and abilities while providing the very best search result possible in all cases. There are several aspects to delivering the best solutions.

Northern Light allows keywords, Boolean expressions (all operators, compound, and nested), natural language, phrase searching, wildcards, and any combination of these. So, for example, queries of the following forms might represent two extremes of training and professional involvement:

Alzheimer's drugs

drugs or pharmaceuticals and ((treatment or regimen) and alz*) and "clinical trials"

While Northern Light will do a “gleeful orbit” job on the second query (because the searcher in this case has fully specified the desired results), frankly we will do a very good job on the first one despite its generality. Keys to being effective at simple keyword queries include

Relevance ranking. Not wanting to repeat the long discussion from a previous section of the white paper on relevance ranking, suffice it to say that many if not most two and three word keyword searches have enough meat in them to get very good documents to the top of the results list if the relevance ranking algorithms are up to the task.

Default to ‘and’ not ‘or’. In large content sets, getting enough results is never a problem. Rather, the problem is getting on-point helpful results. Northern Light’s Web search engine executed billions of end-user queries over its five-year life, and we studied what people search for and consider effective in terms of results. One of the key observations we made based on that experience was that virtually all users prefer 200 highly useful hits to 400 highly and somewhat useful hits spread out through 2,000 less useful ones. Northern Light defaults to ‘and’ in query parsing, and this greatly aides precision in the results list.

Use of Custom Search Folders to disambiguate results. Northern Light presents the top dozen or so subjects represented in the results list most relevant 400 documents to the end-user as a drill-down opportunity. By selecting a Custom Search Folder of interest, the end-user gives us more data to work with on what is desired, and the results in that folder will be more relevant. Our Custom Search Folders make use of our patent, as described below.

Patent On Clustering Results

Northern Light has a patent (Method and Apparatus For Researching a Database of Records. Filing data: 5/1/1997, Patent number 5,924,090, Grant date 7/13/1999) on, among other things the use of metadata post-query to organize results. This means no vendor other than Northern Light can (legally at least) provide you with a search solution that organizes a results list by subject, source, document type, or any other type of metadata.

Security

Northern Light integrates with your authentication practices to insure that end-users only see documents on results lists that they are entitled to. We provide tools that you can use to:

- Define criteria that will be used to restrict results that are returned to a user (security_id.)
- Determine basis for identifying classes or subsets of documents to which security evaluation can be applied (source level security)
- Associate these document groups with security criteria (security groups specifications file)
- Associate security group tags with actual documents in the database (security index)
- Manage these associations (security groups administration interface)
- Means for defining a database as either secure or insecure
- Create a user hierarchy of security group tags (or download an LDAP directory) and assign or delete individuals from security groups

The Northern Light search engine will then authenticate each query by each end-user against that individual's privileges and will only show results that he has permission to see.

Open API

Northern Light has well-documented API's using J2EE standards, XML search results, and JSP sample code that support the integration of Northern Light into corporate applications. It is a relatively simple matter to hook Northern Light to your UI or application, take queries from that UI or application, and return results back to your UI or application. We explicitly designed our enterprise search engine with this functionality in mind in order to make it straightforward for developers to work with.

Content Integration

The data conversion system includes filters for Microsoft Office, PDF, HTML (including JSP and ASP), and text formats including XML. Our published load format specification allows any file type, from any source, located anywhere to be indexed and searched. It is relatively simple to create a customer filter for a proprietary content type or set of metadata. For example, you may have a file of MS Word documents that are customer proposals, and you may want to capture metadata at indexing time about the customer name, business unit, or product line it applies to.

Discovery-based Crawler

Northern Light's crawler follows links on your network to discover content for indexing. The crawler connects via HTTP, HTTPS, FTP, NFS and SMB (Windows) protocols and supports multiple authentication methods. It can be used for Internet, Intranet, or corporate content discovery and crawling.

Administration Tools

We provide a browser-based administration system that includes a basic search UI, a scheduler to manage crawling, data conversion, database loading, and a system configuration manager.

Sparky

