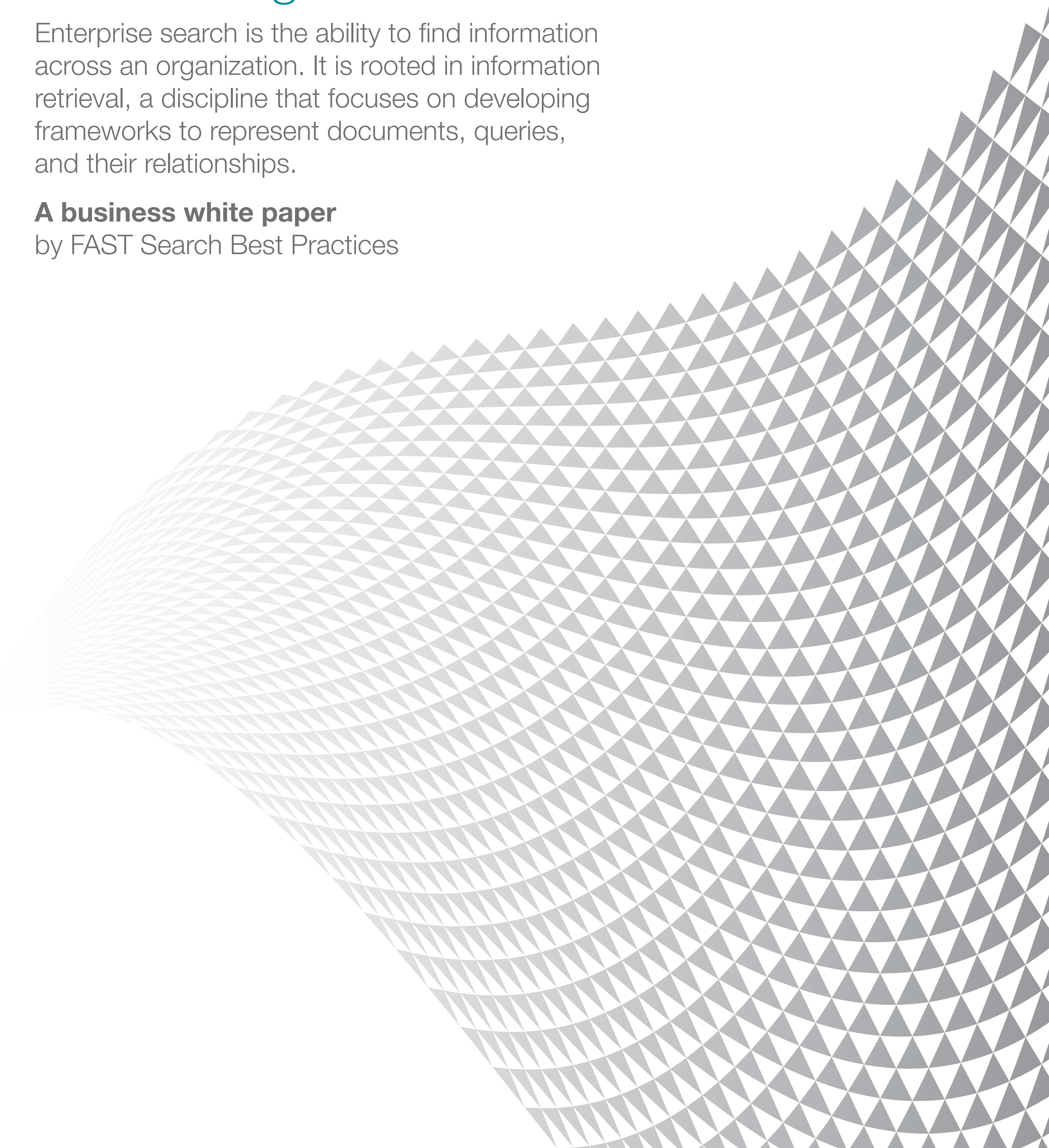
fast SBP™

Introducing search

Enterprise search is the ability to find information across an organization. It is rooted in information retrieval, a discipline that focuses on developing frameworks to represent documents, queries, and their relationships.

A business white paper

by FAST Search Best Practices



What is search?

Traditionally, search is the action that links users to the information they are seeking by means of a query:

- What's on the lunch menu today?
- Which mp3 players have the best reviews?
- How many barrels of crude oil did our refineries produce last year?
- xml:sentence:("D-Day" and scope(date)) and scopenavigator(context=date@base!)

Using a certain syntax searchers explain what they are looking for. On the other side of the fence, search providers design systems which will optimally match information to people.

Today state of the art search technology is used to enable a wide range of applications, some of which do not contain a standard input box: real-time alerting, market and trend analysis tools, data clustering, advanced mobile services, or automatic video and audio broadcast monitoring to name but a few. It is also an essential tool for performing advanced data mining. Search provides the platform for sophistcated real-time analytics of very large data sets. By slicing, dicing and intelligently augmenting structured and unstructured information, highly contextually aware applications can enable everyone from simple web users, to demanding and skilled knowledge workers to maximize the use of the information available to them.

These technologies are all rooted in the science of information retrieval, a discipline that focuses on developing frameworks to understand and represent documents, queries, people, and their relationships. To enable a better understanding of the above, this book takes the bold task of providing an overview of many of the elements of the science of search, why they are important and how they can be uniquely applied to many different business scenarios and industry sectors.

The ubiquity of the technology means that most people will have an idea in their minds of what search is. Words such as "search engine", "crawler" or "inverted index" may be used. Yet whilst at first glance it may appear a trivial and commoditized application, search is actually a flexible and feature rich platform.

From a technical perspective, this starts with mechanisms extracting structured and unstructured content from data repositories, regardless of format or location – for example, messaging systems, television channels, databases, VoIP (voice-over-IP) telephony systems and information archives. The system then processes that data, analyzing and augmenting it as appropriate, before building an intelligent index of the information. The system never moves or alters the original data – it simply stores a reference back to it.

Among other things, the tuning of the index allows search providers to define a ranking model; this is the method of mathematically scoring how well each document matches a query. Once the information is in the index, multiple applications (an intranet portal, a native Java client installed on a smart phone, a personal search tool installed on the desktop) use the query language of the search technology to execute operations and retrieve results.

To optimally match and rank a set of documents related to a query, best in class search products enrich the original content. Enriching techniques include adding metadata to the original document, linguistic analysis, identifying key entities, keywords and concepts, and encoding geographical information.

Along with a set of ranked results, some enterprise search systems return analytics around the query and results set, such as suggested spelling corrections, related concepts and statistical information about metadata. This enables interaction paradigms including drill-downs, refinement, and concept linking, allowing users to refine searches and browse results.

To be fully effective, enterprise search systems must successfully combine technologies from multiple, complex disciplines – for example, computer networking, text parsing and analysis, linguistics, information theory, relational mathematics, statistics, computer software, computer hardware, relational technologies, algorithm efficiency, and human cognitive psychology.

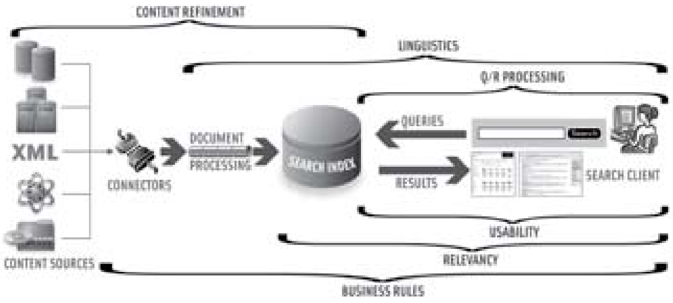
It's important to consider business and technical requirements when deploying enterprise search applications. From the business angle, organizations will leverage the inbuilt features of the search system to promote a better search experience and improve competitive positioning, which will help grow revenue/reduce costs, increase productivity and reduce risk. Features for the business-minded include relevancy, linguistics, query and result processing, content refinement, business rules and the usability of the system itself.

To ensure a best-in-class search application that supports the business environment the technically orientated areas of integration, security, performance, high availability and benchmarking should be considered.

To better understand the dynamics and make-up of search applications, we offer an overview of each of the topics in the pages that follow, describing and analyzing the area, highlighting questions that need to be asked and the best practices around different market segments.

Considering the business aspects of search

Business aspect of search



Relevancy

The search engine must produce accurate and relevant results to meet the expectations of increasingly demanding users and applications. This requires an ability to tune the search engine to meet end-user expectations and business needs.

However, enterprises use search in multiple contexts: commerce sites, corporate information sites, intranets, extranets, portals, etc. Each has distinct objectives, and user communities value content differently. So it makes sense to be able to adjust the yardstick to evaluate content in order to get results that align with the objectives of the search context and with users' needs.

Enterprise search solutions are evolving far beyond standard Web search capabilities. Ranking models are now based on multi-faceted quality measurements of the match between query and document. Relevancy is determined by concepts, and additional levels of abstraction such as context, freshness, completeness, authority, statistics, quality and geography. A ranking model is the independent tuning of each element relative to the business need, to determine whether a document is a good match to a query.

Organizations must understand the underlying information and the search user to achieve the goal of superior relevancy: balancing recall and precision. In general, customers need to strike a balance between finding everything related to a query and only the documents that relate to a given query.

More scientifically, recall is the ratio of the number of relevant records retrieved to the total number of relevant records in the index. Precision is the ratio of the number of relevant records retrieved to the total number of irrelevant and relevant records retrieved.

Linguistics

Linguistics deals with the structure and variation of languages (on the user's query and also the content) to improve the user's ability to find relevant information. This applies to languages as divergent as English and Mandarin Chinese, and also to industry-specific terminology.

The main index-side linguistic optimization tools available include automatic language detection, augmentation of the content via lemmatization, synonyms to correlate words in a much broader sense, and removal of stop words. For Asian languages such as Mandarin Chinese, Japanese, Korean, and Thai, tokenization algorithms (the operation of splitting up a string of characters into a set of tokens) must also be used. More advanced interpretation of language can be carried out using entity extraction, recognition of parts of speech, categorization, unsupervised clustering of documents, and sentiment analysis.

Certain linguistic features focus on improving queries – for example, spelling correction. Queries can also be improved with phrasing (the recognition and grouping of an idiom such as "home run") and anti-phrasing (identifying word sequences in queries that are irrelevant to the search).

A grasp of the linguistic features of advanced search engines, along with data and user knowledge of the search service, can greatly improve precision and recall, therefore improving user success in searching.

Query and result processing

Search engines are faced with two basic information retrieval issues: 1) how to help users craft better questions, and 2) how to provide better results, minimizing what the user has to read through. To deliver consistently superior results, one must understand the intent of the query, know what information is available, know how it inter-relates, and identify where it is located.

These goals can be accomplished by using two key technologies: Natural Language Processing (NLP) and linguistic analysis. NLP interprets queries posed as questions and phrases in part by identifying and stripping out irrelevant terms. Linguistic tools include capabilities such as avoiding word-sense ambiguity to distinguish between, for example, the color orange, the company Orange and the citrus fruit. Search applications house these technologies in the query/result processing stage, where the goal is to analyze human language (the query) to identify the searcher's context and intent, and to return the most relevant set of results (result processing).

Content refinement

The quality of the search index and search experience is dependent on the search engine content. It's important for organizations to consider the quality of the content and prepare it appropriately before it hits the search index. The content refinement lifecycle includes two stages: content aggregation and processing.

Content aggregation brings together content from multiple sources. It is also used to amalgamate search results. Content is made available to the search engine via the content API that acts as an information broker. It pulls content from the data source (database, CMS application, etc.) during scheduled calling requests and pushes content into the search engine.

Document processing is the analysis, conversion, transformation, and enrichment of original content for the purposes of indexing and subsequent retrieval. It can be made up of one or more document processing stages (for instance, language detection, synonyms, spell checking, lemmatization, taxonomy classification, and custom plug-ins). These stages analyze the content and add or remove or transform data accordingly.

Business rules

Organizations are governed by business rules and workflow. The aim is to adjust the business rules in line with market trends, analytic regression patterns, etc to meet the needs of the business and the market. But how do business rules apply to search technologies?

Business rules are algorithms, workflows, or heuristics that are implemented and supported by a software system. For example, a business rule might be that a credit check is not necessary for returning customers. The search application should allow customers to apply business and processing rules at various stages of the search, such as ingestion, ranking, query transformation, or at alerting time (running preconfigured queries and pushing results to the user), to provide the best-in-class search solution.

Analyzing the impact of business rules enables search providers to impact the underlying relevancy model, and direct end users to business-generating landing pages. In the case of a knowledge discovery scenario – at a pharmaceuticals manufacturer, say – analysis of query logs will improve search for users in functions such as R&D, clinical trials, and sales and marketing. The analysis may highlight the need for custom dictionaries (for specific industry terminology) or linguistic capabilities to reduce the number of queries or zero hits. In this scenario, business rules are effectively being used for "fault detection".

Usability

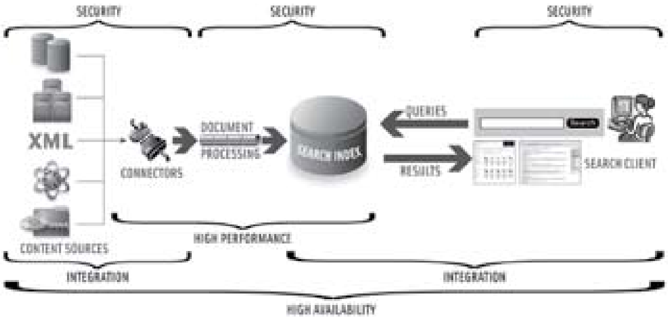
Usability is a key component of a system's total delivered value. User interaction must be aligned with the core purpose of the system itself, and navigation of the system should be obvious to the user. In short, good search usability is an essential part of realizing the system's full potential. It describes how the user is guided through the system from start to finish. When establishing or revising a search-powered system, usability is ultimately judged by the users (business, consumer, administrators, etc.) themselves, so search providers should follow a user-oriented design process. Leveraging reporting and monitoring tools, such as click-through trends, page impressions and abandonment points will provide quantifiable metrics on the success of the system.

Good usability is provided by defining the search experience, aligning the system design with the definition and then letting real end users test and evaluate it.

The site owner obviously wants to make visitors' tasks effective. Search enables workers to quickly and easily locate content relevant to their tasks, and it may assist with the basic analysis of large amounts of information. The user search experience can also impact both motivation and productivity. So it's vital to make search easy to find, to provide different ways to search, and to match advanced search capabilities to users' needs and abilities.

Considering the technical aspects of search

Technical aspect of search



Security

Security guidelines apply in three areas. First, in managing end users – verifying their identity (via the portal application that search is embedded) and the content they're entitled to access. Second, from an application perspective, validating that all communications are issued by authorized clients and that connectors respect a repository's access model. Third, when executing a query, search software must always confirm permission levels on documents.

The most common corporate practice to ensure privacy is using folder- and document-level access control within applications and repositories. This access control logic must then be respected by other applications connecting to the content, including search engines. In this way the search engine becomes the gateway and the gatekeeper to valuable and confidential corporate data.

Integration

Integration deals with embedding search engines in third-party software applications.

There are two areas where this is typical: in an authoring or management application such as a document management system (DMS), where stored content has to be searchable; and an industry-specific workflow and investigation tool (for example, for a law firm or for compliance in the financial-services sector) where multiple external data sources are searched and processed within the application in question. OEM integration requires substantial planning to provide a seamless assimilation of the two technologies. The component architecture has to be understood so that each connection point is identified and treated separately.

There are five main areas to consider: content creation for indexing, index configuration, query logic, user interface design, and administration and configuration. The most important part of integration is deciding which configuration or query features to expose to the content consumers and managers. Top-quality enterprise search engines are flexible tools, designed to cope with many different types of applications and industries – e-commerce, knowledge management, archiving, video search, and more. So for each design and configuration decision made, the OEM must decide whether the decision applies to all clients or whether the option must be left open for systems integrators or IT administrators to fine-tune the system.

Performance

Providing a scalable and high- performance search offering calls for selecting the correct software and hardware configuration based upon the most important performance and fault tolerance requirements.

Identifying the key metrics and the appropriate hardware and software will help with designing a high-performance search service. The performance metrics include the total number of documents to be indexed, the required ingestion rate and acceptable indexing latency, and the number of queries per second (QPS). Similar to the recent trend of grid computing and distributed architectures to support scalable enterprise software, high-performance search grids use replication and distribution of servers to scale document volume and QPS. A third dimension is document ingestion rate, which scales with the resources allocated to content aggregation and processing. Systems using these three dimensions should be able to scale linearly, independently, and simultaneously to achieve the desired performance targets.

Search providers across different business segments have very different needs. News and financial search require fresh data, while litigation support services, for instance, require batched data, indexed once. The key to optimizing a system lies in understanding the user's objectives, while sizing and designing to strike the right balance between speed, size, and cost.

High availability

Critical IT systems are often described as fault-tolerant, redundant, or displaying high availability. In other words, should something go wrong – power cut-offs, hardware failure, corruption of data, say - the systems have been designed to maintain certain levels of service.

The standard solution is to purchase more hardware to mirror vital elements of the system. The downside there is the price of the equipment and the internal cost of managing a duplicate set-up. The extra expense of redundancy should be compared to loss of revenue from the downtime and the odds of fatal errors occurring in the system.

Failures that can impair the search application fall into two areas: backbone failures, and component and service failures. Backbone failures include various hardware incidents and network outages. More often than not there will already be a corporate- or service-wide policy regarding power supply and data-center security, and search will typically comply with these policies. But in cases where uptime is critical, hardware redundancy should be combined with intelligent recovery operations.

Benchmarking

Reporting and benchmarking creates a structured method for measuring and validating the success of search with respect to these varied stakeholders. Without adequate measurements, enhancements may not be possible. The most popular queries, broken links, and user surveys are often used to evaluate the quality of search.

This can be divided into two categories: searchable data and index information, and search engine usage metrics. The first category includes the number of documents in target repositories, an audit of those documents, and information on hardware usage. Usage measurements include hard numbers such as click-counting and subjective measurements about the quality of the interface and the results ranking.

These categories are further split up according to stakeholder groups and the reports relevant to each. Identifying which group has an interest in various metrics helps to assign the ownership of search components to different groups; who is responsible for identifying sources of content, whose role it is to determine what good and bad results are; who is responsible for building and maintaining the platform.

What is good search?

Developing an optimal enterprise-class search capability that delivers complete flexibility is not difficult – it just involves lots of choices that may or may not have to be considered.

There are different subsystems to take into account, and, in a world where one size does not fit all, there are many opportunities to customize each of those to best suit the needs of companies and users in diverse industries, with different requirements, non-uniform usage patterns, and varying goals. However, organizations are starting to realize that the effort is worthwhile; they are beginning to see that the enterprise search platform is every bit as important as the ERP or CRM applications that support their business operations. Getting the design right is achieved with careful planning and with a clear appreciation of the available tools and necessary trade-offs.

In particular the following twelve chapters provide an introduction to understanding the key areas that compose search, and start to underline the important best-practices to abide by when building a state of the art system.

An accurate knowledge and view of the targeted tasks and users of search, as well as the topics of this book, are primordial in selecting the correct balance of the necessary features to assist the users in performing their task. Hence are built solutions that satisfy both searchers and search providers, providing a great user experience and a significant return on investment.

About FAST SBP™ (Search Best Practices)

SBP consulting is a highly focused transfer of search knowledge and experience from FAST to its prospects and customers. SBP workshops aim to help enterprises realize the full potential of search, by creating optimal strategic, functional and technical roadmaps, delivered in the form of business model, solution and architecture designs.

For any feedback or questions related to this paper, please contact us at sbp@fastsearch.com.

Fast Search & Transfer
www.fastsearch.com
info@fastsearch.com

Regional Headquarters

The Americas
+1 781 304 2400

Europe, Middle East & Africa (EMEA)
+47 23 01 12 00

Japan
+81 3 5511 4343

Asia Pacific
+612 9929 7725