THE BUSINESS CASE FOR A FULL FEATURED SEARCH ENGINE IN DATABASE APPLICATIONS

WHITE PAPER

Table of contents

Introduction

Organisations of all sizes and within all industries generally distribute their corporate knowledge amid a variety of heterogeneous database applications. Examples include customer relationship systems, staff directories, content management systems (CMS), electronic document and records management systems (EDRMS), library catalogues and university course databases. In each of these cases, the effectiveness of the database application is often largely determined by how easily information can be retrieved by stakeholders.

IT managers who realise the importance of search will often ask the questions "Will the native database search capability be sufficient?" and "What benefits would be gained by using a third-party search engine instead?" The answers to these questions depend largely on the database technology and the third-party search technology. However, a good quality and full featured search engine offers undeniable benefits and a convincing business case.

This article draws from years of experience working with organisations that have been through this decision making process and lays out the framework for building a strong business case for an external search engine.

1. Result Ranking

The primary responsibility of a search engine is to display all relevant results in a ranked list that minimises the time to locate the most valuable results. In web search, the ranking is largely determined by factors such as anchor text, link graphs, URL length, recency, and term density. However, many of these factors don't exist in the realm of database search.

This lack of evidence often means that no clear definition of relevance exists. Is a database record more relevant because it contains the keyword twice instead of only once? Sometimes this may be the case, but there may be better options available to assist with the ranking. These include:

- Incorporating a recency factor into the ranking. Recency may be based on the database record's last modified date, or it could be the most recent time the associated library catalogue item was borrowed.

- Incorporating an external evidence factor into the ranking. External evidence could include the number of times a library catalogue item has been borrowed, the average user rating of an online product or the inverse geographic distance from the user to the physical object referenced in the database record.

- Incorporating end-user click feedback. End-users are often the best judge of which results are most valuable so why not incorporate their feedback? Click feedback works by recording the number of times a search result has been clicked in response to a query and uses that information to upweight the popular results.

- Up-weighting the importance of particular fields in the database (e.g. records where the query terms appear in the 'title' field are likely to be more valuable than those where they appear in the 'related items' field).

A good search engine is designed to make use of all available information, but also provides a tunable ranking algorithm to produce the optimal ranking for a given information set.

2. Performance and Scalability

A third-party search engine operating on a database application effectively acts as a cache, relieving the database of high query loads and associated performance requirements. Good quality search engines tend to be much more tolerant to query load because of the highly optimised index structures designed for fast lookup. Since the search engine will only access the database when the index requires refreshing, the database load becomes minimal and predictable, avoiding the highly variable loads otherwise experienced.

The performance requirements of the database are even more significant if the database is processing queries over a complex schema and/or performing many joins and aggregations. Costs associated with database optimisation, hardware and software can be reduced if the primary goal of the database is to store and manage information rather than having it accessible in real time.

3. Query Language

A high quality search engine will provide a flexible and powerful query language for posing complex queries. Complex queries can be submitted using either query operators or via an advanced search form. Examples include phrase queries, proximity queries (i.e. find matches where the query terms appear near each other in the record), Boolean operations, fielded search (e.g. search by title) and date range constraints. Other desirable features include the ability to specify optional stemming, case sensitivity and truncation (i.e. wild card) modes.

Although most databases support a subset of these features, a search engine will generally provide an interface that is more consistent with the global web search engines, making users feel more familiar with the interface.

4. Search Form and Results Page Customisation Tools

A full featured search engine will provide tools, APIs or templates for constructing search forms and results pages that integrate seamlessly into any web environment. Anything from a simple search text input field to a highly complex specialist metadata search form can be constructed in minimal time and maintained by in-house staff with no software development skills.

Accessibility compliance is an increasingly important requirement for many organisations, so a search engine that can support this capability out-of-the-box will also be an advantage.

5. Exposure to Web Services

A search engine that exposes an interface such as REST based, .Net or Java can provide a simple and convenient way to integrate the database application with other web services. This enables the database application to be accessible by other corporate or third-party repositories that would benefit from a tighter integration.

6. Federated Search

Enterprise search engines provide the ability to index content from a wide variety of sources including web sites, file shares, databases, corporate applications, email and other data sources. The most powerful search engines will also provide a federated search capability that enables a single query to search all repositories. The major advantages of federated search are:

1. Users don’t have to know which repository to search or use multiple log-ins.
2. Users don’t have to submit multiple queries to discover all information.

An effective federated search tool will also provide the following features:

- The ability to display collection dependent search summaries. For example, a library publication search result summary may comprise the book title, author and ISBN, whereas a staff directory search result summary could comprise a name, title, email, phone number and identification photo.

- Fielded search across repositories. For example, searching by author should match the 'author' column in the library catalogue and the 'DC.Creator' metadata field on the intranet web pages.

7. Spelling Suggestions, Thesaurus Integration and Clustering

An important characteristic of a high quality search engine is that it won't rely exclusively on a ranked result list to ensure the user is presented with accurate and comprehensive information.

A common problem with database search is that the user is left in an unassisted state where it proves difficult to determine whether there is any relevant information on the query topic, or if there is relevant information, which search terms to use. Three tools for avoiding this problem are query spelling suggestions, corporate/industry thesauri integration and best bets (hard-coded results).

A frequent cause of database search applications returning zero search results is the use of complex queries where there are no records that match all of the search terms. A search engine that can present fully matching results followed by partially matching results is more effective at guiding the user through an advanced search query.

Another characteristic of high quality search is that the user is offered ways to discover the existence of related content that may not have otherwise been found by examining a result set. Clustering and sub-topic identification tools can help guide the user to related products or services. This can result in increased online sales and/or better knowledge sharing within the organisation. These tools can also assist with compensating for the lack of ranking information often found in database applications by giving the users a convenient way to focus in on a refined set of search results instead of wasting time navigating through multiple pages of results.

Clustering provides a summary view of query related sub-topics

8. Work Flow and Integration

Full featured search engines provide workflow integration APIs that enable the search engine to be used as a platform for developing custom search applications. Examples of workflow integrations include:

- Incorporation of external metadata or content from other data sources
- Custom translation or pre-processing of content
- Quality checks
- Custom text filtering
- Triggering of index updates

9. Query Reports

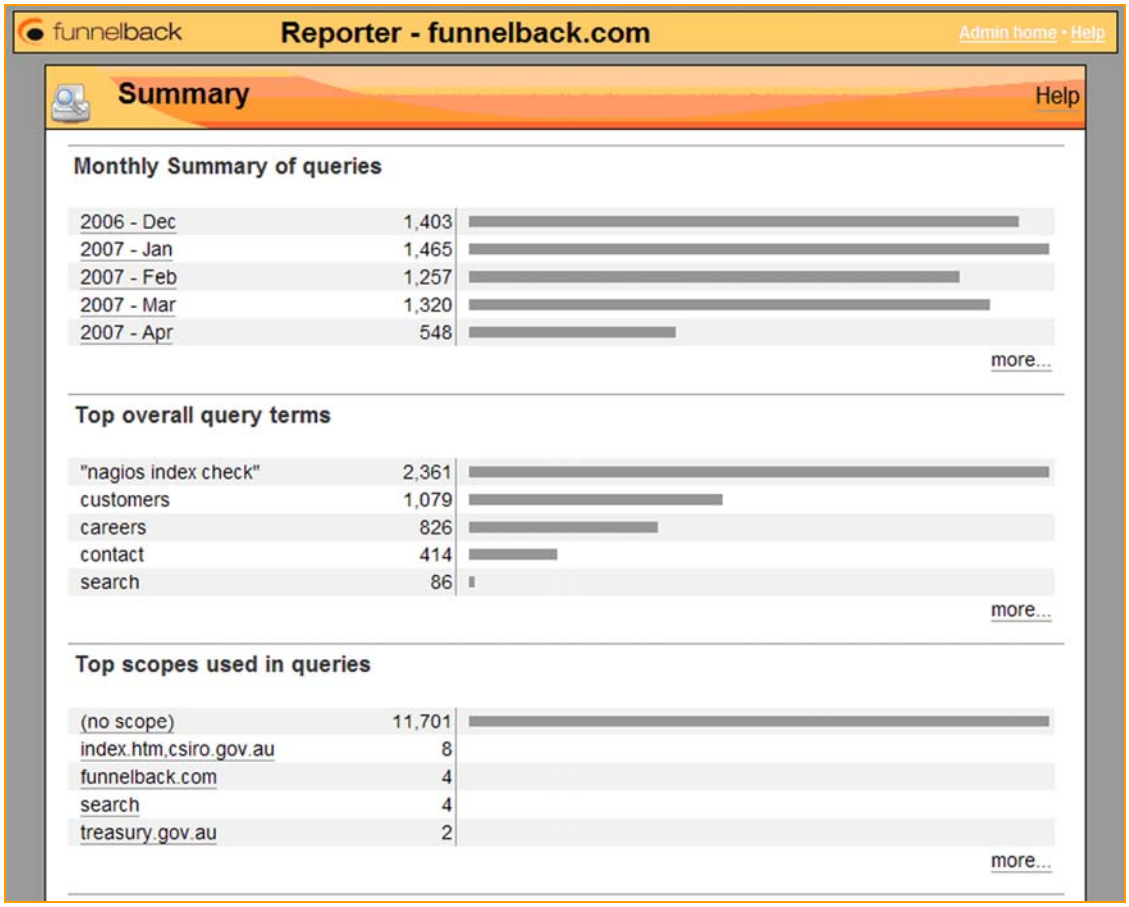
Invaluable business intelligence can be gained from access to detailed query reports. A search engine with an extensive reporting capability will be able to report on:

- The most popular queries within a specified time frame
- Queries that produced no search results
- Queries that did not follow-up with the user clicking on a search result
- Queries that resulted in the user clicking to the next page of results
- Sudden spikes in a particular query
- The list of queries that are guiding users to a particular result

This information can assist organisations to determine what information needs are driving users to use the database application, information needs that aren't being satisfied and queries that are failing to return the most valuable results in the top ranks.

10. Query Expansion

Database search users can often become stranded due to disparities in terminology, query spelling mistakes or information needs that are unsatisfied within the indexed content. These kinds of problems can be identified by examining the query reports to determine common queries that return no search results, generate no result clicks or were followed by a progression into subsequent results pages. Having identified a list of commonly failed queries, use of query expansion mappings will enable the automatic and transparent translation of failed queries into queries that are known to return the desired results or at least results that will assist the user in some way.



The report summary view enables drill-down into more detailed reports

Summarising the Business Case for a Search Engine

The total cost of ownership for a high quality search engine is dwarfed by the following benefits:

- More effective ranking, the presence of spelling & thesaurus suggestions, clustering and query expansion leads to:
    - Increased end-user productivity. Knowledge workers spend 15–25% of their time on unproductive information related activities (IDC White Paper "The High Cost of Not Finding Information", Susan Feldman, 2001).
    - Increased online sales. The Australian Broadcasting Corporation's online shop site reported a 24% increase in online sales after switching from native database search to the Funnelback search engine.
- Decreased database licence and hardware costs due to reduced performance requirements. Also, maintenance of in-house expertise in database optimisation is less important if the real-time database requirements are taken care of by the search engine.
- Valuable business intelligence is gained through detailed query reports.
- Out-of-the-box interface customisation tools reduce initial development costs and ongoing maintenance costs as they require no in-house programming skills to maintain.

- The potential to gain additional value from the search engine by expanding its reach to other parts of the enterprise (e.g. file shares, web sites, other databases and repositories). The ability to search the entire corporate information base with a single query is a challenging yet highly achievable goal with the right search engine.