South African Journal
of Information Management
[ DoE accredited ]

Peer Reviewed Article Vol.1 No.(2/3) September 1999

# Intelligent retrieval on corporate intranets (portrals)

Piet Dempsey
Excalibur Technologies
pdempsey@grintek.com

**Contents**



## 1. Introduction

*An immense and ever-increasing wealth of knowledge is scattered about the world today – knowledge that would probably suffice to solve all the mighty difficulties of our age – but it is dispersed and unorganized. We need a sort of mental clearinghouse for the mind: a depot where knowledge and ideas are received, sorted, summarized, digested, clarified and compared.*
H.G. Wells

Every day, the demand for quality information services grows. More and more information providers – newspapers, research sources, specialty references, regulatory agencies, professional associations and more – are making 'knowledge assets' available to wider and more diverse audiences. Why? The advent of corporate intranets and the Internet itself have made it not only readily possible but almost a requirement to provide quick, easy access to an organisation's documents, speeches, video and audio files, databases and more. The numbers and types of information services are proliferating, with new services and sources emerging for medical research, libraries, electronic publications, legal databases, news archives, government resources and more.

This paper is written for any organisation providing these types of services and will examine the reasons why online services must provide not only comprehensive access to their repositories of knowledge and information, but also why access must be fast, accurate and intelligent. These may sound like fundamental qualities for any business, but in reality very few online services today can truly provide the high performance search capabilities demanded by their customers, whether internal employees or Web site visitors. The bottom

line? Satisfying results keep users coming back.

Top

## 2. Opportunity for profit

The reason for the rush to provide online access is simple economics. By furnishing high quality online content, information providers can expect to:

- Leverage their existing information assets through new, more profitable channels. For corporations this means generating additional revenue through online access at significantly lower costs.
- Comply more efficiently and cost-effectively with government mandates. For government entities this includes meeting EFOIA requirements for public disclosure of information, as well as making regulatory information more readily available.
- Provide a value-add complement to existing resources. If a subscriber to a magazine wants to search back issues on a specific topic, a 'subscribers-only' service, for example, provides the subscriber with value-added capabilities while generating additional revenue for the magazine.
- Offer faster service over traditional methods of printing, binding and distributing information – at a far more compelling profit margin. The latest information can be in the reader's hands in seconds, not weeks or months after a paper publishing cycle completes.
- Expand customer base through the ubiquity, low cost and low barrier of the Internet. Subscribers can join almost instantaneously and get the information they seek immediately.
- Provide members with up-to-the-minute association news. Professional associations, especially, can realise the benefits of giving their members quick, easy access to the latest changes affecting their industry.
- Capitalise on the growing millions of Web-crawling users.

All of these examples – and there are hundreds more – mean the same thing to any organisation: more satisfied customers and lower out-of-pocket production costs. But simply providing access to online content is not enough. Daily advances in technology have made it possible for customers to reach a staggering amount of information. Useful information, however, must be easy-to-reach, accurate once it's found, and provide **exactly** what the customer wants. Raw data and irrelevant information is easy to find; true, meaningful **knowledge** is not.

Knowledge-hungry customers today don't know what to do with much of the information they find. Anyone who has used a rudimentary, Boolean-based Internet search engine has likely been forced to wade through pages and pages of irrelevant search results. Often, even the most obvious answers are buried many 'hits" below. Sometimes, they're not even listed. And sometimes, the whole process is so frustrating that the one thing you want **never** to happen happens: the customer just gives up.

Top

## 3. Search experience is the differentiator

*Search engines on the corporate network now have great recall but little precision. You can get a tremendous range of information, but that's useless if 99 percent of what you get is meaningless. The key to infoglut is refining what is displayed on the return list.*

Hadley Reynolds, Research Director,
Delphi Group

Today's information providers cannot continue to rely on rudimentary search engine technology. The user's overall 'search experience' – the process by which answers are delivered and retrieved – has quickly become just as critical as the desired information itself. The old saying, 'you get what you pay for', is true; by providing a rewarding search experience, customers keep coming back. The real difference between successful and unsuccessful searches lies in the premium aspect of the information. Intelligent, high performance search tools and applications provide intelligent, highly accurate results.

Regardless of whether your service is a searchable morgue of newspaper clippings or a nationwide database of criminal files or an information service for medical professionals or a corporate intranet where employee productivity is paramount, next to the information itself, the search experience is the single greatest differentiator. Even if you offer the most comprehensive collection of information, its value is greatly diminished if your users can't find the exact information they want quickly and easily. In fact, the larger the document collections, the greater the importance of the search experience.

A new class of premium search technologies is needed to power and empower a positive search experience. Solutions capable of 'speaking' a customer's natural language, compensating for misspellings, and understanding the true nature of the search are paramount. The prospect of frustrated, unproductive employees and fed-up customers wading through mountains of data with inadequate search tools means information aggregators need to step up their search offerings. Failing to do so, or being reluctant to bring text-, e-mail-, paper- and word processing-based information online represents a missed opportunity for greater productivity and increased revenue.

Top

# 4. Next generation of intelligent search and retrieval

Information providers in any industry will benefit tremendously from the new class of search technologies available today. The premium nature of general-purpose, as well as specialised, paid-information services rapidly growing in popularity on the Internet today demands extended search capabilities. Meeting all of the new online search 'requirements' can be a difficult challenge for most of today's search and retrieval technology providers, even the ones claiming to offer premium features such as these:

## 4.1 Concept searching

People are starting to think in terms of concepts, not keywords, and their searches are often exploratory in nature. When online users perform a search, they don't want to have to pre-determine what the 'right' keywords will be – or learn what the wrong ones are after numerous fruitless searches. Instead, they want to enter a concept, a concept that the search technology will intelligently recognise, make intelligent assumptions about, and return accurate results.

## 4.2 Pattern or 'fuzzy' searching

In the world of online customer searching, misspellings are a common occurrence. Making matters worse, there may be many ways to spell a customer query. For example, there are more than a dozen ways to spell Khaddafy, the leader of Libya. Entering all 12 in a query is

cumbersome and prone to error. Hybrid, premium search technologies are inherently fault tolerant, compensating for misspellings, spelling variations and even OCR errors. Common Boolean search engines cannot offer this kind of precision.

### 4.3 Natural language querying

Users today form a diverse population with varying skill levels and knowledge of content, and it is difficult, if not impossible, to provide any kind of online training. Most users search the way they speak: in natural language. This can, however, lead to inconsistent results. Loosely structured queries can return thousands of hits that are impossible to sift through. Worse, queries that are structured too tightly can overlook important results. Here is an example of what's required in natural language querying: 'Stock purchases between 100,000 and 500,000 shares in June 1998.' This eliminates the need for operands like and, nor, or, either, not, slashes, quotes and oral algebraic notation. Users can also use idioms such as 'war between the states' to find information pertaining to 'Civil War'. In other words, a first-time user can successfully get results right away, without instructions.

### 4.4 Support for hundreds of data types

While simple text on HTML-formatted pages is by far the most common data format in online databases and services, the amount of other data types is growing rapidly. Video files, pictures and images, relational database tables, sound, e-mail, formatted text PDF files, PowerPoint presentations and hundreds of other data formats are in use today. Modern search tools need to be able to scour all types of data simultaneously. Although many searches are text-based, search and retrieval technology needs to accommodate different asset formats. In the very near future, users will demand to be able to search all these file types in a single query. For example, a news bureau may house text news stories, video footage, and audio clips. A single query on 'Apollo 11' might return detailed stories, footage of the first moon walk, and an audio clip of Neil Armstrong's famous quote: 'That's one small step for man, one giant leap for mankind.'

### 4.5 Hit highlighting

Not all query results are self-explanatory. It can be confusing, frustrating and time-consuming trying to determine why a hit was returned and where the relevant material is located. For example, if a search on 'Apollo 11' returns a lengthy technical article on rocket construction, the reader may have to read through dozens or even hundreds of pages to find the important passages in the file. 'Hit highlighting' simplifies this process by highlighting keywords containing the intended topic in an accented color to take the user to the most relevant portions of the document. It's an overlooked but crucial detail.

### 4.6 Document summarisation

Like hit highlighting, document summaries are a welcome tool for searchers who want a quick synopsis before retrieving the entire hit (and perhaps before paying for that hit, if there is an additional fee). Although some simpler search tools use document summaries, the results can still be lost in virtual 'haystacks' – after the online user has wrestled through the results from a commodity search engine, he or she still doesn't know where the needle is. High performance search and retrieval technology lists the specific information a user has requested – based on the concept search, not the keyword – for instant, easy scanning. There's no combing through piles of hits trying to find relevant passages.

### 4.7 Scalability

One of the barriers that has intimidated many information providers is a concern about scalability. Scalability is more than just fast searching. Rather, it is the ability to maintain search performance even when the demands on the system rise by orders of magnitude. Online services typically have unpredictable numbers of users searching large, heterogeneous document collections. The search tool must continue to perform as the number of users and the size of the databases increase. The growth of multi-media repositories, such as video and image archives, simply ensures that the scale will grow ever larger.

Top

# 5. Future of search and retrieval

Search services have already begun to have even life-saving applications. Physicians Online was launched in 1994 to help doctors, medical personnel, and healthcare institutions search for critical – yet sometimes confusing – medical information located in thousands of databases. Physicians Online also chose Excalibur RetrievalWare for its plain English searching, rich functionality, and its ability to fully exploit the Unified Medical Language System (UMLS) of medical terms. GTE Enterprise Solutions recently introduced the Bastille®, a Web-based service that offers a highly secured application for information sharing and real time communication between law enforcement agencies via the Internet. Individual case histories can be cross-referenced against multiple jurisdictions to examine for patterns and apprehend criminals.

The future of search and retrieval technology will continue to mature to keep pace with the rapid growth of new content and new formats. As we have already discussed, customers will demand not only greater access to new online content, but at increasingly faster speeds and with more accurate results. If their answers aren't found within the first half dozen returns, then the customer will be lost.

Knowledge sharing will dominate the corporate, utility, governmental and non-profit industries into the next millennium. As a critical component of 'knowledge management', those organisations who can more quickly and effectively offer their resources to their customers and employees will have greater chances of success and increased profits. Mediocre search technology is no longer enough; to power a positive search experience, online providers must be powered with premium search tools and applications.

We hope this paper has been a useful introduction to the many applications of search services as well as the requirements for the next level of online search technologies.

Top

# 6. Digital revolution

## 6.1 Video – the real opportunity

Video, television, computers and the Internet are evolving and converging into fascinating new solutions, building on the strengths of each to create the most powerful, extensible communications environment the industry has ever known. Spurred by the proliferation of technical advances and the rapid growth and flexibility of the World-Wide Web, video has become one of the most valuable mediums for communications today.

What makes video such a 'rich' commodity? Video encompasses not only the spoken word,

but sight, sound, images, the textures of light, colour and time, and much more. For those organisations that rely on video, having the means to more quickly, accurate and efficiently deconstruct, analyse, index, access, browse, search and retrieve video assets can offer enormous benefits. Those organisations with flexible, intelligent access to their video-based resources can more effectively leverage those resources to bring new products to market faster. This helps create additional revenue and minimise costs associated with finding, researching and viewing video archives.

The 'problem', however, is that until recently, video-rich organisations have found it cumbersome and difficult to manage. Producing a specific program many times requires expensive offline editing and online finishing. Additionally, to broadcast video over networks requires broad bandwidth and powerful hardware. And to do so requires digitalisation; most video archives today still exist in analog form. Indexing and searching analog-based archives in a streamlined, digital manner have become laborious, involved processes.

Large organisations, such as TV networks or production companies, may have tens of thousands of hours worth of video stored in just one warehouse. In most cases, the only method for searching these vast video libraries is by hand – a time- and people-intensive process that slows production, inflates costs and frustrates everyone involved.

Typically, a producer or director requests screening copies of archived video to preview for possible use in a current project. This sends a worker scurrying through a warehouse to pull backup copies of the video so preview can be made. The producer/director then spends hours manually shuttling the preview tape forwards and backwards to locate and select the specific frames to be used. Once particular segments have been specified, the process begins anew with another request going to the archive to pull particular backup tapes for use in offline editing and online finishing. The entire process is time-consuming, inefficient and expensive. As a result, video assets go under-utilised, production cycles lag, and new production is slow.

The challenge – especially for organisations in the media, broadcast and entertainment industries – is to create new video content and programming within shorter, less-expensive production cycles. That means learning how to efficiently leverage existing video assets for re-use and re-expression, and effectively manage the analysis and production of new video footage. The less time workers spend viewing and previewing videotape is time (and money!) saved in pre-production shot selection, and in expensive editing suites.

ISSN 1560-683X

Published by InterWord Communications for the Centre for Research in Web-based Applications,
Rand Afrikaans University