

Leveraging Content in Enterprise Knowledge Processes

Ramana Rao

“A wealth of information creates a poverty of attention”

Herb Simon

Introduction

Knowledge work depends, almost by its very definition, on the use of information and data. Large amounts of information resides in the form of so-called “unstructured data” available in large organizations. Thus, leveraging available content in the knowledge processes in large organizations has been identified as an important opportunity. Many see content as a strategic element in the processes of capturing, sharing, and reusing knowledge. Yet, current approaches toward content access and use show great limitations even after huge investments over years.

In fact, content may be the most underutilized asset in large organizations today. Organizations typically buy and create large amounts of content at great costs, yet they often fail to truly leverage it. Content refers to collections of electronic textual documents including research reports, product collaterals, development specifications, internal memos, sales materials, patents and invention proposals, press releases, news articles, scientific literature, email messages, and so on.

The causes for content underutilization are complex and varied. First, people, when overloaded by information, tend to detune and make simplifying decisions, often oversimplifying when considered from a broader perspective. Content often lacks sufficient “metadata” characterizing the subjects, the sources, and other facets of the content to support effective access. And, finally, the access tools of search and browse themselves are often create problems, failing by being too brittle, imprecise, hard-to-use, or inefficient. All of these factors lead to users not using potential valuable content, which in turn leads to broader organizational problems of redundant work, costly mistakes, and missed opportunities.

This chapter focuses on a more intelligent approach toward content access and use. Access refers to not just finding relevant content, but also understanding what has been found. Furthermore, other kinds of content use applications beyond information retrieval, strictly defined, show great potential. As we approach ten years of mainstream retrieval experience on the Internet and Intranets, the limitations of traditional search and browse systems are now quite clear, particularly in the face of the increasing complexity and competitive pace of the world.

The key insight marking the path forward is to respect the reality that content is not in fact, “unstructured data,” as is characterized in the industry, but in fact, expressions of human language. The next section takes a look at shifting this perspective and others. Following that we look at broader classes of content use applications, the underlying technologies of content analysis, and finally, at specific examples of enterprise applications. We conclude with a few key recommendations.

Shifting Perspectives

Shifting our perspectives on the nature of content, the use of content in knowledge work, and the requirements for access technologies show the path toward better leveraging of content.

From Unstructured to Richly Structured

The phrase “unstructured data” is meant to contrast documents to data as it is typically stored in relational databases in rows and columns. This characterization reveals a technology bias, since in fact, to humans, documents are nicely structured, whereas databases are generally unfathomable! The problem is that there is not enough time or money to access the structure within and across documents using humans. Thus the game is now about finding ways to analyze and organize content using software-assisted processes that utilize human effort efficiently and effectively.

The content management paradigm tends to view documents as if they are records, with one big field or blob which is the content of the document. Content analysis is about drilling inside the boundary of the document to extract or analyze specific aspects of the content, particularly recognizing that content is linguistic in nature. This granular data then allows for a better contextualization of documents into conceptual spaces and connecting of the document with other documents. The shift in perspective is from regular tables of documents with minimal file system style metadata to association networks across and within documents. It is this kind of rich, not impoverished, structure that any human quickly appreciates about the “hyper-structured” web.

From Transactional to Knowledge Work

More important than the stuff (i.e. content) are the work processes which the stuff serves. Could you imagine calling knowledge work, “unstructured work”? Indeed, it is much harder to characterize what is actually happening in knowledge work, but still there is an opportunity to better support these more open processes. The content management perspective here also limits the use of content: the focus has typically been on production and control processes defined as transactional, tightly-scripted, repetitive workflows. Thus it tends to be used in well-defined, late stage processes, for example, technical publication, insurance forms processing, and new drug submissions.

In contrast, knowledge work, by its very nature, can not be tied down in strict workflows. The processes are inherently exploratory, creative and analytical in nature. The challenges become evident even in finding documents, never mind using them. Search

tools are brittle in that they provide almost no value when you don't really understand what you are looking for. Besides trying to repair this problem and open up the possibility of other applications beyond retrieval, there is also an opportunity to provide "loose coupling" in collaboration and communication processes that are critical in knowledge work. Content can be a bridge between people across time and space and social structure.

From Finding to Using

Search has been the focus of past efforts to leverage organizational content. Yet, besides the challenges of using search tools to find relevant content, the user is still left largely unsupported with the work of understanding what is found. In actuality, users are not typically interested in the documents per se, but rather what the document say about the world. Here again, the key insight arises: content is made out of human language statements about the world. Right now, the use of the statements is left only to humans, which greatly limits the use of content.

Thus, fully utilizing content will depend on technologies that focus on processing the statements in the content not just tools for the finding of documents. Not just single statements that stand out for their uniqueness or relevance to our pursuits, but also patterns over entire collections. There is signal and meaning in the stocks and flows of content, and we can go after these with software. For many, this will conjure up the spectre of solving the grand scientific challenge of natural language understanding by machines, but there are sound and viable approaches that lie in a happy middle between leaving it to humans alone and relying on natural language understanding by machines.

Classifying Applications

All content use applications have something to do, not surprisingly, with content and with users. Particularly, they all enable some kind of interaction between the information needs of humans and the meaning-bearing streams of content. Differences in the nature of interaction and handoff between the system and the user define distinct types of applications. First, activity may be driven by the user, or instead, by the flow of content. Second, the focus may be on providing documents to the user, or rather on analyzing or processing the contents of statements contained in the documents. These two distinctions capture four basic types of applications:

Retrieval – users find and understand relevant documents

Routing – system routes relevant documents to people

Mining – users explore or analyze collections or flows

Alerting – system generates events or reports sent to people

In retrieval applications, activity is user-initiated based on information needs that arise during tasks or projects. Retrieval applications are certainly the most widely-deployed and understood type of application. Information retrieval has been an active field for almost the entire history of computing, and the Internet has catapulted it into the mainstream. Though the focus with retrieval is on finding documents, the requirement of relevance underscores the importance of knowing what a document is about. So, even here, the use of content analysis can dramatically improve retrieval systems.

Routing flips the retrieval paradigm by turning the pull of retrieval into the push of content-triggered delivery, for example, to an email box. Routing makes sense when information needs are not just one-time, but rather, recur based on broader roles or organizational needs. A simple example is a syndication service that matches new documents against saved queries or users profiles. Broader organizational applications include routing of documents to the right people for further processing e.g. routing patents to examiners or support cases to relevant specialists. Because routing "pushes" content at people, it requires finer discrimination on what the documents is about, otherwise the push quickly feels like shove.

While retrieval and routing applications can be improved by finer-grained processing of contents, mining and alerting applications absolutely require such processing. Mining applications enable users to explore the statistics of content collections or flows looking for interesting patterns or occurrences. Mining applications turn text documents into structured data that can be combined with other data sources and integrated into statistical or business intelligence applications. Alerting applications are the routing style obverse of mining. They notify users when particular patterns or events occur in content flows, or regularly route canned analysis (i.e. reports) to the users.

Analyzing Content

All the types of applications described above depend on technologies for analyzing content to "understand" some portion of the meaning of its statements. Somewhere between one extreme of completely depending on humans to extract meaning and the other of expecting machines to fully understand content themselves (whatever that may mean), approaches for extracting particularly useful aspects of meaning are now becoming quite viable.

Content analysis can be viewed as the processing of content into structured representations or databases that captures some aspects of the meaning of the content's statements. To get at meaning, we can ask about what is the statement talking, and about that, what is it saying? These questions highlight the two basic mechanisms for meaning in statements. Statements "refer" to objects in the world and they "say" something about them.

A search index can be seen as a trivial example of such a structured database. It provides a table of how many times and where words are used in the documents of a content collection. Its model of the world is that the world has documents in it, and that the words used in a document tell you what the document is about. At the other extreme is a rich

semantic network of the type typical of knowledge-based systems in artificial intelligence. Such semantic networks try to model a more complete "meaning" of the statements to support machine reasoning systems.

In between these two structures, we can imagine a database that like the semantic network truly is referring to objects in the world, but that makes more limited types of statements. These statements are of high value in particular applications and can be reliably generated from textual content. Again, it's about looking for sweet spots that balance utility and viability.

For example, consider a collection of articles about company events. The world covered by the statements in the collections is familiar. It includes people, companies, roles people play in companies, corporate event (e.g. founding, bankruptcy, mergers and acquisitions) and so on. A structured database over this space of objects and relationships would capture more meaning than a simple word index while not provided the structure to answer arbitrary questions that could be answered based on the articles.

Such information extraction technology isn't yet applied in most industries, but it has become mission-critical in government intelligence and is quite common now in the publishing and pharmaceutical industries. It includes what is called entity extraction, figuring out about what objects in the world a statement is talking about, and fact extraction, figuring out what the statement is saying about them. It focuses on the meaning of statements in content and on the problem of graspability rather than that of findability.

Another key technology, Automatic Categorization, which is more common, is really about the mapping of document into conceptual spaces. Particularly, categorization is about the filing of documents into an organized classification structure (typically called a taxonomy). For example, the filing of books into the Dewey Decimal System, or of Web sites into Yahoo! The value here is in automating the filing process, so that enterprises can affordably and reliably categorize their private content. This is can never be a fully automated process, because a dead taxonomy stops being relevant in the same way as a dead language does, but human involvement can be optimized to make the overall process effective and workable.

Supercharging Retrieval

Retrieval applications, search and browse style applications, can be improved dramatically using taxonomies and extracted information. In this section, we will illustrate this point using an application based on Inxight SmartDiscovery. In a new more powerful browse style interaction, a user accesses a collection by navigating across a taxonomy. In fact, this is a familiar paradigm, not just from the organization of physical information resources as in libraries, but also in the electronic world. The early popularity of Yahoo! demonstrates the appeal and usefulness of high-quality taxonomy, organized by human catalogers. Automatic categorization enables this style of interaction in settings where full human cataloging is not an option.

A visualization suited to large hierarchies is ideal for helping the user quickly understand the taxonomy as well as drilling down to categories of interest. The taxonomy and the visualization together are effectively operating as a conceptual and perceptual map of the content collection. As with real maps, these structures allow a user to get an overview of the territory as well as navigate to specific areas of interest. In Figure 1, Inxight Star Tree is used to show the structure of the taxonomy.

Also, as with maps, visualizations can provide a backdrop for showing search results. For example, a search for ‘war’ shows that there are matching documents in numerous categories. By selecting one or more of the categories containing matches, the user can filter the results based on their understanding of the taxonomy. This kind of filtering is effectively equivalent to doing an advanced query, but the user sees what he wants rather than thinking of what he wants before he says it. Conversely, by looking at where there are matches in the taxonomy, the user learns about the taxonomy during the process. Thus search teaches the user to browse better.

The screenshot shows a web browser window displaying the Inxight search interface. The search query is 'war'. The interface features a 'Filter Results Mode' section on the left with a 'News Navigator' tree. The tree is a radial diagram with 'war' at the center, branching into various categories such as 'violence', 'terrorism', 'justice and rights', 'economy, business and finance', and 'social issues'. The right side of the interface shows search results for 'war', including document titles, previews, top entities, and categories. The browser's address bar shows the URL: <http://localhost:7474/News1collexp/jsp/two-frames.jsp?sh=kw=war&view=taxonomy>.

Figure 1: A visualization can provide a perceptual map of a taxonomy which in turn provides a conceptual map of a content collection.

The right side of Figure 1, shows the documents of a selected category as a result list. In addition to a title and link for matching document as typically shown, the previews show “a little bit more but not too much” about each document. The previews include a query-sensitive summary of the document as well as a list of entities (e.g. organizations, people, places) mentioned in the document and all categories that the document matches.

This query returned a large number of documents and it would still take a long time to look through all the results. The document list view, shown in Figure 2, provides additional tools based on extracted information. As with a back of the book index in a book, the additional indices help a user understand an entire result set. In particular, three index types are shown. The first index shows the concepts related to war found in the results, the second shows the matching categories from the taxonomy, and the third shows the entities of different types found.

The screenshot displays the Inxight search interface for the query 'war'. The top navigation bar includes the Inxight logo, a search input field, and options to search within results. Below the search bar, there are tabs for 'VIEW' (taxonomy, document list, document) and a 'HELP' icon. The main content area is divided into three vertical panels on the left and a large preview area on the right.

- concept filter:** A tree view showing hierarchical concepts related to war, such as World War (6), Gulf War (5), Vietnam War (1), and Star War (2).
- category filter:** A 'News Navigator' tree view showing categories like 'crime, law and justice (11)', 'politics (20)', and 'act of terror (171)'.
- entity filter:** A list of entities categorized by type: Company/Organization (Iraqi Information, etc.), People (Bulgarian Deputy F..., Bush, etc.), and Place (Abu Dhabi, Afghanistan, etc.).

The right-hand preview area shows a list of document titles and snippets. Each entry includes a title, a brief description, top entities, and categories. For example, the first entry is 'dywni-bc-lebanon-arabs-spot.html' with the snippet 'Arab officials warn of Iraq war impact' and top entities 'Iraq (2) | Kuwait (1) | Arab League (1)'. The interface also includes 'PREVIEW OPTIONS' (set to 'complete') and pagination controls for 'Results 1-10 of 404'.

Figure 2: A search for the keyword “war” generated 404 ‘hits’. The tools on the left are both views and filters that help users understand the entire result set, then focus in the portions of interest.

These indexes are live filtering tools that can be used to refine the query, thus through the availability of metadata allows users to create advanced queries on the fly as they better understand what they are looking for, and learn about the available content. As users narrow the search results list, the index/filter tools dynamically update to show the information about the refined result list. When the user finds a document they are interested in either by browsing the taxonomy or refining a query, the user can focus in on the document and continue to leverage extracted information within the boundary of the document itself as shown in Figure 3.

The screenshot shows the Inxight Collection Explorer interface in Microsoft Internet Explorer. The browser title is 'Inxight Collection Explorer - Microsoft Internet Explorer'. The interface includes a search bar with the text 'SEARCH' and a search button. Below the search bar, there are navigation tabs: 'VIEW' (selected), 'taxonomy', 'document list', and 'document'. A 'HELP' button is also visible. The search results show '[SEARCH: iraq] + [CONCEPT: Scud Missile Range Of Iraq]'. The main content area displays the document 'Document 3 of 4' with the title 'Rumsfeld irate over secret plan on Iraq' and a URL 'C:\data\News31203\cleaned_news_feb2003\dysk7-bc-us-rumsfeld-germany-3rdld.html' dated '02/08/2003'. The document text discusses U.S. Defense Secretary Donald Rumsfeld's reaction to a proposal for beefed up U.N. arms inspections in Iraq. On the left side, there are filters for 'in this document' (CITY: Berlin (1), Kyoto (1), Munich (1), Paris (1); COUNTRY: Afghanistan (3), America (1), France (7), Germany (10), Iraq (17), Israel (1), Libya (1), Macedonia (1), Russia (1), the Netherlands (1), Turkey (8), U.S. (5), United States (13), Yugoslavia (1); DATE: 02/08/2003 (1), 1995-07-10 (1), 2001-09-11 (1); DAY). On the right side, there are sections for 'similar documents' (listing various document IDs) and 'new topic search' (with filters for Company/Organization, Place, and News Navigator).

Figure 3: The document level view allows quick grasping of the document content, and using the document as a source of queries back out to the collection.

On the left side, an index of the people, places, dates, measurements, etc., discussed in the document provides a quick way of locating specific information in text. On the right side, the document is used as a springboard to find other documents either using a “more

like this” or “relevance feedback” capability, or by using the categories and entities of the document to start a new query.

Leveraging Content

The work in most industries can be characterized as a chain of activities starting with science or engineering or design or development at one end and ending with marketing and servicing at the other. For example, the work of a large pharmaceutical company starts with fundamental science, flows through drug discovery to drug development to clinical testing to drug approval to commercial exploitation. These stages are typically seen as distinct and often content from one is not leveraged in other stages. Better content analysis can enable the repurposing of content across different stages in a number of applications that apply in many industries. Many of these applications address typical knowledge management concerns:

Mergers and acquisitions. Mergers depend crucially on being able to integrate the content resources of multiple organizations, particularly because large mergers are usually followed by attrition and headcount reduction. Meanwhile, the new organization typically has to handle all the same workload, so it becomes all the more important to be able to understand what information is available and to use it after the merger.

Corporate Licensing. Many large corporations accumulate large intellectual property (IP) portfolios through Research and Development as well as Mergers and Acquisitions. Increasingly, corporations look to external sources to license key technologies and look for revenue opportunities from licensing their own IP. Beyond the patents of a company, this activity requires dealing with other internal documents, the patents of others, and external scientific, technology, and marketplace documents.

Competitive intelligence. Monitoring the market for competitive and marketplace dynamics is one of the oldest applications of search technology. Yet, this application is fundamentally about the fine-grained understanding of the interactions between the players, products, technologies, strategies, actions and so on in the marketplace. In the past, large companies tended to serve this function through small departments staffed with skilled research librarians and competitive intelligence specialist that followed well-defined methodologies. This approach hasn't been able to keep pace with the increasingly complex competitive and marketplace landscape, nor with the increasing variety or amount of available information and user needs across large global organizations.

Product Development. Companies produce large amounts of content during research and development as well as attain publicly or commercially available content. For example, life sciences companies leverage public content funded by government agencies, e.g. National Institute of Health, as well as content from large electronic publishers. The pressures in the pharmaceutical industry are rapidly mounting to improve their drug discovery and development processes. Though work has gone into integrating and curating structured data sources (e.g. experimental data), internal textual content remains relatively underutilized.

Marketplace Feedback. Internet content sources and customer email and surveys contain valuable feedback to an organization. Monitoring statements made about a company or its products in the press, on websites, in blogs, in discussion groups, and directly to the customer support organization can help evaluate brand perception and company reputation. Such monitoring can help tune corporate and product marketing activities, as well as help focus product development efforts on important areas for improvement or greater opportunity.

Supplier management. Large organizations that provision products and services from a large number of suppliers often struggle with product documentation and service level agreements. Internal users (e.g. product development) often must depend on an internal service department to figure out how to find necessary information. Suppliers and products and agreements are constantly changing, so manual organization efforts quickly fall behind.

A number of other applications, outside of knowledge management areas, are becoming important in large enterprises. These applications are well worth understanding because of their urgency:

Regulatory compliance. Increasingly, large businesses or organizations are being regulated by laws or proactive policies to disclose various communications or documents to the public or to governmental agencies; or to monitor or restrict certain communications with their customers; or to retain or destroy documents for some period of time or under certain conditions. Examples of regulations include the filing requirements on customer complaints related to pharmaceuticals, HIPAA in the healthcare industry, and of course, the most visible of such regulatory acts, namely, Sarbanes-Oxley in the area of corporate accountability. A typical example of an application of extraction technology in this arena is to monitor emails between brokers and their client for inappropriate messages and forward them to compliance officers.

Legal Discovery. In preparing for litigation, law firms, on behalf of their clients, dig through thousands or millions of documents looking for evidence to build their cases. Indices of the people, organizations, and subjects and maps of the communications can help focus or prioritize discovery work. As a case develops, it also becomes important to re-search based on new lines of thought. Because many of the documents are informal and are created by different people, it is important to be able to deal with vocabulary and name variation. These highlighted aspects of legal discovery also apply to many of the other collection-oriented applications below.

Customer Self-service. All successful product companies must ultimately focus on support costs for their products. One strategy that many companies are pursuing is to publish product and support information through interfaces that allow their customers to retrieve relevant support information directly. Besides mitigating costs for the company, a positive user experience that leads to solving the customer's problem can also enhance the company's brand.

Conclusion

The use of content analysis technology and deployment of content use applications beyond standard search or browse applications are just now starting to become viable in enterprises. Much can be learned from the experience of leading adopters in government intelligence and law enforcement agencies. The pressures on these organizations are extreme and they can not afford to not leverage available content in urgent missions.

A number of intelligence agencies are focused on the mission of counter-terrorism, while law enforcement organizations continue their pursuit of criminals or even better their activities prior to the committing of their crimes. These missions have access to huge repositories of textual content gathered by multiple agencies and departments including field reports, collection summaries, immigration records, Web page content, emails, message traffic, open source news feeds and the like. Though flexible and powerful user-driven retrieval applications are important, routing, mining and alerting applications based on content analysis and information extraction are tantamount to directing human resources to consequential activities. In this arena, a number of important principles appear underscored:

Humans Matters. For the foreseeable future, it's unlikely that any automated approach is going to succeed in truly understanding natural language documents. Thus the question isn't about eliminating human intelligence, but rather about how to design overall systems that arm humans with effective computational tools at the end user task and organizational levels. Years of investment in artificial intelligence have only clarified the absolute necessity of using human intelligence in government intelligence and law enforcement.

Language Matters. Content is, once again, made of human language. There's no avoiding the fact that language is made up of words in natural languages like German, Korean, Farsi, and English, but it goes beyond that to all the specialized languages we speak, from Terrorimese to Xeroxese to Medicalish. Furthermore, for the same reasons that there are so many languages, most terms have many surface forms. For example, consider the spelling of a foreign name in English, or the same molecular compound in various scientific names and in drug/brand names. Any system that ignores this is missing the fundamentals of *natural* language.

Architecture Matters. Often there is great debate about algorithms and their accuracy, but since full text understanding isn't around the corner, the focus should be on the overall effectiveness of the system. This realer, bigger concern points at the importance of the architecture of systems. System must support the blending of human resource and computational processes, and the overall management of multiple and various algorithms and their associated data.

Speed Matters. Again, since text will not be processed once and its full meaning captured forever, text will often be processed again and again for different purposes or projects or as new algorithms or language models are updated. Certainly, the actual text stream available to intelligence and law enforcement streams is torrential, but that same text

stream is being processed over and over. Cascaded architectures based on sound and speedy processing at different levels can support overall efficiency, allowing simple changes to be made more aggressively, while still allowing for more complex updates.

As enterprises looked to deploying content analysis infrastructure that can be used across a variety of content use applications, they should carefully consider these principles. To distill this further, the key observation of this paper is we must keep our eyes focused on two essential realities if we are ever to truly leveraging content. One reality, about content, is that content is text is language. The other reality, about leveraging or use, we have to think about the overall effectiveness of systems at organization and at user levels, not the narrow level of algorithm accuracy or other technological virtues.