inxight

Rich Interaction with Content

A Technology Paper by Ramana Rao, CTO of Inxight Software, Inc.

Summary

Content is perhaps the most ineffectively used assets in large organizations today. The primary reason is that content is not provided with structure and tools that enable employees to quickly find, understand, and utilize the revelant information when, how and where they need it. Search and browse, rather than being the two alternatives, are really suggestive of two endpoints of a spectrum. A number of new technologies for structuring and accessing content fill in this spectrum. Integrated or blended with conventional search and browse, these new technologies enable the kind of rich interaction with content that is necessary for solving the problem of "underutilized" content.

The Business Problem

Businesses buy and create huge amounts of potentially valuable content. However, documents are typically used once and then lost, and never to be used again, at the significantly lower cost of even the second turn. In fact, content including email, working files, analyses, plans and so on, may be the most ineffectively utilized asset in companies with a typical Information Technology infrastructure.

The problem appears at two levels: the user problem that is related to the obstacles faced by an individual knowledge worker, and the problem that is related to the loss of a business from aggregate behavior. The user problem is often labeled "Information Overload", conveying a picture of a poor user flooded by ever-growing waves of content and data. A spray of statistics (itself part of the flood) indicate hours of time wasted by employees searching for information, or sorting, assessing, and understanding information after it is found. The corresponding business problem is the cost associated with the wasted labor and lost productivity.

A bigger concern is based on the proposition that most people don't waste their time fighting systems or tasks that don't pay off. Very few users are willing to spend the time learning even the simplest of advanced search features. People almost always move on when they can't find useful information quickly. And, they are unlikely to grind away at digesting poorly organized or apparently featureless piles of documents.

Thus information that could potentially make a difference in a particular decision, task or project is not utilized, leading to uninformed decisions being made, risks being overlooked, and opportunities lost. This, in my mind, is the most compelling business problem. It is as much about top line as bottom line, even in tight times.

The Nature of the Solution

If the problem is that content is underutilized, then the solution has to be about utilizing it more. Sure. So what does that entail? Quite simply, I think it means that we must enable richer interaction with content than we currently do. To understand this, it's important to appreciate a few points about the nature of content and interaction.

First, regarding content, despite often being called unstructured data, content is hardly unstructured. To us humans when we look at it or read a specific piece of it, it is not a still, clear pool of liquid information. Rather, it is shaped both by intrinsic aspects of representation and expression, and by the social context in which it is produced and consumed.

Consider the example of a magazine. You would hardly call it "unstructured". You can flip through it quite quickly, rattling off observations, with hardly a glance here and there, as fast as you can speak. Here's the table of contents, and here's the editor-in-chief. "Oh, I hate that color, but what a nice dress." This is an ad, this is a feature article and this one is really confused. Our problem, as humans, starts not with one magazine, but with a stack, as they pile up across months of being just too busy to keep up with the *Mother Jones*.

The phrase, unstructured data, is meant to point out that computers have difficulty routing and accessing content, slicing and dicing it, and manipulating it in all the ways they can; what's stored in rows and columns in the sacred vault of information technology - the relational database. The implied challenge is about devising methods for computers to extract the latent structure embedded in content. Without access to such structure, computers can't be as useful in helping us deal with the volume of information.

Turning to the nature of interaction with content, we can readily see parallels. Interaction with content is embedded in larger processes of work. And again, intrinsic aspects of interaction and work, and their social context, impact the effectiveness of the interaction. For example, we are often under time pressure or other constraints related to our social context and often the tools at hand are disconnected from and inadequate to our tasks, leaving us to our own resources.

Certain kind of highly structured routine work that aligns well with databases has been automated to a greater degree. However, workplace anthropologists commonly point out that even after automation there is much more "unstructure" in these jobs than the boss knows or perhaps, wants to know. In contrast, knowledge work has much more structure in it than current tools support, yet to date the support for it is quite limited.

We access information for a variety of purposes and in a variety of ways according to our purpose. Sometimes we are surveying an area of knowledge, trying to get a general understanding of what it's about or what's available. At other times we are searching for specific answers. Sometimes we wander with a vague sense that something important is on

the verge of crystallizing in our understanding. And, at other times we are skiing a series of tight turns, narrowing in on our target.

It is this range of purpose and context that the conventional tools of information access, search and browse, are meant to address. In fact, as a pair, they seem to bracket the examples above quite well. Yet by now, we are all familiar with their strengths and weaknesses. They have opposite profiles.

Search is precise yet brittle, while browsing is robust, but vague. Search works quite efficiently when it works at all. If you know what you want and how to say it, and if it is there, then nothing works better. How often is this the case? Search is precise like a scalpel or laser, but can it alone help you discover or understand the disease?

Meanwhile, you can always browse, but it's not clear what you are doing. You scan a page, you read a little, you click, you're somewhere else, you aren't sure where you are, you lose track of time.

Rather than think of conventional search and browse as the only two alternatives, it is more productive to see them as two ends of the spectrum. The new technologies described below are all about extending, blending, integrating, or optimizing search and browse in various ways. These unstructured data management (UDM) technologies in their various ways get at the rich structure of the content to enable richer interactions with large amounts of content.

Automatic Categorization

Simply, categorization systems assign documents to categories organized in a hierarchical structure typically called a “taxonomy”. A categorization system creates and maintains a taxonomy and assigns documents to categories. A typical application is to use the taxonomy as a navigable directory for a high value collection of private content, a "*Yahoo!*" for your content.

Taxonomies are typically blends of classification schemes (e.g. the Dewey Decimal system) and controlled vocabularies (e.g. Library of Congress Subject Headings) used in systems for cataloging and indexing library collections. They arrange subjects or topics into a hierarchy from general to specific categories. For example, at the top you might see Art, in the middle, Postmodern Neon Art, and at the bottom, American Postmodern Neon Art in the 21st Century.

The effectiveness and the design of a taxonomy depend on many factors. It's quite easy to get lost in a thicket of theoretical concerns, but it helps to keep in mind that the ultimate evaluation of a taxonomy must be made at the point of use. Thus what the user will be doing, who the user is and what he knows and understand is much more important than abstract properties of knowledge.

General collections and general tasks aren't likely to be the highest value applications for enterprise categorization in the near term. Would Pfizer want to see random content on their intranet in a Reuter's view of the world? Or rather, some private collection of content in the world according to Pfizer? Perhaps the answer is yes to both questions, but the latter seems to hold more obvious value.

After all, it is exactly a privileged understanding in a focused area of knowledge or practice that creates a competitive advantage.

So it is likely that a general-purpose taxonomy would be less useful than appropriate, specialized or even private taxonomies. Not only are focused taxonomies likely to make finer grain discriminations on topics in more specialized collections, but they are also likely to better match the language and the purposes of specialized users and uses. For example, the Medical Subject Heading (MeSH) taxonomy which has been used and evolved over many years in medical research is more suited for organizing Pfizer's reports for its thousands of drug researchers.

Beyond defining and managing a taxonomy, a categorization system must be able to accurately assign documents to one or more categories. Most of the leading products employ a mixture of multiple methods, including linguistic analysis, statistical inference, machine learning, and rule-based processing to perform this. There are likely to be ongoing debates about the relative value of different methods and the best ways of blending them. For the moment, probably many other aspects of the categorization system beyond core accuracy are likely to make the critical difference in choosing the best product.

Once a document has been assigned to categories, these category tags or codes can be used for a number of applications:

- **End User Browsing**. As the analogy with *Yahoo!* suggests, a hierarchical directory is a way to allow browsing while gaining some of the precision and control of search. It effectively turns the command line interface of search into a hierarchical menu of topics. Browsing the taxonomy imparts an overall sense of what's available and how it is organized, and then allows drilling down to specific topics of interest.

- **Document Routing**. If end user browsing is about the user accessing a pool of past content, this application is about processing the stream of new content and triggering appropriate enterprise action or response. A user can be notified automatically as documents are assigned to categories that relate to their work. A document can be routed to the appropriate employee for processing.

- **Enhanced Search and Browse**. Search can be blended with a taxonomy in a number of ways that enhance search itself or that enable a tighter interplay between searching and browsing. Showing search results organized by the categories of the matching document helps a user more quickly digest the results. The search results can also become the point of departure for browsing, since a document's category is likely to contain other relevant documents that might not have matched the query.

A related technique likely to be popular is to search from the top of taxonomy to discover relevant categories and then navigate up and down from the matching categories.

Information Extraction

Information extraction systems extract elements of information from documents and collections. This technique is akin to what we do as we scan an article to assess its relevance to our goals. Based on linguistic analysis and scanning for patterns among words and phrases, extraction identifies many of the things people can quickly find in text without reading for deeper meaning. Two levels of extraction are common:

1. **Entity extraction** focuses on identifying "named entities" like people, organizations, products, and places. Also important are other special noun groups including large noun phrases that indicate topics or concepts, as well as entities like dates, quantities, and measurements. Even this basic level of identification can dramatically improve higher-level capabilities including categorization, search, and automatic summarization.

2. **Fact extraction** spreads out from entities and topics, connecting and contextualizing them in relationships, thus expressing facts.

   The simplest level of fact extraction is to determine the roles played by entities and the relationships among various entities. For example, that John Laing is the CEO at Inxight and he is also a board member for another company. This form of link creation (often called link analysis) quickly allows us to use facts in documents as a way to understanding the connections in the larger world.

   Topics, too, can be linked and contextualized in various ways. That this document is about one topic and another might be the key reason to select it. The fact that two topics appear more and more regularly together in documents might suggest a synthesis or emergence of an important new perspective.

   Another important class of facts is to identify significant events or event classes. For example, the announcement of a merger and acquisition, along with filling in the roles played by organizations as acquirer and acquired.

Extraction can provide a rich stream of additional "tags" that can be stored as metadata in a content management system or database along with a document to support a variety of purposes. Further, fact extraction can enable more specialized forms of querying or action triggering. Example applications include:

- **Previewing Documents.** Extracted entities and facts can provide clues to the usefulness of a particular document to a specific task when displayed as part of search results, directory entries, or other places where documents are referred to in user interfaces. By seeing *a little bit more but not too much*, the user is able to

efficiently decide whether it might be worth examining the whole document or complete their task using just the preview.

- **Content Packaging and Routing.** As an extension of document routing enabled by subject categorization, the richer set of tags obtained by extraction can enable more sophisticated forms of packaging and routing. For example, a publisher can slice and dice their content offering into new channels or product offering. Not only does this create opportunities for new revenue streams, but also for enhancing the experience of existing customers, increasing their loyalty and their revenue potential over time.

- **Surveillance.** New documents can be scanned for new events of interest. For example, a financial industry application might detect mergers and acquisitions or management personnel changes in new streams or Web sites. Other examples include noticing "suspicious" patterns of activity in intelligence information.

- **Question Answering.** Often specific questions we ask can be expressed as a query structured over roles, relationships, and entities. Combining search with fact extraction on result sets provides a way of answering many kinds of questions using a document collection, or at least of narrowing in the most relevant content.

Information Visualization

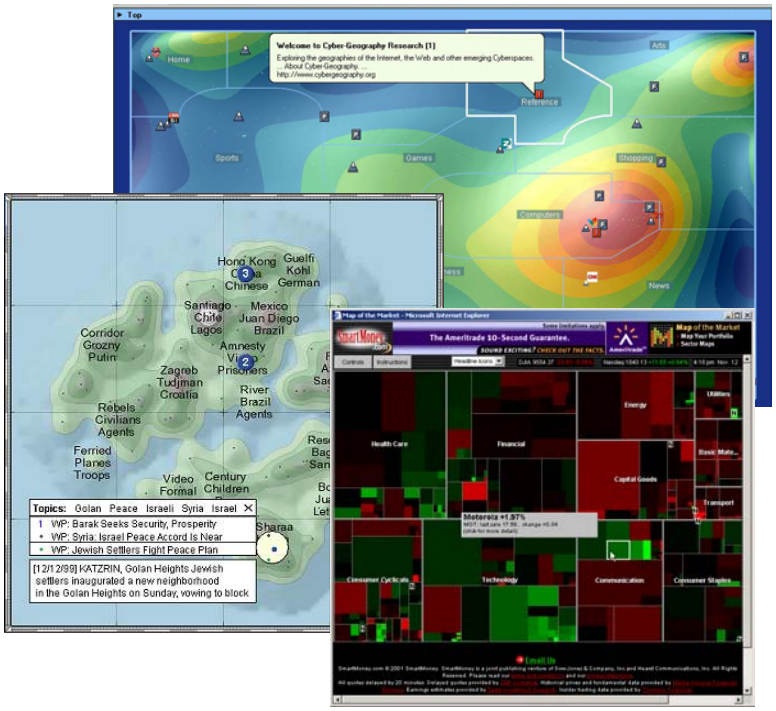
Information Visualization is about the use of visual techniques to increase the bandwidth of interaction with information. Though in many ways, visualization can be thought of as supercharged browsing, current Web browsing and directory navigation often require much more reading and mechanical effort than would be necessary with the use of visualization.

Visualization takes advantage of our evolved pre-attentive, automatic skills for processing large amounts of visual information, and for staying oriented in space and retaining spatial information. For example, we quickly spot a single red point among thousands. We can see small discontinuities or changes of texture or shade. Groups, parallel structure and shapes "pop out" at us. We are quite good at remembering that something was in the upper-left corner or was located about so high on the page. Taking advantage of these inbuilt skills is the design opportunity of visualizations.

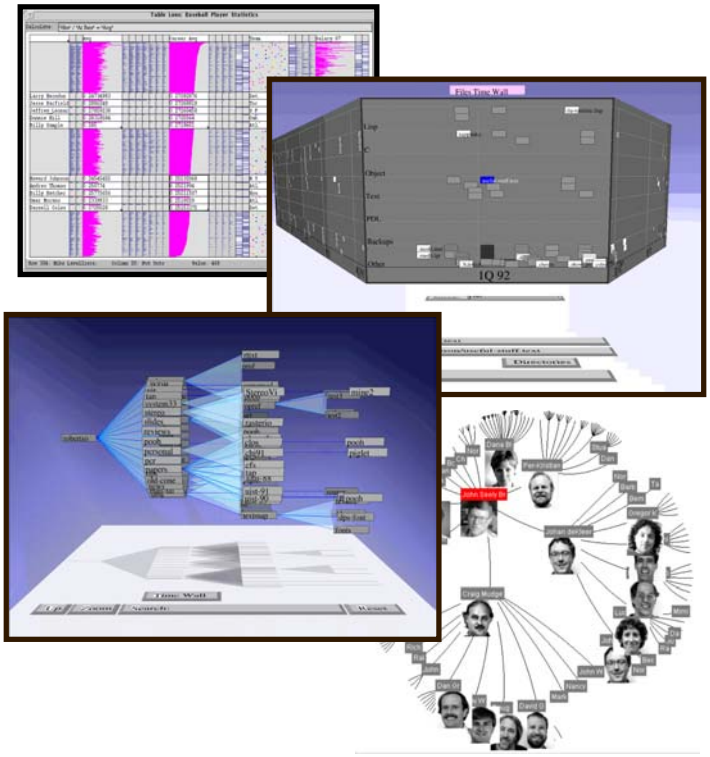
Effective visualizations function like maps. Generally, maps allow you to survey or assess the entire territory and also navigate to specific locations. You can understand the neighborhoods and terrain of San Francisco as well as plan a route to the Moscone center from your hotel.

As with maps, different visualization techniques might emphasize one task or another. Some techniques emphasize big picture thinking, while others might emphasize the detail and control necessary to navigate tightly or answer specific questions. Two kinds of visualizations are increasingly being applied to content collections:

1. **Content Terrain Maps** -- directly analogous to geographic terrain maps, representing sections and subsections of content is generated, typically using automatic analysis to assign locations and area based on the properties of documents in the collection, and to render glyphs for subsections or items in the collection. Various forms of interactions are supported with the map surface and with extra tools that might search or highlight sections or items on the map. A typical example is Smart Money's Map of the Market or Antarti.ca's maps.



2. **Wide Widgets** -- like user interface objects that make up the graphical user interface, but typically showing many more objects at a time and provide much more effective spatial management. They are typical based on showing an interactive, visual structure that mirrors some central spine in the information. A typical example is the Inxight Star Tree or TheBrain.



Visualizations support number of uses that tie search and browse together:

- **Collection Overview.** By showing a rendering of the entire collection or the several levels of taxonomy at once, a visualization can help a user get an overall sense of the collection and its structure much more quickly. The concreteness of an effective visual structure has much to offer as a mental map that can act as resource for interaction over time.

- **Rapid Navigation.** The visualization can tightly integrate navigation and interaction into the act of observation. For example, regular use of Map of the Market makes it quite easy to access the information of the most important companies really quickly, whether they are important today or important in a broader now. Similarly, Star Trees enables a user to rapidly move up and down a taxonomy several levels at a time as the user learns how the taxonomy is organized or sees visual features that represent some underlying properties of the categories.

- **Interpreting Items in Context.** "Spotlighting" search results on a graphical visualization, by marking matching items with salient visual features, is a powerful technique for helping users interpret search results. Such spotlights enable users to quickly see clusters of matches and thus focus on areas of the collection that are most relevant. The misses in the area might in fact be quite relevant. Similarly, isolated matches might either be quickly dismissed or possibly, because of their unusual context, embraced as the most interesting. Result marking in this way is a specific example of the more general power of visualizations to help users interpret local features in their context.

Conclusions

As a mind experiment, let's consider the bare bones of search. At indexing time, the content is analyzed and an index is built of the words in each document in the collection. Then at access time, the user types a query, which is matched against the index to return a list of documents. Viewing search stripped down so, it's easy to see its limitations and to imagine improvements. We can do more than catalog the words used in a document, and clearly we don't have to start with a query or leave the user unaided in digesting results whole. An analogous mind experiment or browsing, leads to similar observations.

Neither search or browse, nor both together, capture the richness of methods that a typical knowledge worker uses in interacting with documents in their physical workspaces. We are doing many different things as we access content. From understanding the big picture to scanning for specific answers, from forming a theory to populating a case with evidence, from collecting documents to read at the beginning of a project to finding a document we need at the end.

The technologies described here begin to throw off the limits as seen in our mind experiments, and to address the needs across this diverse range of information access tasks.

About the Author

Ramana Rao is the CTO and Founder of Inxight Software and the editor of the monthly email newsletter, Information Flow. More information can be found at www.inxight.com and www.ramanarao.com. You can reach him at rao@inxight.com.