Increasing the Information Density in Digital Library Results

Gio Wiederhold
26 May 2003

The problems addressed in this note focus on reducing the human overhead in obtaining information from Digital Library and general web resources, while retaining the valuable contents. It is intended to deal with a problem recognized many years ago by Herb Simon:

> "What information consumes is rather obvious; it consumes the attention of its recipients. Hence a wealth of information creates a poverty of attention, and a need to allocate that attention efficiently among the overabundance of information sources that might consume it."

However, allocation is not a simple task. Getting the right stuff to the right person has many aspects, and means supporting a variety of technologies, and understanding their benefits, costs and interactions. The total of available attention in the world may well be less than the total available information. We talk about billions of on-line webpages, and a hidden web that is yet larger. And yet, because so much potentially valuable information is lacking, many initiatives are funded to put more on the web. A crucial task is hence the reduction of available information to actionable information, i.e., the specific information that will cause a change in behavior, a reduction in further work, or the making of decisions.

Many technologies to filter information have been investigated in the past, and we list some of those, rapidly moving to harder and more speculative tasks. Several are in routine use

1. Ranking by document contents. Associated with ranking is the assumption that the consumer will only consider a few documents on the top of list.
2. Ranking by authority. Giving preference to documents published at a site that is valued in a context, for academic data that would be a journal versus a workshop report, for many other sources it would be a recent document.
3. Ranking by reference authority -- Google's page ranking algorithm extracts communal knowledge as evidenced by references given.
4. Elimination of redundancy. If similar documents are retrieved present either the latest one, or the one ranked higher by a suitable criterion.
5. Differences among documents: obtaining what is different between a known document and a new one. The task may be as simple as looking for additional material in a new version or as hard as requiring a deep analysis of both before a comparison on a higher level of abstraction.
6. Determining the novelty of a new document in respect to a given document collection. That task can be seen as a generalization of the high level of abstraction approach when comparing two documents.

7. Determining novelty individual with respect to an individual. If all the knowledge held by an individual can be captured then one could truly find out what material might be useful. Since there would be too much, domain emphasis is needed, and the unsolved (unsolvable?) problem of `common knowledge' should be avoided,
8. Abstraction of textual documents to retain essentials. There has been work on domain-independent abstraction selecting sentences that appear to represent the contents; better abstractions can be gained for domain specific texts, as pathology reports. An interesting task here would be automatic annotation of gene-sequences from relevant papers.
9. Abstraction of the contents of document collections is an obvious generalization. That task will require integration, and also semantic matching if the sources used autonomous ontologies.
10. A complementary source here is data-mining. I'd keep data-mining as such out of the scope of digital library initiatives, but linking data-mining results with information from textual sources would strengthen the users explanatory capabilities.
11. Reduction of textual information into a visual presentation is a step that is yet harder, and would require the competence of doing abstraction and the ability to place the result into some model. I visualize this task mainly for data that has a temporal or spatial aspect: Progress notes for a patient, description of an exploratory journey, or the progress of a scientific project.
12. Moving yet to a greater level of difficulty is populating an analytic model with such information. Here again domain specialization will be needed to demonstrate success. Having an analytic model will allow manipulation, not only to discern novelty, but also as a representation of normal behavior, if the domain can be well characterized. Domains that may lend themselves relatively conveniently to such a task are corporate finances from 10-K and similar documents. Harder would be domains as ecological processes, a global change. A challenge would be metabolic models, needed to formulate an understanding of food, drug, and environmental effects on organisms.
13. Having populated models will allow support of two further challenges. The first one is support for predictive tasks. Current information technologies, databases, data-mining, and digital libraries are seen as supporting decision-making, but fall short of providing the needed infrastructure. The decision maker today will copy the resulting information into a spreadsheet, and then add formulas to make extrapolations into the future for various scenarios of investment, probabilities of outcomes etc. It should be obvious that information systems should not terminate their support with the past, but be able to extrapolate the modules into the possible futures. And the outcomes of alternative futures should be readily comparable.
14. The second, yet harder challenge is the discovery of abnormal situations. That need is heavily emphasized today, when we try to use information systems to look for terrorists. Traditional search tasks, oriented to finding interesting, i.e., reasonable frequent relationships among data and, by abduction the processes that generate those data, can serve marketing folk, but not intelligence tasks. One can locate unusual or abnormal behavior by having a pre-existing model, perhaps one based on recent incidents, as the linking of flight-schools enrollments and behavior to terrorism. Finding unknown

abnormal linkages requires populating a large model with normal findings, since unknown abnormalities can only be identified if normality can be quantified. Unfortunately, such models will be large since observed data, say travel patterns, are the aggregate of activities in many domains, here business, holidays, and family visits and emergencies.

Many of these task may be handled semi-automatically, i.e., with human supervision, before full automation can be achieved. But semi-automatic systems should have the capability to learn from those interventions, so that the human load is reduced over time.

All these tasks can be expanded by adding adjectives as `distributed', `multi-media', or `ubiquitous', but those won't change he scientific import greatly.

In order to assess the cost/benefits of alternative technologies the setting has to be quantified. In some settings the cost of missing a source entry (Type 1 error) is high; in other settings the cost of having to reject irrelevant entries (Type 2 errors) is high. For instance, we the cost of missing a terrorist is indeed high, but many schemes now being considered fail because technologies that have a low rate of Type 1 errors are typically associated with a huge rate of Type 2 errors, so that even at a low cost per error rejection there may be no acceptable cost/benefit ratio.

To support web-based businesses, as envisaged in the semantic net initiatives, a very low rate of Type 2 errors, false hits will be needed. Businesses already today routinely pre-qualify suppliers that they will consider dealing with. The potential cost of getting the wrong stuff, getting it late, or obtaining the wrong information about stuff, is so much higher than the benefits of `getting a good deal', say getting some supplies at 5% less, are relatively negligible. Here certainly smaller is better.

Assignment of costs to these two types of errors also depends on ones background. Often senior people, having grown up in an information-poor setting will want to get all the information. It is often the generation in the trenches that realizes that there is too much to devote attention to.