Beyond Full-text Search: AI-Based Technology to Support the Knowledge Cycle

David M. Steier, Scott B. Huffman, Douglas I. Kalish
Price Waterhouse World Technology Centre
68 Willow Road
Menlo Park, CA

Abstract

From the mounds of raw information available electronically today, what professionals really need are targeted, timely nuggets of knowledge that can guide the solution to business problems. Today's common information tools –Web full-text search engines and the like – do not fully support this conversion of raw information into knowledge. In examining the common knowledge management problems faced by Price Waterhouse professionals, we have found that converting information to knowledge requires not only *finding* raw information, but also *filtering* through it for relevance, *formatting* it appropriately for the knowledge task at hand, and *forwarding* it to the right people. A fifth stage, *feedback* from the users, can allow the effectiveness of each stage to increase with time. In this paper, we describe each stage of this *knowledge cycle* and discuss the potential role that AI-based technology can play in its automation. We illustrate the possibilities through case studies of deployed knowledge management tools we have built at Price Waterhouse. These tools demonstrate that for targeted business tasks, AI-based technology can potentially facilitate much of the knowledge cycle, providing users with useful business knowledge that provides competitive advantage.

Introduction

In recent years there has been an explosion in the availability of electronic information. The World-Wide Web, newswire feeds, SEC filings and other corporate reports, government documents, litigation records, and much more are all available electronically and inexpensively. However, this treasure-trove of raw information has proven difficult to exploit. It is often difficult to find the information relevant for a particular task or decision. Even if relevant information can be found, it is often in the wrong form, requiring significant collation, reorganization, etc., to be useful. Information from different electronic sources must be combined, and this can require time-consuming conversion and normalization to make figures comparable and terminology consistent. Finally, although it is easier than ever to share information electronically (through email, electronic bulletin boards and databases, etc.), in a large organization it can be difficult to get new information into the hands of those who could use it best.

In our experience at Price Waterhouse, AI-based technology can play a key role in dealing with these difficulties in managing knowledge. In this paper, we introduce the concept of the *knowledge cycle* – the path from raw information to useful knowledge – and use it to highlight the key technological needs in knowledge management. Information tools that are typically used today, such as full-text search engines on the Web, are useful to a point, but support only the early stages of the knowledge cycle. To demonstrate what can be done for the other stages, we present case studies of knowledge management tools that we have developed and deployed for specific business tasks within Price Waterhouse. Our goal, however, is not so much to describe the specific techniques, algorithms, etc., used by these tools – they have been described elsewhere – but rather to illustrate how well-targeted AI-based technology can significantly impact knowledge management problems in large organizations. The paper concludes with several principles to inform the design of future applications that draw useful, targeted business knowledge of various kinds from large volumes of raw information.

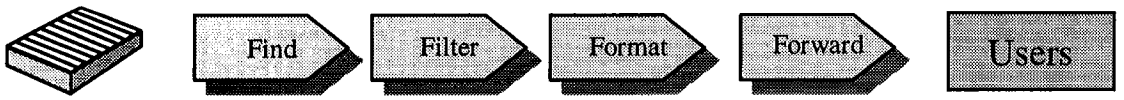

Figure 1: The Knowledge Cycle Value Chain

The Knowledge Cycle

Anyone who has had a full-text search tool give them thousands of "hits" in response to a query will agree that transforming raw information into knowledge involves much more than searching for a few words or phrases.[1] Rather, the transformation can be broken into a cycle of four general stages:

- **Find** sources and documents containing the needed raw information, in a timely fashion. This can involve general queries such as full-text searching over large document collections, or lookup in structured catalogs and directories that organize sources and documents into pre-determined useful categories.

- **Filter** the information from those sources and documents to extract only what is relevant to the knowledge task at hand. This can include applying more stringent relevancy tests to whole documents, to rank them, categorize them, etc. It can also involve filtering within each document to find and extract only those sections, sentences, etc., that contain the information needed. For textual documents this can include using natural-language processing techniques for *information extraction* [Hobbs, 1993].

- **Format** the filtered information for effective communication. This can include collating information across documents, "data cleansing" and normalization of information from multiple sources, and presenting the results appropriately through text formatting, summarization, use of graphs, charts, spreadsheets, multimedia, etc The appropriate use of formatting and charting allows users to identify important relationships within the information much more easily than they could from text alone [Larkin and Simon, 1987].

- **Forward** the formatted results to the person or group of people who can best use them. This involves determining who should receive the information and delivering it through various media – summaries in e-mail, personal databases, attached documents, fax, phone, pager, etc. Some researchers have approached this problem through systems that try to automatically produce a "profile" of each user's interests based on the documents they read (e.g., [Bloedorn *et al.*, 1996]), or by using key terms in documents users write in discussion databases [Krulwich and Burkey, 1996].

The Find/Filter/Format/Forward stages represent a general "value chain" in converting any information into knowledge, as shown in Figure 1. For a given information to knowledge transformation, the effectiveness of each stage and of the cycle as a whole can be evaluated by performance measures including:

- Time: How long did it take to get a question answered? Was the information timely enough?

- Completeness: Did the "knowledge user" get all, and only, the information needed?

- Accuracy: Was the knowledge provided correct, in the most useful form, to the right people?

- Cost: Was the knowledge created and delivered in the most cost-effective manner?

As an organization acquires experience in converting information to knowledge, it can function more effectively as future knowledge needs arise. Because the set of information sources and knowledge needs is diverse and constantly changes, it is impossible to anticipate all of the processing that will be required in the knowledge cycle for a given organization. A fifth stage, **feedback**, may provide the ability to adapt the first four stages to new circumstances. Feedback evaluates the performance of the previous stages in terms of performance measures such as

[1] In fact, experimental studies have shown that full-text search over large collections of documents is typically quite inaccurate. For instance, Harman [1993] indicates that full-text search techniques from information retrieval generally produce about 30 relevant documents out of the first 100 retrieved in tests using ad-hoc queries.

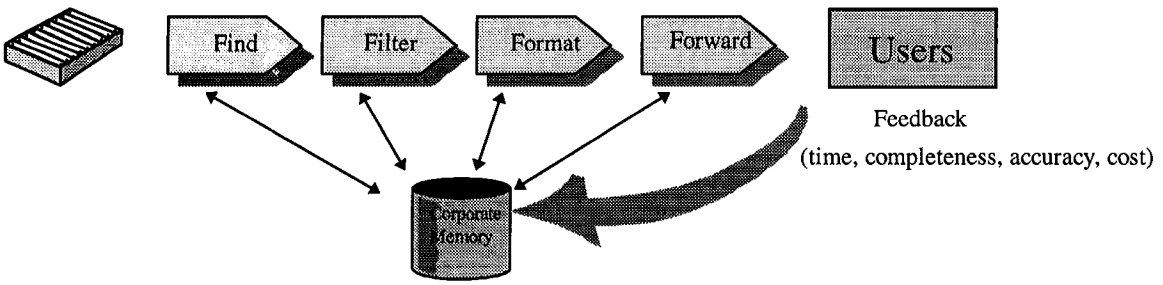

Figure 2 Continuous Improvement in the Knowledge Cycle

those mentioned above. Together, the five processes act on the corporate memory as shown in Figure 2.

To manage knowledge effectively, organizations must identify their highest value knowledge-based tasks, and for those tasks, they must identify and address the major bottlenecks in the knowledge cycle. Some of these bottlenecks will be organizational, and can be addressed by restructuring, retraining, corporate policy, etc. Other bottlenecks will be addressable by technology. Below, we discuss how technology can reduce bottlenecks in each stage of the knowledge cycle. We ground the discussion by presenting case studies of two applications we have developed and deployed at Price Waterhouse that automate the knowledge cycle for specific business knowledge needs.

Technology Enablers for the Knowledge Cycle

What technology will impact the knowledge cycle in a large organization like Price Waterhouse? We will begin with some very general answers, and then examine the case studies. General technology issues for each stage include:

- **Find**: PW professionals, like those in other large organizations, use information from a wide variety of sources. We have hundreds of Lotus Notes databases replicated throughout the firm, thousands of local documents, and external sources like the Web, CD-ROMs, and newswires. We need powerful, easy to use search capabilities that can search across this variety of source types.

- **Filter**: Given the growing volume and diversity of information sources, PW professionals have time to review an increasingly small fraction of that information. Filtering technology is needed that can process documents for a variety of stringent relevance conditions, with high accuracy. Conditions of relevance important to our organization include financial criteria in financial statements and their footnotes, corporate information in a variety of countries and industries, and business events of particular types – e.g., mergers, management changes, new products, legislation – in companies, technologies, and markets that impact our clients.

- **Format**: A large stack of documents is often the least useful form in which information can be delivered. Rather, the business needs that PW professionals address require knowledge in a variety of forms. Knowledge cycle technologies for formatting should support target formats such as graphs, presentations, spreadsheets, rich-text documents, hyper-linked document collections, mail messages, and even multimedia documents with interactive visualizations of complex material.

- **Forward**: Business opportunities can be lost if relevant information within an organization is not forwarded to the right people in time. In a large organization like PW, it is impossible for any individual to be aware of all of the others in the organization for whom a piece of information could be relevant. Forwarding technology could address this need by automatically notifying people of possibly relevant information, based on personal profiles that are generated either manually or automatically.

- **Feedback**: As we address new needs for knowledge, we encounter opportunities for improving the efficiency of the knowledge cycle when similar needs arise in the future. Utilizing feedback automatically could allow changes to knowledge cycle technology with a minimum of programming burden.

Because of the diversity of knowledge cycle tasks, these general issues are hard to grasp outside of the context of specific applications. Next, we will describe two sets of these applications as case studies to ground our discussion of technology to support the knowledge cycle.

Case study #1: EDGAR data and financial benchmarking

Price Waterhouse, like many financial and consulting organizations, makes heavy use of the various types of information contained in the corporate SEC Filings of U.S. public companies. In the past, Price Waterhouse has spent large amounts obtaining filings in paper form from third-party data providers, and thousands of hours of staff time searching through these filings, rekeying, analyzing and formatting financial statement information and other portions of the text into reports, spreadsheets and charts. The information is used for a number of purposes, primary among them the analysis of company financials, and the benchmarking of one company's financials against other comparable companies within their industry. In recent years the SEC has made corporate filings available electronically through their EDGAR program. Because of PW's heavy use of this information and the large cost of performing financial benchmarking manually, this was a good target for the application of technology to the electronic filings.

For SEC filings, the knowledge cycle takes the following form:

- ***Raw information input:*** SEC filings; primarily 10-K and 10-Q filings (annual and quarterly reports). These filings are in raw ASCII form; a typical filing has over 100 pages of text. Tables, sections, and footnotes within the text are not formatted or labeled in any regular fashion across the filings – and are sometimes labeled inconsistently even within a single filing. Although there are some SGML tags specified by the SEC for indicating the locations of tables and other items, they are used inconsistently within different filings (so much so that our automated systems simply remove them before processing).

- ***Knowledge output:*** For individual companies, our users need financial tables and footnotes within a spreadsheet; and full filings and/or sections of filings (such as the Management Discussion and Analysis section), either as raw text or nicely formatted for printing in a rich-text format (e.g. MS-Word format). For benchmarking multiple companies, our users need easily-generated spreadsheets and graphs that compare companies across a wide variety of financial measures. The data in these spreadsheets and graphs must be correctly normalized for different scaling factors (e.g. reports in thousands of dollars vs. those in millions), different line labels within the original financial tables (e.g. "revenues" vs. "Net sales"), etc. Advanced features that users have requested include the ability to automatically find the financial measures for which a company significantly differs from the other companies being compared, and the ability to automatically generate output in the form of a presentation (e.g. as a set of MS-PowerPoint™ slides).

- **Find:** Users need the ability to find companies and which filings are available for them. This can be either by simple search for the company name, by SIC code, or by standard measures of company size such as total assets or revenues. For benchmarking, users need to find sets of related companies.

- **Filter:** For individual companies, filtering technology must extract section boundaries, important financial tables, and table footnotes from the ASCII text of 10-K and 10-Q filings. Of these, extracting tables and their footnotes is the most difficult. Tables must be found in running text by analyzing whitespace patterns, parsing potential table titles and line items, etc. For multiple company benchmarking, the filtering task also includes *normalization* (conversion to a uniform, directly comparable form) of the extracted financial items across companies. This includes normalization of line labels, scaling factors (thousands vs. millions), cross-checking what is extracted for each company within and across different financial tables, etc.

- **Format:** Users want the text of either whole filings or filing sections, formatted nicely in a rich-text format. For financial tables and their footnotes, users want a spreadsheet format compatible with programs like Excel. For benchmarking, users want the ability to generate attractive graphs, charts, and presentations comparing companies along various financial measures.

- **Forward:** Once financial information is extracted and normalized, it is possible to monitor filings as they come in for particular sets of financial conditions. For example, one PW group needs the ability to monitor companies within certain industries and size parameters for a set of conditions that may indicate financial distress. Reports of companies meeting the conditions are automatically forwarded to a special database.

We have built specialized technology for each of the Find/Filter/Format/Forward stages for this knowledge task, focusing in particular on the technology needed to find and interpret financial tables. We observed that keyword-based processing (for instance looking for phrases such as "total revenues") was not powerful enough alone for substantial automated analysis, but was much more effective when used in combination with expectations of the structure of SEC forms and of financial statements. In particular we found that incorporating knowledge of expected arithmetic relationships between line items, both

within and across tables, allows for very precise interpretation of financial statements. We call this technique *constraint-driven table parsing.*

The technology that parses EDGAR filings has been deployed as the basis of several applications within PW. The EDGAR filings database is a Lotus Notes database containing an index to the over 20,000 filing entities so that people can find companies and request filings, portions of filings, or benchmarking information. This database has thousands of users within Price Waterhouse and processes over 200 requests a day. Several other tools provide benchmarking capabilities. EdgarScan (accessible on the Web as http://edgarscan.tc.pw.com) offers the user the ability to access filing sections, analyzed financial statements and their footnotes, financial ratios, and some rudimentary charting abilities. A Windows-based application called Benchmarking Assistant™ provides our practice with the ability to perform more sophisticated benchmarking based on data from EDGAR and CompuStat®. Another specialized benchmarking capability allows members of PW's Tax practice to request a detailed comparison of companies' tax reconciliation tables, found in the income tax footnotes of their annual Form 10-K filings. The results are returned in the form of a spreadsheet, showing the components of the tax reconciliation, such as state taxes, normalized to a percentage for easy comparison, and a hyperlinked file showing the tax footnotes as they appear in the 10-K. With all these applications, automating search, extraction, normalization, and formatting of SEC documents has given PW a substantial gain in efficiency, and consequent reduction in "time-to-market," for new knowledge drawn from financial statements.

Case study #2: Management changes extracted from newswires

In addition to corporate financials, it is important for PW professionals to track other business events as they occur. One type of event that is particularly important for PW is executive management changes at large companies. These are reported in press reports and newspaper articles, available electronically via newswire feeds.

Due to the huge volume of newswire articles each day (one service PW uses provides over 5000 articles daily), it would be extremely difficult to track management changes manually. Conventional technology, such as full text search, can reduce the number of articles but cannot pinpoint those that definitely contain management changes, and does not extract relevant information such as the company and person involved. Third-party "clipping services" can find the relevant articles, but often not in a timely manner, and not with relevant information extracted and searchable in an electronic form. Therefore, we decided to build specialized technology for tracking business events like management changes.

For this task, the knowledge cycle takes the following form:

- ***Raw information input:*** Newswire articles in ASCII text, typically a few paragraphs each. Our newswire service provides a keyword filtering capability; filtering for keywords that must appear in management change articles reduces the input volume from 5000 down to about 1000 articles per day.

- ***Knowledge output:*** Management change reports, organized by date, company, person, and new management position. For PW's use, it is important not only to find the articles reporting management changes, but to find them in a timely manner (e.g. the day they are first reported in the press); to extract specifically the relevant information (e.g. the company, person, and position involved), as opposed to just producing a stack of articles; and to organize what is found by company, by category using lists of companies important for various PW uses, by industry groupings, by geographic region, and even by cross-reference to information in other sources, such as EDGAR filings.

- **Find:** Users need to be able to quickly find the recent management changes of interest to them. Typically, this means either changes at one of a specific set of companies, or changes at companies in a particular industry group and/or geographic region. Our application, therefore, organizes management changes as a Lotus Notes database with a variety of "views" that list changes in the various categories of interest. Although there are typically 75 to 100 new management changes found per day total, for a particular user interested in an industry group or region, there are typically only a dozen or so changes per week.

- **Filter:** Filtering is the primary part of the management changes task. Given input of 1000 or so articles each night, an information extraction system called ODIE (for *On-Demand Information Extractor*) scans each article for management changes. For each found, it extracts the company, person, and new position reported in the text.

- **Format:** Each management change report is formatted into a Lotus Notes document in a database devoted to them. These documents are then cross-referenced and classified using other data sources, such as company lists that include industry group and geographic

information. This classification process allows us to build the various views described in "find" above. The classification is performed using heuristic matching techniques to cross-reference management changes with entries in other data sources [Huffman and Steier, 1995].

- **Forward:** We have experimented with the capability of automatically emailing users reports of management changes at companies of interest to them. However, for this application we have found that our users prefer to monitor the management changes database themselves, without automatic forwarding, using the various views that the database provides.

The ODIE extraction system [Huffman, 1996] has been the key to automating the management changes knowledge cycle. ODIE exploits the fact that the language used in business news articles is stylized, and uses a relatively small number of syntactic patterns to express most instances of particular types of business events. The system performs a shallow, efficient linguistic analysis of newswire texts, to find specifically those syntactic patterns that indicate a business event such as a management change. Extraction patterns are represented as paths through a non-deterministic finite state machine; embedded finite state machines are used to recognize syntactic relationships. The overall technique has some similarities to SRI's Fastus [Hobbs et al., 1992] and UMass's CIRCUS [Lehnert et al., 1993] extraction systems. ODIE has been applied to filter both European and US newswires. In addition to management changes, we have experimented with filtering for other types of events, such as corporate acquisitions.

An important bottleneck in applying extraction systems like ODIE to the extraction of new events is determining the set of extraction patterns that indicate the event's presence in the newswire texts. As a step towards overcoming this difficulty, we have produced a **feedback** system for ODIE. The feedback system, called LIEP, allows a user to input examples of texts paired with the events that should be extracted from them [Huffman, 1996]. LIEP analyzes each input text and uses a combination of analytical and inductive machine learning techniques to induce patterns that will extract the indicated events. Thus, ODIE can be "trained" to extract new kinds of events without further programming effort. We used this feedback method, for instance, to train ODIE to extract simple corporate acquisitions. Related systems for learning text extraction patterns include AutoSlog [Riloff, 1993], AutoSlog-TS [Riloff, 1996], PALKA [Kim & Moldovan, 1995], and CRYSTAL [Soderland et al., 1995].

Discussion

To date we have focused primarily on automated assistance for filtering and formatting, and less for the other portions of the knowledge cycle; this balance will shift in future work. We are particularly optimistic, based on our experience with the management changes task, about using Feedback to shorten the technology development life cycle. We plan to explore the potential of Feedback in developing applications that track events such as new legislation, new product announcements or industry developments, etc.

Generalizing over these applications, the following principles emerge that provide guidance for future development of knowledge management tools:

- **Exploit task constraints in developing knowledge management technology**: General tools such as full-text search may provide a starting point for finding information, but they need to be augmented to create usable knowledge efficiently. Optimal processing for the knowledge cycle must take advantage of regularities present in data sources, whether in document format, in language used, or in arithmetic relationships between items in tables. From the perspective of the end-user, interfaces that are oriented towards a particular business problem, such as financial benchmarking or client monitoring, are more convenient than those in generic systems.

- **Target high-value knowledge:** Developing technology to automate bottlenecks in the knowledge cycle is costly. Therefore, it should only be considered for knowledge that is of high value to your organization. Good metrics include how many people use the type of knowledge within the organization (or would use it if it were freely available); how much value the knowledge adds (or would add) to the tasks those users perform; and how much is currently spent to produce the knowledge manually. EDGAR filings and management changes passed these metrics within our organization, so that automating them was cost-effective. For more generic, lower-value information tasks, it may be more cost-effective to utilize manual processing or third-party tools.

- **Give users a way to "drill-through" to the source**: Some guesswork is involved in transforming all but the most highly-structured information sources into usable knowledge. Even if the transformation can be automated, there will be times when users want to see where an extracted data item came from, either to verify the extraction or to get more background. For example, for the benchmarking application, a significant issue is ensuring comparability of numbers drawn from different contexts (industries, accounting

policies, etc.). With hyperlinks from an number to the original source material, users can form their own judgments of comparability.

These principles, together with the knowledge cycle introduced in this paper, provides a framework for using technology to remove the bottlenecks in efficient knowledge creation and use. Our case studies have demonstrated the importance of considering technological support for all stages of the knowledge cycle. Efficiencies at each stage multiply out to reduce the amount of user effort by several orders of magnitude. For instance, the management changes application reduces the effort from reading five thousand articles received on a daily newswire to only looking at a few management change events with the right information extracted and categorized. The reduction in effort translates to an earlier time to market with new knowledge. Similarly, people within Price Waterhouse can get the financial data extracted from EDGAR filings weeks or even months before it is available from on-line services. Granted, the extensive task analysis and support required for knowledge cycle automation can be arduous and is only worth undertaking for very high-value applications. For wisely chosen applications, however, organizations that invest in appropriate AI technologies for knowledge management – parsing, information extraction, intelligent search, user profiling & document forwarding, and the like – will receive major returns.

References

[Bloedorn *et al.*, 1996]. Bloedorn, Eric, Mani, Inderjeet, and MacMillan, T. Richard, "Machine learning of user profiles: Representational issues," in *Proceeding of the Thirteenth National Conference on Artificial Intelligence (AAAI-96),* AAAI Press, 1996.

[Harman, 1993]. Harman, Donna, "Overview of the first Text Retrieval Conference," in *Proceedings of the sixteenth annual ACM conference on research and development in Information Retrieval*, pp. 36-47. Association for Computing Machinery, 1993.

[Hobbs et al., 1992]. Hobbs, J. R., Appelt, D. E., Bear, J. S., Israel, D. J., and Mabry Tyson, W. "FASTUS: A system for extraction information from natural-language text." Tech. Report No. 519, SRI International, 1992.

[Hobbs, 1993]. Hobbs, J. R. "The generic information extraction system," in *Proceedings of the Fifth Message Understanding Conference (MUC-5). Morgan Kaufmann,* San Mateo, CA, 1993.

[Huffman, 1996]. Huffman, Scott B. "Learning information extraction patterns from examples", in *Connectionist, Statistical, and Symbolic Approaches to Learning for Natural Language Processing*, ed. S. Wermter, E. Riloff, and G. Scheler, pp. 246-60. Springer-Verlag, Berlin, 1996.

[Huffman and Steier, 1995]. Huffman, Scott B. and Steier, David, "Heuristic joins to integrate structured heterogeneous data," in *Working Notes of the AAAI Spring Symposium on Information Gathering in Heterogeneous Distributed Environments*. American Association for Artificial Intelligence, 1995.

[Kim and Moldovan, 1995]. Kim, Jun-Tae and Moldovan, Dan I. "Acquisition of linguistic patterns for knowledge-based information extraction." *IEEE Transactions on Knowledge and Data Engineering,* 7(5):713-24, 1995.

[Krulwich and Burkey, 1996]. Krulwich, Bruce, and Burkey, Chad, "The ContactFinder agent: Answering bulletin board questions with referrals" in *Proceeding of the Thirteenth National Conference on Artificial Intelligence (AAAI-96),* AAAI Press, 1996.

[Larkin and Simon, 1987]. Larkin, Jill H. and Simon, Herbert A., "Why a diagram is (sometimes) worth ten thousand words." *Cognitive Science, 11(65-99)*, 1987.

[Riloff, 1993]. Riloff, Ellen. "Automatically constructing a dictionary for information extraction tasks" in *Proceedings of the Eleventh National Conference on Artificial Intelligence (AAAI-93),* pp. 811-16, 1993.

[Riloff, 1996]. Riloff, Ellen. "Automatically generating extraction patterns from untagged text" in *Proceedings of the Thirteenth National Conference on Artificial Intelligence (AAAI-96),* AAAI Press, pp. 1044-49, 1996.

[Soderland et al., 1995]. Soderland, S., Fisher, D., Aseltine, J., and Lehnert, W. "CRYSTAL: Inducing a conceptual dictionary" in *Proceedings of the 14th International Joint Conference on Artificial Intelligence (IJCAI-95),* pp. 1314-9. Morgan Kaufmann, 1995.