From: IAAI-92 Proceedings. Copyright © 1992, AAAI (www.aaai.org). All rights reserved.

Making Sense of Gigabytes: A System for Knowledge-Based Market Analysis

Tej Anand and Gary Kahn, A. C. Nielsen Company

Market researchers are looking more and more to point-of-sale scanners in retail outlets as a rich source of high-quality and timely sales data. Unfortunately, the sheer volume of such data makes analysis and interpretation an overwhelming task. Consequently, there is a demand for software tools that can automatically provide an analysis of large volumes of data. SPOTLIGHT is a knowledge-based product that enables A. C. Nielsen clients to understand what is significant in databases of point-of-sale scanner data. Using SPOTLIGHT, manufacturers and retailers track the sale and movement of products, assess the effectiveness of various promotional strategies, and compare the performance of competing products and product segments.

SPOTLIGHT provides an alternative to the classical approach, where market analysts provide custom interpretations and reports, typically using spreadsheet tools such as Lotus as aids. SPOTLIGHT turns a task that takes 2 to 4 weeks into a task of 15 minutes to several hours depending on data volumes. Although previous attempts were made to automate the analysis of market data, within this domain, SPOTLIGHT represents the first commercial use of an expert system shell and the first deployment of sophisticated analytic capabilities directly onto a large number of widely distributed personal computer (PC) platforms.

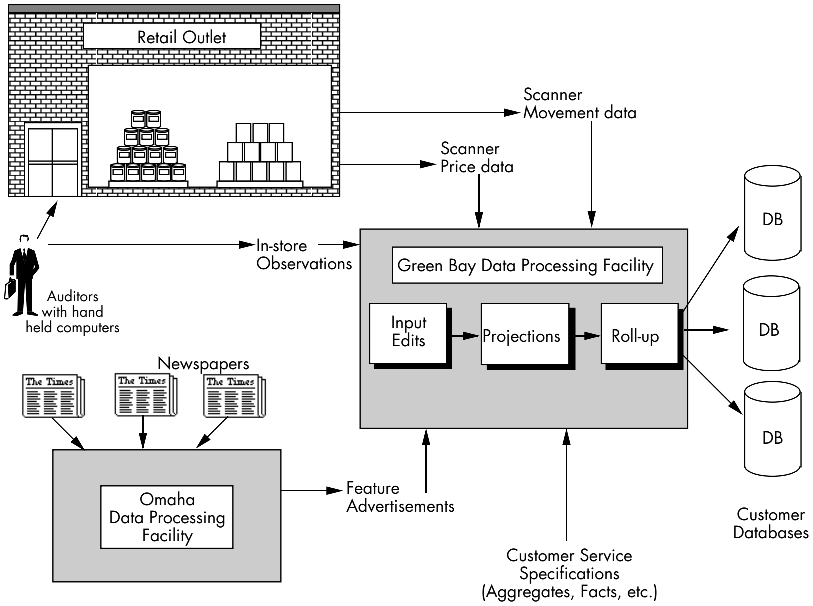


Figure 1. The Information Factory.

Task Description

Over 1 billion dollars a year is spent by consumer-packaged goods companies to understand why their products are succeeding or failing in the marketplace. Proctor and Gamble, Kraft/General Foods, and others devote significant resources to purchasing and analyzing market data pertaining to thousands of products sold across hundreds of retail organizations to millions of consumers. A. C. Nielsen addresses the needs of such companies by providing custom and syndicated data services.

Among the most significant data collected are universal product code (UPC) volumes and prices taken from cash register scanning systems in a representative sample of grocery stores. In addition, data are collected by field auditors visiting individual stores to record in-store promotional conditions such as displays, discounts, shelf placements, and coupons. Data about market conditions, such as retailer advertising in newspapers and flyers, are also collected.

Sample data are projected to a market-level assessment using statistical techniques and are made available through online databases typically organized by four dimensions: product, market, period, and measure.

The lowest level of granularity at the product dimension is UPC; at the market dimension, an individual store; and at the period dimension, a week. The measure dimension consists of facts such as price and distribution. Nielsen organizes its product databases into categories such as coffee and carbonated beverages. Figure 1 provides an overview of the flow of data from the retail outlet through Nielsen's information factory.

Nielsen databases are customized according to its clients and contain aggregated facts over the product and market dimension based on knowledge supplied by the client. Examples of aggregates over the product dimension are brand totals and totals over product characteristics, such as total caffeinated coffee. Examples of aggregates over the market dimension are totals for user-defined geographic boundaries. These boundaries are usually arranged hierarchically, for example, total United States, northeastern region, New York, and Buffalo. A user can define custom aggregates over the product and market dimensions and also request facts to be aggregated for any number of weeks.

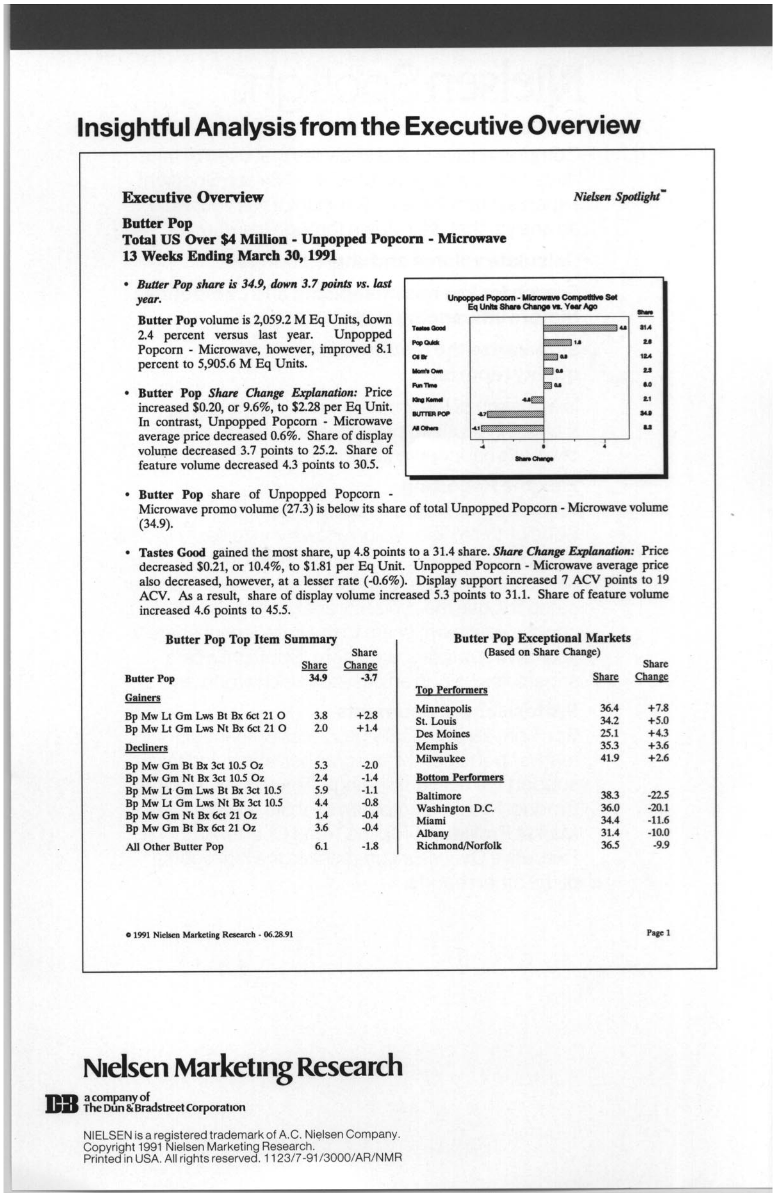
The databases described here have over 100 factual measures for thousands of products across hundreds of geographic areas for at least 125 weeks. These are large databases with millions of records. Nielsen has over a terabyte of online data. Some clients regularly review several gigabytes of data, some megabytes.

Extracting meaningful information for decision-making purposes is a difficult task. As scanners make the collection of high-quality data possible, the analysis and interpretation of the resulting volumes of data becomes an overwhelming task. Consequently, there is a growing demand for software tools that aid in interpreting marketing data.

Application Description

SPOTLIGHT is a knowledge-based product that enables A. C. Nielsen clients to understand what is significant in large databases of point-of-sale scanner data. SPOTLIGHT recognizes significant shifts among product segments (share of caffeinated coffee decreased while share of decaffeinated coffee increased), the impact of promotional programs, and changes in distribution and price that lead to a reportable shift in market share or volume. SPOTLIGHT also tries to find patterns of common behavior across a set of competing products.

SPOTLIGHT processes large amounts of data into five brief, clearly understandable reports. These reports allow manufacturers and retailers to track the sale and movement of their products, assess the effectiveness of promotional strategies, and compare the performance of competing products and product segments.



Insightful Analysis from the Executive Overview

Executive Overview *Nielsen Spotlight*™

Butter Pop
Total US Over $4 Million - Unpopped Popcorn - Microwave
13 Weeks Ending March 30, 1991

- ***Butter Pop share is 34.9, down 3.7 points vs. last year.***

  Butter Pop volume is 2,059.2 M Eq Units, down 2.4 percent versus last year. Unpopped Popcorn - Microwave, however, improved 8.1 percent to 5,905.6 M Eq Units.

- **Butter Pop** ***Share Change Explanation:*** Price increased $0.20, or 9.6%, to $2.28 per Eq Unit. In contrast, Unpopped Popcorn - Microwave average price decreased 0.6%. Share of display volume decreased 3.7 points to 25.2. Share of feature volume decreased 4.3 points to 30.5.

- **Butter Pop** share of Unpopped Popcorn - Microwave promo volume (27.3) is below its share of total Unpopped Popcorn - Microwave volume (34.9).

- **Tastes Good** gained the most share, up 4.8 points to a 31.4 share. ***Share Change Explanation:*** Price decreased $0.21, or 10.4%, to $1.81 per Eq Unit. Unpopped Popcorn - Microwave average price also decreased, however, at a lesser rate (-0.6%). Display support increased 7 ACV points to 19 ACV. As a result, share of display volume increased 5.3 points to 31.1. Share of feature volume increased 4.6 points to 45.5.

Butter Pop Top Item Summary

	Share	Share Change
Butter Pop	**34.9**	**-3.7**
Gainers		
Bp Mw Lt Gm Lws Bt Bx 6ct 21 O	3.8	+2.8
Bp Mw Lt Gm Lws Nt Bx 6ct 21 O	2.0	+1.4
Decliners		
Bp Mw Gm Bt Bx 3ct 10.5 Oz	5.3	-2.0
Bp Mw Gm Nt Bx 3ct 10.5 Oz	2.4	-1.4
Bp Mw Lt Gm Lws Bt Bx 3ct 10.5	5.9	-1.1
Bp Mw Lt Gm Lws Nt Bx 3ct 10.5	4.4	-0.8
Bp Mw Gm Nt Bx 6ct 21 Oz	1.4	-0.4
Bp Mw Gm Bt Bx 6ct 21 Oz	3.6	-0.4
All Other Butter Pop	6.1	-1.8

Butter Pop Exceptional Markets
(Based on Share Change)

	Share	Share Change
Top Performers		
Minneapolis	36.4	+7.8
St. Louis	34.2	+5.0
Des Moines	25.1	+4.3
Memphis	35.3	+3.6
Milwaukee	41.9	+2.6
Bottom Performers		
Baltimore	38.3	-22.5
Washington D.C.	36.0	-20.1
Miami	34.4	-11.6
Albany	31.4	-10.0
Richmond/Norfolk	36.5	-9.9

© 1991 Nielsen Marketing Research - 06.28.91

Page 1

Nielsen Marketing Research

a company of
The Dun & Bradstreet Corporation

NIELSEN is a registered trademark of A.C. Nielsen Company.
Copyright 1991 Nielsen Marketing Research.
Printed in USA. All rights reserved. 1123/7-91/3000/AR/NMR

Figure 2. Example SPOTLIGHT Report Photo.

Each report is one to two pages and has a specific role to play in supporting the analysis of product behavior:

Executive overview: This report provides key information that is further detailed in subsequent reports.

Product profile: This report summarizes sales and merchandising information for a selected product and category and identifies significant segment shifts.

Competitive profile: This report summarizes and explains sales and merchandising performance of competitive products.

Exceptional market profile: This report summarizes events in the most volatile markets and identifies events that might explain volatility across markets.

Product trend chart: This report encapsulates a factual summary of merchandising volume, distribution, price, and promotional events.

The first four reports are output as composite documents with text, tables, and business graphics (see, for example, figure 2). The fifth report is a complex table.

SPOTLIGHT Characteristics

SPOTLIGHT uses expert heuristics to determine possible causes for the sales of a product. An example of a heuristic is, If the share of a product has increased over a time period and its distribution has also increased during that period, then the increase in share can be attributed to the increase in distribution. All these heuristics are parameterized such that they can be customized by the user for a particular category. Default parameters are built into the system for categories where customization is not necessary. Where customization is desired, the user can control SPOTLIGHT's selection criteria for key competitors, key product segments, and volatile markets.

Users of SPOTLIGHT have different needs based on their role within marketing or sales. SPOTLIGHT accommodates different analytic needs by providing the user with the flexibility to select the granularity of analysis along the product and market dimensions. The product targeted for explanation might be a unique product UPC, but it could also be a brand grouping or all of a manufacturer's products, that is, many brands, each containing many UPC items. Similarly, the user can select the target geography to be the total United States, a particular region, or an individual city market. In addition, SPOTLIGHT has limited interactive capabilities that allow the user to start at one level of granularity and then drill down to the next, as needed, for example, going from a brand to its UPCs to further isolate reasons for poor performance.

Differences in information and presentation needs are addressed by providing a selection of reports and display formats. Reports produced by SPOTLIGHT can be used to generate status reports for the performance of various products, evaluate the impact of promotional programs, or diagnose the performance of products. The graphs and tables incorporated into SPOTLIGHT reports can be generated individually in enlarged formats to assist the user in making effective presentations.

SPOTLIGHT Innovations

SPOTLIGHT provides an alternative to the classical approach where market analysts provide custom interpretations and reports, typically using spreadsheet tools such as Lotus as aids. SPOTLIGHT turns a task that can take 2 to 4 weeks into a task of 15 minutes to several hours depending on data volumes. Although previous attempts have been made to automate the analysis of market data, within this domain, SPOTLIGHT represents the first commercial use of an expert system shell and the first deployment of sophisticated analytic capabilities directly onto a large number of widely distributed personal computer (PC) platforms.

By using a modular rule-based architecture, SPOTLIGHT provides more functions and higher-quality reports and achieves greater maintainability than previous efforts to develop similar products.

SPOTLIGHT is innovative as well in its tight integration with third-party software, distributed architecture, and approaches to product development, as discussed below.

SPOTLIGHT System Description

The SPOTLIGHT system architecture is designed to access large mainframe databases, provide an untethered report-generation capability, take advantage of a low-cost central processing unit that processes millions of instructions per second, and minimize mainframe connect time. SPOTLIGHT achieves these objectives by downloading a filtered set of data to PC for extended analysis and report generation. Filtered data sets typically measure about one one-hundredth of the original data file. Once they are downloaded, a variety of SPOTLIGHT reports can be run untethered from the mainframe.

The PC component of SPOTLIGHT is delivered on 286- and 386-based PCs and PC compatibles. The system runs under conventional memory; that is, extended memory is not needed. After the downloading of data from the mainframe, the system completes the analysis and output generation phases in approximately one minute on a 386-based 33-MHz PC.

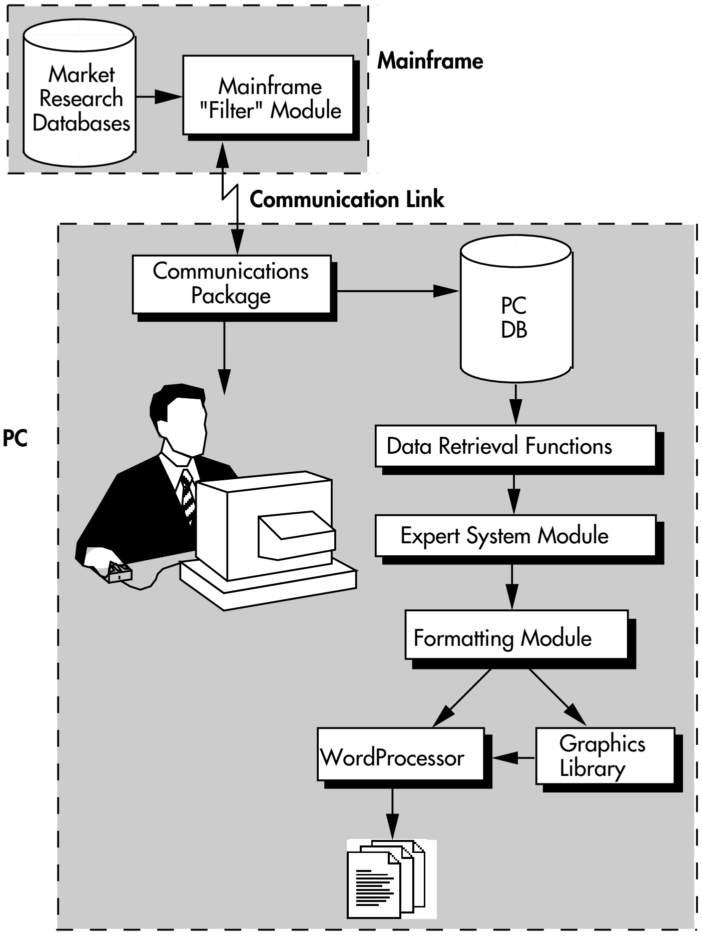


Figure 3. System Overview.

System Overview

The major components of the SPOTLIGHT system are shown in figure 3.

Mainframe Filter Module: This module extracts data about the products and the key competitors in the target geographies. It can automatically determine key competitors and exceptional geographies. Important product segments and aggregates are also identified. The extracted data are stored in a binary indexed file and downloaded to PC.

Communication: SIMPC, a commercial communications package, provides the cooperation between the mainframe and PC. Both messages and files are transferred.

User Interface: The user interface in SPOTLIGHT is mouse driven, with the user making selections from pick lists. The user is guided through the selection of a target product and a target geography by the system. All other selections are made at the user's initiative. The user can access and change the default selections in a SPOTLIGHT analysis by selecting menu items.

Data-Retrieval Functions: These functions are a library of access functions that allow the expert system to query the database that is downloaded from the mainframe to the PC disk. The expert system module queries for data, as needed, depending on dynamic needs.

Expert System Module: This module incorporates heuristic rules for analyzing the data extracted from the mainframe, selecting among graphs to present these data, and generating text to describe results of the analysis. This module is implemented with a rule-based expert system tool. The role and design of this module is described further in the next subsection. Additional detail is provided in Anand and Kahn (1992).

Formatting Module: This module generates the final SPOTLIGHT output. Currently, this module uses CHARTMAN, a proprietary Nielsen graphics package, to generate the graphs and WORD PERFECT to generate the compound document. This module makes all the composition and layout decisions. Integration with third-party word processing and graphics software allows users to customize the reports using editors that they are familiar with.

Expert System Module Discussion

The functional breakdown of the expert system module is shown in figure 4. The architecture of SPOTLIGHT is similar to applications that use the PENMAN system (Mann 1983; Springer, Buta, and Wolf 1991): An expert system is supplemented by text-generation and control utilities. In addition, SPOTLIGHT is modular with respect to the analytic and output-generation components. As a result, the user can configure reports with different contents without changing the underlying approach to analysis.

The result of a SPOTLIGHT analysis is a series of reports consisting of text, tables, and graphs. Within the expert system module, the contents of the reports are represented as instantiations of bullet objects, paragraph objects, graph objects, and table objects. Each of these objects has a *condition* a*ttribute,* which specifies when the object is relevant, and a *parts attribute,* which defines the contents of the object; for example,

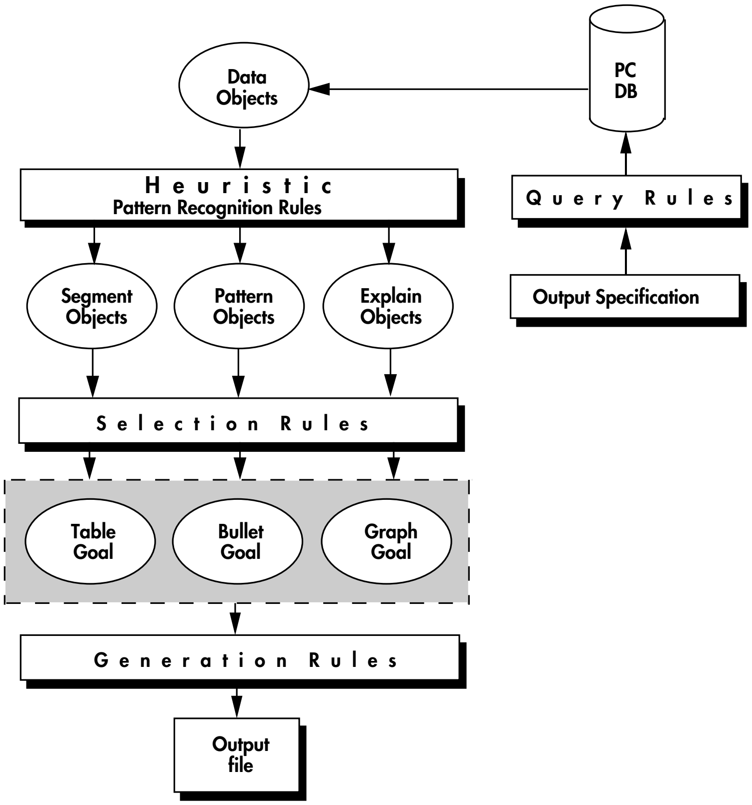


Figure 4. Expert System Module.

the value of the parts attribute of a particular bullet instance is one or more sentence types.

A report is defined within SPOTLIGHT by specifying instances of bullets, graphs, and tables, including values for the condition and parts attributes of each of these instances. These instances are loaded into a working memory agenda. Query rules take the instance on the top of the agenda, determine the data needs of this instance, and formulate queries to the database. Once the data are retrieved, the system enters the analysis phase.

Data retrieved from the database are mapped onto data objects. The internal representation of a data object is such that the relationship between the target product and its competitors, the target product and its component products, and the target geography and its component geographies is preserved. The system forward chains on these data objects and uses heuristics to produce explain, pattern, and segment objects. *Explain objects* represent causes for the performance of products. *Pattern objects* represent correlations among market share and causal factors across the set of all products. *Segment objects* represent a significant shift in volume among product segments.

Associated with a data object representing a product, there might be one or more explain objects representing causes for the performance of the product. These explanations are represented as a *directed acyclic graph* of interrelated concepts. This graph is a model of how the causal factors affect the performance of a product. The product performance in need of explanation is attached to a top-level entry node. Subordinate nodes represent explanations for the entry node. Links from the entry node represent a partial preference ordering of explanations. Nodes closer to the top are stronger explanations. Once the strongest explanation is found, weaker explanations are not considered.

When all relevant explain, pattern, and segment objects have been created, the system shifts from the analysis phase into the output-generation phase. This phase is driven by the specification of the report and controlled by the objects produced in the analysis phase.

Based on the data, explain, pattern, and segment objects in memory, the system selects a bullet, table, or graph instance whose condition attribute is satisfied and establishes this instance as a goal. The values of the parts attributes of this goal then constitute subgoals recursively until one or more generation rules is fired to generate the necessary output. Only subgoals whose condition attributes are satisfied are established as goals. Selection rules determine whether a condition attribute is satisfied. For example, the parts of a bullet are various sentences that can be produced conditioned on the presence or absence of specific data and explain objects. If the conditions are satisfied, then the appropriate sentences are output.

SPOTLIGHT includes rules that are capable of producing a variety of sentences that describe the performance of a product, explanations for the performance, and so on. These rules generate text by completing a wide variety of context-dependent templates. The output of this module is a file of report contents and format commands. The formatting module reads this file and generates WORD PERFECT files for display and printing.

Similarly, SPOTLIGHT includes rules that are capable of producing different types of graphs and tables. The system selects among alternative graphs based on the relationship between the objects produced by the analysis phase. For example, a pattern object that represents a correlation between market share and price generates a substantially different graph than a pattern object that represents a correlation between market share and distribution.

The expert system module consists of approximately 200 parameterized rules, 70 different sentence types, 10 different graph types, and 10 different table types. These units can be controlled flexibly to produce selected reports or define new reports.

Why Knowledge-Based Techniques?

The heuristics changed as the expert validated the results of initial implementations. We used rule-based techniques to facilitate representation of the domain knowledge and allow iterative refinements. As we expected, a rule-based representation also made it easy to allow user customization of key parameters: Custom values instantiated parameterized rules.

A commercial forward-chaining expert system tool, ECLIPSE (Haley 1991), greatly reduced the development cycle by providing an inference engine that enabled experts' rules to be encoded in a straightforward manner. In addition, the use of ECLIPSE reduced the complexity of the code relative to a procedural language. ECLIPSE is a commercially supported, more efficient version of CLIPS (Giarratano 1991). ECLIPSE was preferred to alternative products because it uses a standard (CLIPS-ART) syntax, requires substantially less memory, and is faster. ECLIPSE enabled delivery on a conventional PC/286, 640K platform.

Application Use and Payoff

SPOTLIGHT was released on 15 July 1991. Response from customers in the marketplace has been overwhelming. A large number of Nielsen clients have taken delivery of SPOTLIGHT. As a result, development costs were recovered in six months, and product revenues are currently growing ahead of plan. With the success of SPOTLIGHT, Nielsen is planning further knowledge-based applications.

Most users have found that SPOTLIGHT provides a quick, effective overview of the market and pinpoints areas for richer, in-depth analyses. Client feedback includes comments such as "we have an effective tool to realize incremental benefits of the data we are purchasing from Nielsen" and "our analysts do not have to spend all their time browsing through data and making pretty graphs and tables."

SPOTLIGHT is designed for easy installation and minimal training. Nielsen field representatives provide training for clients, usually lasting no more than a day. Because of its ease of use, SPOTLIGHT is rapidly spreading through client organizations. For example, one major manufacturing organization has already deployed SPOTLIGHT to 250 field sales representatives and 19 regional centers. The regional centers produce reports for hundreds of other sales representatives that need the information to understand brand behavior in their territories. It appears that this client is saving hundreds of hours a month relative to the time it used to take to analyze data with spreadsheet tools.

Application Development and Deployment

About 48 person-months were spent on SPOTLIGHT over a period of 7 months from concept to delivery. The developers, a core of six, were split into three teams. One team had responsibility for the expert system module, one for mainframe extraction, and one for conventional PC software. The efforts were cost justified by expected revenues and client demand for interpretive tools that could enhance the value of Nielsen data.

Validation of the application was done by running dozens of case studies for review. The development team worked closely with the marketing organization. Domain experts from marketing were responsible for validation. SPOTLIGHT might be one of the first systems to explore a unique relationship between marketing and product development—one that goes beyond traditional roles of requirement generation and system analysis to those of domain expert and knowledge engineer. Approximately three months of effort were required to roll the product through Nielsen's entire sales organization.

Maintenance

We are already realizing the tremendous advantages of using knowledge-based techniques in the ease with which we maintain the system and the rapidity with which we respond to requests for enhancements. Based on client feedback, two new releases were made within six months of the original release.

SPOTLIGHT's design addresses maintenance and upgrade needs in several ways. For example, a clear separation exists between analytic capabilities and presentation generation. This design makes it possible to embed SPOTLIGHT capabilities in other delivery vehicles, such as planned interactive decision support systems. In addition, text is clearly separated from the rules that decide which concepts to express. This separation between text and rules makes it easier to package SPOTLIGHT for international delivery.

Maintenance requires the ability to produce subsequent releases with minimal costs. One area of concern, particularly for rule-based systems, is regression testing. Because rules are evoked depending on the state of data, it is difficult to guarantee regression with large databases. Consequently, we created a continually growing library of test cases that were manually validated by the experts. This library includes cases that account for all possible input combinations. We developed utilities to ensure that this library of test cases leads to 100-percent coverage of the rules.

Whenever changes are made to the rules, all cases in the library are run in a batch mode. The resulting output is compared with the previous output, and the differences are presented to the expert. If these differences are acceptable, the library is updated. After every change, the library is reviewed to ensure coverage of all the rules. If necessary, new cases are added to the library.

Finally, attempts were made to enable limited end user maintenance. Where possible, the behavior of the expert system can be controlled by using an interactive setup facility that allows custom definitions of key concepts and threshold values.

References

Anand, T., and Kahn, G. 1992. SPOTLIGHT: A Data-Explanation System. In Proceedings of the Eighth IEEE Conference on AI for Applications, 2–8. Washington, D.C.: IEEE Computer Society.

Giarratano, J. C. 1991. *CLIPS Reference Manual.* Houston, Tex.: National Aeronautics and Space Administration.

Haley, P. 1991. *ECLIPSE Reference Manual.* Sewickley, Pa.: Haley Enterprise.

Mann, W. 1983. An Overview of the PENMAN Text-Generation System. In Proceedings of the Third National Conference on Artificial Intelligence, 261–265. Menlo Park, Calif.: American Association for Artificial Intelligence.

Springer, S.; Buta, P.; and Wolf, T. 1991. Automatic Letter Composition for Customer Service. In *Innovative Applications of Artificial Intelligence 3*, 67–83. Menlo Park, Calif.: AAAI Press.