September 2001

Knowledge Discovery: Some Empirical Evidence and Directions for Future Research

Meliha Handzic
University of New South Wales, m.handzic@unsw.edu.au

Aybüke Aurum
University of New South Wales, aybuke@unsw.edu.au

In: Buhl, Hans Ulrich, u.a. (Hg.) 2001. *Information Age Economy*; 5. Internationale Tagung Wirtschaftsinformatik 2001. Heidelberg: Physica-Verlag

ISBN: 3-7908-1427-X

Knowledge Discovery: Some Empirical Evidence and Directions for Future Research

Meliha Handzic, Aybüke Aurum
University of New South Wales

Summary: Large amounts of data generated by electronic commerce are becoming an increasingly important source of knowledge to support organisational decision making. An empirical study was conducted in a simulated electronic commerce environment to examine people's ability to discover varying associative patterns in transactions data, and utilise that knowledge to support product sales forecasting. The results of the study indicate that people were able to reasonably well discover valid associations among data items and consequently improved performance over naive forecasts. The results also indicate that people were more successful in recognising and using stronger rather than weaker associative patterns. However, they failed to reach optimal performance.

Keywords: Knowledge discovery; Knowledge management; Electronic commerce; Judgemental forecasting

1 Introduction

The world is experiencing a major transition from an industrial to a knowledge based society. Parallely with this transition comes a growing recognition among researchers and practitioners alike of the importance of knowledge management with respect to the struggle for economic success. Knowledge has been described as the principal fuel and currency that will drive the new economy [Devl99]; a key resource of tomorrow's organisations in the most competitive economy we have yet known [Druc93]; and a hidden gold embodied in the minds and hands of organisational participants [Stew97]. Knowledge management (KM) is an emergent response to the need to accelerate both the creation of knowledge and its application to physical resources in the battle for competitive advantage or survival. The central task of those concerned with knowledge management is to determine ways to better cultivate, nurture and exploit knowledge to improve performance.

In order to effectively manage knowledge, companies need to consider many avenues for accumulating knowledge. Recent developments in information and communication technologies, particularly the flexibility and user-friendliness of the Internet, have made it possible for modern organisations to increasingly conduct

their transactions with their customers, suppliers and other trading partners by means of electronic commerce (EC). The growing number of EC transactions among organisations and their customers generate large volumes of electronic data that are becoming an increasingly important new source of organisational knowledge [Blan00]. Vast amounts of data accumulated in organisational databases may contain potentially valuable new knowledge. One of the main KM issues with respect to EC is the discovery of knowledge that is implicit in the EC transactions and utilisation of the extracted knowledge to support decision making. The purpose of this study is to address this issue in the context of product sales forecasting.

2 Literature Review

Knowledge discovery has been described as the nontrivial process of identifying, valid, novel, potentially useful and ultimately understandable patterns in data [Fayy96]. Two main goals of knowledge discovery include: description and prediction. Description is concerned with identifying patterns for the purpose of presentation to users in an understandable form. In prediction oriented datamining, the patterns are being discovered for the purpose of predicting future values of some variables. If the discovered knowledge is going to be used for judgement and decision making, the comprehensibility of the extracted pattern is considered to be crucial. There are four major categories of KD approaches: classification, association, sequence and cluster [Mara99]. The focus of the present study is on association analysis. Such analysis can discover highly useful and informative patterns within data that can be used to develop predictive models of behaviour in a wide variety of knowledge domains. For example, if one discovers that changes in ice-cream sales are related with changes in air temperature, sunshine hours and tourists numbers, one can use that knowledge to make more accurate estimates of future sales, translate them into more suitable production plans and, as a result, minimise losses due to overproduction or missed sales opportunities.

A generic term ‘contextual knowledge’ is used in the forecasting literature to denote the knowledge of a variety of cause-effect relationships and specific environmental and organisational factors such as market intelligence and promotion plans. The value of such knowledge is seen in its ability to explain past and anticipate future changes in the behaviour of the variable being forecast and to enable a decision maker to deal more competently with his or her decision task. However, despite plausible theoretical arguments for the importance of contextual knowledge, it remains unclear whether and how well can individuals discover, extract and utilise contextual knowledge contained in electronic data. Past empirical studies on the issue indicate mixed and inconclusive findings. Several studies using real world settings support the notion that the possession of contextual

knowledge improves the quality of human judgement in product sales and marketing areas [Edmu88; Sand92; Fild91; Math86, Math89; Wolf90]. In contrast, a more recent field study of sales forecasting by Lawrence et al. (1995) revealed that bias and inefficiency could mask any contribution of seemingly helpful contextual knowledge to the accuracy of judgemental forecasts. This study compared the accuracy of judgementally produced monthly product forecasts with naive forecasts of thirteen companies. Empirical evidence from a number of laboratory studies also casts doubt on people's ability to effectively process contextual/causal knowledge embedded in electronic databases. [Andr91; Harv94; Lim95, Lim96a; Lim96b; Good99].

In view of inconsistent prior findings and concerns expressed, it is of particular interest to this study to examine the situation in which judgemental forecasters are provided with the same amount and regularity of EC transaction data with valid but varying associative patterns (weak or strong). The question is (i) whether and how well can individual forecasters discover the varying associative patterns in data and (ii) and whether and what impact that knowledge may have on their subsequent forecast accuracy.

3 Research Methodology

The experimental task in the current study was a simulated forecasting activity in which subjects made estimates of daily sales of fresh ice-cream. Participants assumed the role of Manager of a fictitious firm that sold ice-cream products in a beach suburb of Bondi in Sydney. The manager made daily sales forecasts as a part of the production planning process. The company incurred equally costly losses if the production was set too high (due to spoilage of unsold product) or too low (due to loss of market to competition). The manager's goal was to minimise the company's total loss, by minimising forecast errors. During forecast preparation, the manager could consult the company's database containing aggregated EC transactions data on past sales. The database also contained regular weather (ambient air temperature, amount of sunshine) and tourist forecasts obtained from EC transactions with Meteorology Bureau and Tourist Board respectively. Temperature, sunshine and tourist data was used to simulate the effects of continuous contextual factors on the sales time series. The subjects repeated the task for thirty consecutive simulated days. Before commencing the real task, they were allowed to make five trial forecasts, for practice purposes only. Throughout the experiment, instructions were provided to inform subjects of the task scenario and of the requirements. Performance feedback was provided to enable participants to analyse their earlier performance and to adjust their future strategies.

A laboratory experiment with random assignment to treatment groups was used, as it allowed greater experimental control and made possible drawing of stronger in-

ferences about causal relationships between variables. The only independent variable was *associative pattern* (weak or strong) in data. The condition of weak and strong associative pattern was achieved by adding different error terms to sales series data. Error terms were produced by drawing random values from normal distributions each having mean 0 and a standard deviation of 1.66 and 5.87. The first standard deviation corresponded to strong association (r=0.95), and the second standard deviation corresponded to weak association (r=0.65) between individual factors and sales variables. Product sales series figures were produced by drawing random values from a normal distribution with a mean of 25 and a standard deviation of 5 (hundred units).

Task performance was evaluated in terms of forecast accuracy operationalised by cumulative relative absolute error (CumRAE). It was calculated as a ratio of the cumulative absolute error of judges' forecasts and the corresponding error of the 'random walk' (naive) strategy [Amst92]. A naive forecast is one that simply determines the next day sales as equal to the current day's sales. Such strategy makes no use of any contextual knowledge and typically produces poor performance. Relative measure was used to assess improvement in the quality of forecasts due to knowledge extracted. Scores equal to 1 indicated no improvement, while scores smaller than one indicated improved accuracy. In addition, optimal forecast errors were calculated to assess how much of the maximum hidden knowledge was extracted and used by the subjects. Optimal strategies are modelled using stepwise regressions with three contextual variables (temperature, sunshine, and tourist) as independent variables and sales data as the dependent variable. The optimal response integrates all three variables into a single response using regression weights and produces the best possible performance given the available patterns in data. Scores can vary from 0 (ie naïve) to 1 (ie optimal).

The subjects were thirty-two postgraduate students from the University of New South Wales, Australia. Sixteen subjects were assigned to each of the two treatment groups. They participated in the experiment on a voluntary basis. They received monetary rewards for the best three results in each of the two treatment groups. Some previous studies indicated that postgraduate students were appropriate subjects for this type of research [Asht80; Remu96; White96]. The experimental session was conducted in a microcomputer laboratory. On arrival, subjects were assigned randomly to one of the treatment groups, by choosing a microcomputer from a number of units set for the experiment. Before commencing the task, subjects were briefed about the purpose of the experiment and read case studies descriptions incorporated in the research instrument. They then performed the task. The session lasted about one hour.

4 Results

Mean performance scores (CumRAE) of experimental subjects by associative pattern groups are presented graphically in Figure 1. Mean values of naive and optimal forecast errors are also presented for comparison purposes. The collected data were further analysed by one-way analysis of variance (ANOVA) method. The analysis found some significant results.

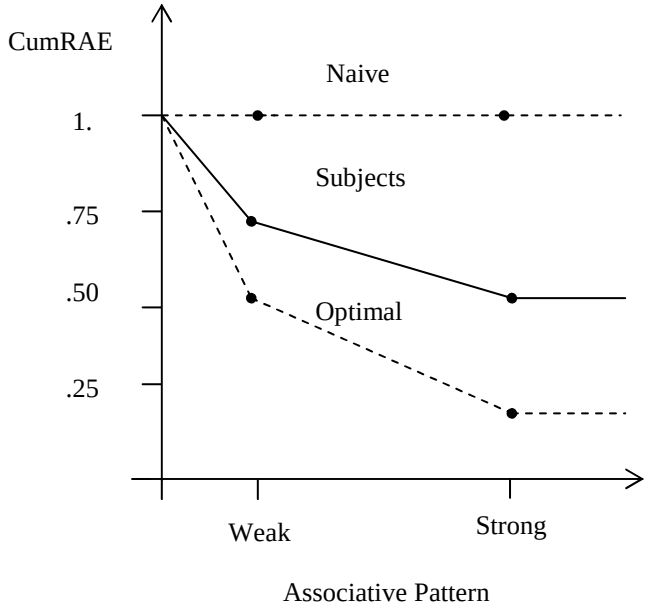

Figure 1. Subjects' forecast errors (CumRAE) by associative pattern

First, error scores of all subjects were lower than 1. Subjects tended to make smaller than naive forecast errors both in weak (0.75) and strong (0.49) associative pattern conditions. These results indicate real improvement in performance due to knowledge extracted. The analysis found that error scores dropped by 25% and 51% in weak and strong association groups respectively. Second, the results of the analysis performed indicate significant differences in error scores between subjects in the two associative pattern groups. Subjects in strong associative patterns condition tended to make significantly smaller forecast errors than their counterparts in weak associative patterns group (0.49 vs 0.75, $F(1,30)=12.809$, $p=0.001$). Smaller relative errors indicated better knowledge extraction and greater improvement in the quality of actual forecasts over naive forecasts. Third, Figure 1 shows that subjects tended to make greater than optimal errors in both weak ($0.75>0.56$) and strong ($0.49>0.19$) associative pattern groups. Further analysis

showed that subjects discovered and extracted on average 57% of the hidden knowledge in weak and 63% in strong associative pattern groups.

5 Discussion

In summary, the main results of the study indicate that people are reasonably good at discovering and utilising valid knowledge about associative patterns in data and improving their forecasting performance. However, gains were less than theoretically possible given the objective relationships among data provided. In addition, forecasters tended to make significantly better forecasts when provided with data containing stronger rather than weaker associative patterns.

The fact that the participants made better than naive forecasts indicates that they were able to extract some knowledge about the associative patterns in available data. As a result, they tended to improve their subsequent performance. In real terms error scores dropped by 25% and 51% in weak and strong associative patterns groups respectively. Such findings seem to contradict the overly pessimistic picture of human ability to utilise contextual knowledge painted by earlier laboratory research in judgemental forecasting [Andr91; Harv94]. One potential explanation for the difference may be in the characteristics of the tasks used. Participants in the current study were provided with a meaningful task context, a small number of data items with valid causal relationships, and forecast values to suggest future behaviour. It is also possible that a graphical form of presentation facilitated discovery of associations and enabled the subjects to better judge the right direction of future changes. This is consistent with Lawrence et al. (1985) findings that graphical presentation enhances the accuracy of novices.

Furthermore, the results of the current study indicate that the strength of association present in the available data had a significant impact on forecasters' working knowledge and performance. The participants with stronger associative patterns tended to perform substantially better than those with weaker associative patterns. This was demonstrated by significantly smaller forecast errors found among subjects from the strong compare to the weak association group. It is possible that the subjects with weaker associative patterns perceived their task as more complex due to increased predictive uncertainty and consequently tended to choose simpler strategies to deal with their problem. A contingent model for choice of forecasting strategies [Beac86] identifies task complexity as one of the major contibuting factors to systematic deviations from the optimal behaviour by human judges. Findings from a large number of past empirical studies in behavioural decision theory also provide strong support for the contingent nature of human decision making upon task complexity [Payn88]. The other possible reason for difference in performance found could be the use of a pattern matching strategy. In pattern matching, which can be seen as a form of the representativeness heuristic, a single

past case which will contain an element of noise, is used as the basis for the prediction, while general tendencies such as long-run time series trends are ignored. Hoch and Schkade (1996) found evidence of pattern matching in a judgemental time series forecasting task and suggested that pattern matching is a fairly good strategy in highly predicable environments, but is deficient when the environment contains high levels of noise.

Despite this, the results indicate room for further improvement. Subjects were found to make greater forecast errors than optimal. Greater than optimal errors indicate that the subjects tended to uncover less knowledge than they possibly could from their available data. Further analysis revealed that on average they extracted and used between one half and two thirds of the maximum knowledge about associative patterns in weak and strong conditions respectively. One potential explanation for the observed suboptimality is that subjects lacked vital analytical and procedural knowledge. All subjects had an opportunity to learn from their own experience through task repetition and from feedback. However, it seems that the period of 30 trials was too short to induce more effective learning. Earlier studies on learning from feedback [Klay88] indicate that while people can well learn the existence and direction of cue-criterion relations over a larger number of trials (ranging into the hundreds) they generally have difficulties in learning their shape. As a result they tend to perform suboptimaly.

While the current study provides a number of interesting findings, some caution is necessary regarding their generalisability due to a number of limiting aspects. One of the limitations refers to the use of a laboratory experiment that may compromise external validity of research. Another limitation relates to artificial generation of data that may not reflect the true nature of real business. The subjects chosen for the study were students and not real life forecasters. The fact that they were mature graduates may mitigate the potential differences. Small monetary incentives offered to the subjects for their effort might not have been sufficient to motivate them to try as hard as they could. Most decisions in real business settings have more significant rewards. Further research is necessary that would extend the study to other subjects and environmental conditions in order to ensure the generalisability of the present findings.

Although limited, the findings of the current study may have some important implications for organisational decision making in general and EC management in particular. Firstly, they raise the awareness of a contingent nature of the knowledge discovery process by humans upon the complexity of the pattern in data, and they provide a warning regarding the systematic deviations from optimal behaviour. Secondly, they suggest that individual sales and marketing personnel could potentially benefit from those KM initiatives that would enhance their understanding of the existence and the form of relationships among various EC events, and facilitate optimal performance.

One possible solution is to provide training in analytical and statistical reasoning to all knowledge workers. Alternatively, organisations may employ specialists trained in these areas who would perform knowledge extraction process for other professional and managerial knowledge workers. Another possible solution is to create working contexts that encourage communication and culture of knowledge sharing among employees. The spiral knowledge postulates that the processes of sharing will result in the amplification and exponential growth of working knowledge [Nona98; Nona95]. Alternatively, organisations may choose to use automated data mining tools to more efficiently discover patterns in data. However, human judgement will still be needed to interpret them. Finally, combining and integrating various KM initiative may create synergy effect and even higher levels of knowledge . According to Davenport and Prusak (1997) only by taking holistic approach to managing knowledge it is possible to realise the full power of information ecology. Future research may look at some of these issues.

6 Conclusions

The main objective of this study was to investigate the ability of people to discover varying associative patterns in EC data and utilise that knowledge to improve product sales forecasting. The findings indicate that participants were able to reasonably well discover and utilise some, but not all knowledge hidden in EC data irrespective of the pattern strength. Accordingly, they achieved some performance improvement, but failed to accomplish what was possible. They also recognised and used stronger associative patterns in data significantly better than weaker patterns. Although limited to the specific task and context, these findings may have some important implications for organisations. In particular, they suggest the need for other knowledge management initiatives to further enhance knowledge and performance. Therefore, more research is necessary to systematically address various initiatives in different tasks and contexts, and among different knowledge workers, if a better understanding of the area is to be achieved.

References

[Arms92] Amstrong, J. S., Collopy, F.: 'Error Measures for Generalising about Forecasting Methods: Empirical Comparisons', *International Journal of Forecasting*, 8 (1992), 69-80.

[Andr91] Andreassen, P. B.: *Causal prediction versus extrapolation: Effects on information source on judgemental forecasting accuracy*, working paper, MIT, 1991.

[Asht80] Ashton, R. H., Kramer, S. S.: 'Students as Surrogates in Behavioural Accounting Research: Some Evidence', .*Journal of Accounting Research*, 18, 1 (1980), 1-15.

[Beac86] Beach, L. R., Barnes, V. E., Christensen-Szalanski, J. J. J.: 'Beyond Heuristics and Biases: A Contingency Model of Judgemental Forecasting', *Journal of Forecasting*, 5, 3 (1986), 143-157.

[Blann00] Blanning, R.W.: '*Knowledge Management and Electronic Commerce*', Position Papers on Future Directions in Decision Support, IFIP WK8.3 Working Conference on DSS, Stochholm, 2000.

[Dave97] Davenport, T.H., Prusak, L.: *Information Ecology*, Oxford University Press, Oxford, 1997.

[Devl99] Devlin, K.: *Infosense:Turning Information into Knowledge*, W.H. Freeman and Company, New York, 1999.

[Druc93] Drucker, P.F.: *Post-Capitalist Society*, Harper Business, New York, 1993.

[Edmu88] Edmundson, R. H., Lawrence, M. J., O'Connor, M. J.: 'The Use of Non-Time Series Information in Sales Forecasting: A Case Study', *Journal of Forecasting*, 7, 3 (1988), 201-211.

[Fayy96] Fayyad, U., Piatetsky-Shapiro, G., Smyth, P.: 'Knowledge Discovery and Data Mining: Towards a Unifying Framework', *Proceedings of the Second International Conference on Knowledge Discovery and Data mining, KDD-96*, Oregon, 1996.

[Fild91] Fildes, R.: 'Efficient use of information in the formation of subjective industry forecasts', *Journal of Forecasting*, 10 (1991), 597-617.

[Harv94] Harvey, N., Bolger, F., McClelland, A.: 'On the nature of expectations', *British Journal of Psychology*, 85 (1994), 203-229.

[Klay88] Klayman, J.: Learning from Experience in Brehmer, B. and Joyce, C.R.B. (eds) *Human Judgement. The SJT View*, North-Holland, Amsterdam, 1988.

[Lawr85] Lawrence, M., Edmundson, B., O'Connor, M.: An Examination of Accuracy of Judgemental Extrapolation of Time Series, *International Journal of Forecasting*, 1, 25-35, 1985.

[Lawr95] Lawrence, M., O'Connor, M., Edmundson, B.: *A Field Study of Sales Forecasting: Its Accuracy, Bias and Efficiency*, Working paper, School of Information Systems, The University of New South Wales, July, 1995.

[Lim95] Lim, J. S., O'Connor, M.: 'Judgemental Adjustment of Initial Forecasts: Its Effectiveness and Biases', *Journal of Behavioural Decision Making*, 8 (1995), 149-168.

[Lim95a] Lim, J. S., O'Connor, M. J.: 'Judgemental Forecasting with Time Series and Causal Information', *International Journal of Forecasting*, 12 (1996), 139-153.

[Lim96b] Lim, J. S., O'Connor, M. J.: 'Judgemental Forecasting with Interactive Forecasting Support Systems', *Decision Support Systems*, 16 (1996b), 339-357.

[Mara99] Marakas, G.M.: *Decision Support Systems in the 21 st Century*, Prentice-Hall, New Jersey, 1999.

[Math86] Mathews, B. P., Diamantopoulos, A.: 'Managerial Intervention in Forecasting: An Empirical Investigation of Forecasting Manipulation', *International Journal of Research in Marketing*, 3, 3-10, 1986.

[Math89] Mathews, B. P., Diamantopoulos, A.: 'Judgemental revision of sales forecasts: A Longitudinal Extension', *Journal of Forecasting*, 8, 129-140, 1989.

[Nona95] Nonaka, I., Takeuchi, H.: *The Knowledge Creating Company: How Japanese Companies Create the Dynamics of Innovation*. Oxford University Press, New York, 1995.

[Nona98] Nonaka, I.: The Knowledge-Creating Company, *Harvard Business Review on Knowledge Management*, HBS Press. Boston, 1998.

[Payn88] Payne, J.W., Bettman, J.R., Johnson, E.J.: 'Adaptive Strategy Selection in Decision Making', *Journal of Experimental Psychology: Learning, Memory and Cognition*, 14(3), 534-552, 1988.

[Remu96] Remus, W.: 'Will Behavioural Research on Managerial Decision Making Generalise to Managers?', *Managerial and Decision Economics*, 17 (1996), 93-101.

[Sand92] Sanders, N. R., Ritzman, L. P. 'The Need for Contextual and Technical Knowledge in Judgemental Forecasting', *Journal of Behavioural Decision Making*, 5, 1 (1992), 39-52.

[Stew97] Stewart, T.A.: *Intellectual Capital: The New Wealth of Organisations*, Doubleday, New York, 1997.

[Whit96] Whitecotton, S. M.:'The Effects of Experience and a Decision Aid on the Slope, Scatter, and Bias of Earnings Forecasts', *Organisational Behaviour and Human Decision Processes*, 66, 1 (1996), 111-121.

[Wolf90] Wolfe, C., Flores, B.: 'Judgemental Adjustment of Earnings Forecasts', *Journal of Forecasting*, 9, 4 (1990), 389-405.