Tom Breur

runs the consulting firm XLNT Consulting (www.xlntconsulting.com) dedicated to helping companies make more money with their data. His fields of interest span data mining, analytics, data quality, IT governance, data warehousing and business models.

Papers

Data analysis across various media: Data fusion, direct marketing, clickstream data and social media

Tom Breur

Received (in revised form): 22nd August 2011

Abstract

As the volume and diversity of digitally stored data keeps growing, the importance for the business to create a holistic picture of the customer continues to grow. In this paper, we discuss the possibilities available to researchers with four different streams of data. First, we discuss data fusion, merging of proprietary databases with market research data, which enables appending attitudinal data to customer behaviour. Secondly, best practices in direct marketing leverage analysis of customer response patterns to provide the right offer to the right customer at the right moment. Thirdly, analysing internet usage data allows many opportunities for small-scale, agile marketing pilots. And finally, social media are discussed, one of the newer opportunities to collect data about customer sentiments that relies heavily on text mining analytics.

Journal of Direct, Data and Digital Marketing Practice (2011) **13**, 95–105.
doi:10.1057/dddmp.2011.32

Keywords: data analytics, data fusion, direct marketing, internet usage, social media, text mining

Introduction

We are living in an ever-more ‘digitized’ society: just about every possible step we take gets recorded in some electronic format. This provides a tremendously rich trail of data, mapping out our existence in great detail. For the first time in history, more data *about us* are being recorded than we produce *ourselves*.¹

Companies employ multiple channels to market to and stay in touch with their customers. To get a holistic view on their customers, businesses need to paint a picture across *all* touch points. Market research and direct marketing have been around for a while. And every company has a website these days that provides a rich source of information that may or may not have been fully leveraged, yet. After social media really took off, many companies are now seizing that opportunity to stay in touch with their customers through that channel as well.

The onus is on data management professionals to compile a 360° view on customer behaviour, and set their research agendas

A 360° view requires consolidating *all* touch points

Tom Breur
XLNT Consulting
Langestraat 8-03
Tilburg 5038SE, The Netherlands
Tel: +31-6-463 468 75
E-mail: tombreur@xlntconsulting.com

accordingly. How else can you arrive at a comprehensive assessment of the impact marketing has on your customers? In this paper, we will address the following questions:

- How can you fuse market research with (proprietary) databases so as to capture all value contained in surveys?
- What are the best practices to capitalize on wisdom contained in the contact history of your direct marketing campaigns?
- How can you extract value from internet usage data?
- What do social media have to offer for our research agenda? And how do we extract meaning from it?

The amalgam of data across all these channels provides many opportunities for cross-fertilization. By blending different types of research data, our information becomes more useful. But at least as important, it allows us to provide a vivid context for 'plain' numbers so that they can speak for themselves. For example, when you append sentiments that were extracted from social media research to clickstream data that describe website visitor behaviour, the whole becomes more than the sum of parts.

In this paper, we describe benefits from capturing the full extent of interactions with your customers, to avoid missing out on potentially valuable information. Given the strategic importance that is attributed to data assets, it is important to follow contemporary developments and trends. Davenport and Harris² have sparked a wave of interest in business analytics. Others like Redman,³ Taylor and Raden,⁴ and Ayres⁵ have expanded on this view.

Besides market research and database analysis, we are finding new ways to leverage dialogues with our clients on the web. Internet is the ultimate pilot market playground once you know how to gather and use these data. Social media provide another new avenue to explore, with valuable and timely information.

New sources of data will always emerge. The present day challenge is to integrate new channels in some efficient way. This should enable businesses to extract economic value from these data. Just because data are available does not necessarily make them useful. And even *if* data are useful, storing and managing them still needs to make economic sense.

Fusing market research with proprietary databases

Most companies have access to databases (or a data warehouse) with customer data. It may contain client characteristics, and quite likely a recording of product ownership and/or usage. Valuable as these data are, the picture is still incomplete. One thing that is missing from the equation is how customers *feel* about your business. This can be quality perception, satisfaction with your service, likelihood to recommend, etc.

Market research comes in a variety of flavours. It can be aimed at scanning the market as a whole through studies that explore not only captive customers, but also non-customers, which therefore makes it more representative of the market as a whole. In the context of data

Attitudinal data can enrich behavioural (database) data

fusion, the objective of market research is usually to combine behavioural database findings with attitudinal data. The latter are sometimes referred to as 'soft' data.

Because these 'soft' customer measures have a significant role to play, many companies have installed dedicated market research teams. Market researchers and database practitioners tend to be a very different breed. But obviously, there is value if you can somehow bring these data sources together.

In particular, when market research or lifestyle data can be tied down all the way to transactional behaviour, this link becomes very valuable. By merging customer attitudes or sentiments to a description of behaviour (eg transactional data), you prove a richer context for research data. This begs the question: *how* can we combine results from internal database systems with external market research?

One can distinguish between two kinds of data fusion. Either the relation between market research data and database entities is associative or deterministic. Associative data fusion is when a statistical model provides a chance connection between records coming from two different sources. Therefore, the relation is based on a high (but not 100 per cent) probability that two records belong to the same individual. In the latter case (deterministic fusion), a one-on-one match is possible using unique identifiers that are available in both the market research and customer data set. This has been called the exact matching problem.⁶

Different terms have been used for data fusion, like: micro data set merging, statistical record linkage, multi-source imputation or ascription. In data fusion, the matched subset of database records that are connected to market research is referred to as 'workbench data'. And in both associative and deterministic fusion, the matching process needs to preserve privacy of research participants, as well as confidentiality of the database source.

Privacy protection is of utmost importance when fusing data sets

Privacy protection is of paramount importance. Customers place their trust in companies when they volunteer sensitive information. Companies put their reputation on the line when conducting such research. The outcome should lead to an overall better customer experience, and should steer shy from exploiting short-term sales opportunities that violate the customer's trust.

Associative data fusion has been developed to enable an assumption-based overlay of, for instance, market research and database contents. But the same process can also be applied to lifestyle data, or more generically any kind of data enrichment. Since there is no match possible at the individual level, overlaying is based on the so-called linking of common variables in donor records. Fuzzy matching is most often performed using k-nearest neighbour techniques, where the objective is to find records that are closest in geometric hyperspace. These geometric 'neighbours' are then deemed a match.

The full details of how these algorithms work goes beyond the scope of this paper. The interested reader is referred to van der Putten⁷ for an in-depth treatment of this topic. The fundamental principal underlying all of these data fusion algorithms is that there needs to be a set of

‘overlapping’ (common) variables that are known for your database customers, and which have been gathered from your market research sample. Examples could be fields like age, gender, which products customers hold, etc.

These so-called ‘matching fields’ are subsequently used for extrapolation of the full richness of market research data onto the rest of the customer base. On the basis of the linked workbench data, one can determine the most predictive variables in relation to market research or lifestyle characteristics.

Analysis of the workbench data set will converge to an optimal choice of model(s). After (optimal) models have been established, you can now deploy these models on the entire customer database. The assessment of what constitutes the ‘optimal’ model for data set matching depends on the algorithm used.

Often, data fusion algorithms are distance-based, comparable to K-means clustering.⁸ Then, your optimal model will be the one that minimizes distances, after taking into account certain variables (so-called cell variables) on which a match is mandatory. An example of a cell variable might be gender. This is typically done to avoid predicting something nonsensical like percentage of pregnant males. Overviews of the data fusion literature have been provided, among others, by Radner *et al.*,⁶ van Pelt,⁹ Raessler¹⁰ and D’Orazio *et al.*¹¹

This allows inference of most likely market research segments or lifestyle clusters. Now you can infer ‘soft characteristics’ with some statistical likelihood, even though these customers have not been involved in the market research.

One of the unique possibilities data fusion offers is to infer market characteristics like Share-of-Wallet (S-o-W) for the entire customer database. Predictive models that link database attributes (like transactional behaviour) to market research data provide powerful insights into drivers of ‘soft’ measures like satisfaction, attitude, or S-o-W. This will help the business identify observable indicators for latent or intangible key success factors (like brand advocacy, etc).

Applications that have benefited from this approach are diverse. In the UK, a number of years ago a well-known financial segmentation project was undertaken on the basis of NOP’s Financial Research Survey data. Using data fusion with exact matching, FRuitS, a customer segmentation was devised.¹² One of the unique benefits of this FRuitS segmentation is that it is based on individual characteristics rather than neighbourhood (ZIP-code) profiles. This FRuitS scheme classifies inhabitants of the UK as either Plums, Pears, Lemons, etc, where each cluster has been appointed a descriptive name, relative to their financial position and family characteristics. Later successful projects include the BARB/TGI data fusion, which merged media and product usage,¹³ and Touchpoints/IPA.¹⁴

Best practices in direct marketing

Roughly in the 1990s, direct marketing came to the fore.¹⁵ It had been ‘invented’ before the term ‘direct marketing’ was coined in 1967 by

Direct marketing enables ROI calculations from marketing spend

Lester Wunderman, but it was not until computers enabled wholesale capture and storage of large volumes of customer profiles that this discipline reached the mainstream of business practice.

Direct marketing's claim to fame is that it is the only way to provide measurable, accountable advertising. Initially, this favoured direct mail, but with the rise of the internet (and email) even cheaper cost-per-contact opportunities became available. For a (short) while, this seemed marketer's Walhalla, until response rates dropped significantly. Nevertheless, direct marketing is here to stay, and for very good reasons.¹⁶

One of the reasons why direct marketing has remained so popular is because it enables exact measurement of the return on investment for your marketing campaigns. By employing sophisticated testing designs,¹⁷ you can progressively fine-tune your efforts to maximize yield.

Testing is an integral part of direct marketing, often referred to as 'closing the loop'. When properly designed, each campaign will enhance your learning about the market, and how to best communicate your offering. By feeding this learning back into the organization, this becomes an ever more powerful, adaptive mode of interacting with your customers.

The term dialogue marketing has followed concepts permission marketing¹⁸ and one-to-one marketing,¹⁹ describing constant 'offer-response-adjusted offer' cycles. Best-in-class companies use dialogue marketing to improve ROI by mass-customizing individual messages to leverage their marketing budget.

A central tenet behind Customer Relationship Marketing was to record interactions with your customers, and subsequently *show* them you remembered their preferences.²⁰ This elegantly fits in with principles of direct marketing, and with researchers' needs to gather intelligence about customers in the course of executing a (direct) contact strategy.

Every campaign provides an opportunity to learn

The overarching philosophy is to treat every campaign as an opportunity to learn. The majority of commercial activities can indeed be leveraged as 'live' tests of customers' preferences and willingness to buy. Direct marketing was initially 'invented' as a means to economically reach very large numbers of customers.

Researchers can get highly relevant business information out of this process that gives up-to-date information on a company's leverage on the market. Conversely, when it comes to designing direct marketing test grids, and determining group size using statistical power analysis,²¹ researchers can play an important advisory role.

For most business people, the relationship between statistics of direct marketing tests and commercial objectives is either not so obvious or buried in a long-forgotten mathematics class. These applied statistics play a pivotal role in rational trade-offs between short-term sales per campaign, and medium-term budget-level optimization of marketing ROI.

Extracting value from internet usage data

The internet offers many opportunities to leverage customer data. Although major strides have been made, it still feels like we have only

Analysis of clickstream data facilitates 'live' testing of product offers

scratched the surface. Analysis of clickstream data (recordings of web users' activity) is of recent origin, yet already a commercial marketplace for these data has been established when internet service providers provide aggregations of so-called server logs to data resellers. Needless to say, this has caused some privacy concerns: internet loggings provide a detailed trace of one's surf activity.

When you plot a maturity curve that businesses typically follow, there seems to be a dominant succession of stages that companies go through. This runs from plain vanilla 'hits' or 'clicks' reporting, to usage of more advanced (multi-variate) statistics, to web visitor life cycle modelling, and eventually leveraging 'live' experimental results.

Clickstream analysis has so far been mostly employed in the e-business world. But there is no reason to limit its application there. The internet has proven to be the greatest experimental 'playground' for businesses to test their products.²² You can set up 'live' experiments in real-time, exposing your customers to many product or advertising variations, getting immediate and highly valid responses. When customers respond to offers that are put in front of them in real-time, spending their own money, this constitutes the most compelling evidence of their preferences. Customers vote with their feet anyway. Why not 'help' them do so, and on the back of this gather live data on their preferences?

Setting up a marketing pilot in the bricks and mortar world can be quite a challenge. On the web, this is considerably easier. What it calls for is a professional content management system, skilled marketing resources, IT and analytics staff. Manipulating experimental control in 'the real world' poses many challenges. On the web, it consists of no more than serving up slightly different pages to different website visitors after the necessary methodology and IT have been put in place.

As an example, advertising research (in particular ad tracking and brand awareness) is rather expensive when conducted in a 'traditional' manner. Clickstream analysis has turned this (previously lucrative) market upside down. Panels of internet research subjects provide much larger samples than previously available. Nonetheless, one always needs to remain critical of the risk of inherent sampling bias with internet panels, respondent bias, and response completeness and quality. But given the substantially larger samples possible with internet panels, you can have elaborate statistical designs tested at economic rates.

Moreover, by rotating ad concepts or art concepts, the audience's response can be measured — rapidly! Instead of 'guessing' what kind of creatives appeal to your audience, why not show them side-by-side to successive website visitors? By monitoring clickthrough rates, you get an excellent idea about audience preferences.

Whereas 'traditional' market research relies on statistical tests with often marginal differences between groups, testing on the web for corporate websites with significant traffic quickly leads to (very) large samples that mostly obviate the need for any statistics at all. If you find any differences, they will by highly significant due to the large sample

sizes. For websites that do not draw a lot of traffic, the test will need to run a bit longer, in order to draw sufficient test subjects, as dictated by statistical power needs.²¹

One way to enrich clickthrough rates is to add short pop-up questionnaires. This can provide some additional context to customer behaviour. Clickthrough rates of course only describe surfing behaviour. Generally, we want 'sticky' sites, meaning that customers stay for as long as possible, and visit as many pages as possible. A longer stay is associated with higher purchasing rates.

By analysing typical sequential patterns in page visits, and combining this with contextual questionnaire results, marketers can manipulate content to appeal to segments they are persuading to buy, because of course only the eventual *purchasing* behaviour will generate ROI.

When 'live' tests on the website lead to inconclusive results, you may switch to qualitative research as needed. One of the biggest gains there is that subsequent focus group research is strictly targeted at specific questions, and only gets applied where it is really needed. Such a set-up provides optimal synergy between quantitative and qualitative research.

Social media and research

In the midst of our contemporary data deluge, we are witnessing a social media revolution. Twitter, Facebook, LinkedIn, weblogs, text messaging, wikis, all engage a lot of consumers. Many of these companies are only a few years old, yet already have market capitalizations in the billions now.

Adoption has been fast. Twitter was launched in 2006, had hardly any users until 2009, and since then has seen a steady rise in its user base expecting 200 million users this year. Facebook is expected to hit one billion users in 2011. LinkedIn is preparing an IPO.

From a business perspective, this growth provides tremendous opportunities to reach potential buyers. The nature of social networks provides fertile ground for viral marketing. When consumers who have no apparent vested interest rave about your product or service, this can provide an enormous boost to your sales. Conversely, a 'hate' campaign can ruin your business very quickly.

Analysing social media relies heavily on text mining technology. This is necessary to qualify text fragments. There has been a divide in this science, roughly, between two schools of thought. There are natural language processing (NLP) methods, which analyse the structure and content of words.²³ These methods analyse grammar, and treat words as they occur in structural relation to each other. And on the other end of the spectrum you have brute force 'number crunchers' that treat texts (almost) like a random 'bag of words'.²⁴

For problems where it is important to retrieve all relevant entries (high recall), NLP methods are (still) superior. So if you want to be sure that you get (almost) all entries for a particular query, this is your preferred approach. This might apply, for example, when a pharmaceutical researcher wants to investigate existing patents to ensure that she has truly discovered a new compound.

Analysing social media relies heavily on text mining

The structure of social media can be likened to mathematical graphs

Another example might be an attorney preparing a lawsuit. In order to familiarize herself with *all* the rulings of judges in past cases, she will need to be sure that no relevant jurisdiction escapes her attention. She will take for granted that she might need to sift through some irrelevant cases.

The internet, on the other hand, is mostly a big numbers game, where often high precision is more important than high recall. When you perform a web search using Google, most users find it important to get *only* relevant hits on the first page. The fact that you might not be seeing a particularly relevant search result that is 'buried' in one of the later search results is less of a concern. For these kinds of applications (precision over recall), 'number crunching' approaches to text mining seem to have gained the upper hand. This difference is caused partly because NLP approaches tend to be more labour intensive, and partly because computing power has gotten so cheap.

Most commercial social media text mining applications need to be built on a (very) limited budget. In addition, business constraints tend to favour precision over recall. Therefore, the balance of business applications is based predominantly on 'brute force' text mining solutions, with an emphasis on flexible reporting capabilities and a versatile mining interface.

Mathematically, you can compare the structure of social media data with a network that is made up of nodes and links: the persons are nodes, and connections between people are links. Sometimes, the nodes are called vertices, and the connections between them are referred to as edges. Together, they form what mathematicians call a 'graph'.

This approach to analysing social media data opens up a vast territory of relatively new but well-established science. Euler first started publishing about this in the early 18th century (he published on the seven bridges of Königsberg²⁵), and has spurred many innovations since. You look at the network without regard to the distances between nodes. The term Graph theory was not introduced until the late 19th century. For social networks, Graph theory is particularly well suited.

So what is a relatively 'new' science for mathematics already provides a rich source of knowledge and algorithms since its inception in modern mathematics. Have you ever wondered how LinkedIn can 'guess' which people you might know? Or why Facebook can recommend certain friends you are likely to know? These graph algorithms provide the answer.

The way this works is by looking at the distance between two nodes (people, typically), and then applying multiple functions to calculate the separation between them. If two people are connected by one mutual friend, there is less of an association compared to another pair that has two mutual friends between them, etc.

The early success of Google's search engine was based largely on graphs. Sergey Brin and Larry Page²⁶ reasoned that if a website has many links pointing to it, it probably means this is a more relevant site to present in relation to your search query. Despite many innovations

to Google technology since their initial Stanford publications, this was, and still is, a cornerstone in their algorithms.

Data warehousing is maturing, which has made new avenues of data storage viable. By employing iterative development methods (Agile BI) that are appropriate for the risks and uncertainties involved in building data warehouses, project risks are mitigated. An early and steady flow of user functionality in predictable cycles helps the organization steer development to maximize value to the business.

Social media are an example of 'alternative' data sources that are finding their way into the data warehouse. To this end, we need to employ text mining so that we do not just read texts but also can fathom associated customer sentiments, and monitor audience response(s).

Text mining relies on a taxonomy of topics that is tied to a thesaurus with relevant words and their connections within each domain. Developing a state-of-the-art domain specific taxonomy and thesaurus (still) requires a large amount of manual work, and hence is usually too costly for 'mainstream' commercial purposes.

Using artificial intelligence, one can mimic this process by 'feeding' a text mining tool semi-structured texts like books that contain a table of contents. Of course these need to be *relevant* books for the domain at hand! Kiddo speak or slang is usually not available in useful formats for text mining purposes, which makes it so hard to create a useful taxonomy and thesaurus for such domains.

Some unique challenges for text mining of social media content are that taxonomies can evolve. And taxonomies for popular slang evolve quickly: not only do we find 'new' words coming in vogue and existing words acquiring new meaning, but the relations between words changes rapidly as well.

To complicate matters further, this type of 'popular' content tends to be very, very 'dirty': people do not take the time spell properly, so texts are chockfull of typos and severe grammatical errors. In fact, the grammatical 'errors' themselves sometimes need to be mined as content, as potentially meaningful vernacular. For example, not all social circles would use a phrase like 'my bad', so depending on the context, this may or may not be a significant entry.

There have been some compelling studies into predictive powers of social media usage. At Pace University,²⁷ a relation between stock price and activity on social media was established. A group at Indiana University (led by Bollen) showed that mood as measured on Twitter predicted (rather than reflected!) movement of the Dow Jones.²⁸ At HP Labs, Asur and Huberman²⁹ demonstrated that Twitter could predict box-office performance of new movies.

From these studies, we now know that social media *can* effectively be used for predicting. But what might be at least as important from a business perspective is that tapping into social media provides an ear to the ground. The turnaround time between a new advertising campaign or a product launch and responses from the marketplace is a matter of hours — or less.

Conclusion

These are exciting times for data analysts. Data warehouse technology is maturing, and the cost of storage is rapidly coming down. This has made many new data integration projects economically viable that previously were technically possible but not (yet) feasible from a business standpoint.

Web services are continuously expanding, which by itself is creating yet more data. This has resulted in exponential growth because of the internet, and more specifically social media. This explosion of so-called 'big data' offers huge opportunities along with serious challenges when we try to turn all these data into useful information. Traditional tools have been geared predominantly to analysing structured data, and not so much at the unstructured or semi-structured data we are facing today.

In this paper, we have identified four major streams of research data that have been shown to be worthwhile to pursue. All have a proven track record of analytic power. First, we have discussed data fusion, which can be used to augment existing data sources, or used to apply 'soft' measures from market research to proprietary in-house databases.

Secondly, we discussed best practices in direct marketing, which has been around for a while, and continues to play a strong commercial role. The main reason for this is that so many of these activities take place under voluntary control of corporations. They can 'design' their interactions not only with short-term commercial objectives, but also with long-term knowledge management in mind.

Thirdly, we discussed the usage of clickstream data, another relatively new avenue. Since you can collect these data 'for free' on the back of every web server, this has opened up enormous opportunities. Again, by systematically manipulating web content, site owners can use the browsing experience as a source of valuable learning and data generation.

Fourthly, and last, we discussed the rise of social media, and its impact on the research landscape. Although new and quirky, already some compelling evidence has been gathered regarding the predictive powers of these data and their applicability.

Researchers play a pivotal role in two distinct ways. They are in the driver's seat when it comes to extracting value from all these disparate data. And on top of that, they have an important responsibility in setting their research agendas, thereby guiding IT investments to advance available knowledge on how to capture and integrate these vast oceans of potential corporate wisdom.

References

1. IDC report. (2007) 'The diverse and exploding universe', available at <http://www.emc.com/collateral/analyst-reports/diverse-exploding-digital-universe.pdf>.
2. Davenport, T. and Harris, J. (2007) *Competing on Analytics: The New Science of Winning*, Harvard Business School Press, Boston, MA.
3. Redman, T. (2008) *Data Driven: Profiting from Your Most Important Business Assets*, Harvard Business School Press, Boston, MA.

Research drives both exploration and exploitation of value in data

4. Taylor, J. and Raden, N. (2007) *Smart Enough Systems: How to Deliver Competitive Advantage by Automating Hidden Decisions*, Pearson, Boston, MA.
5. Ayres, I. (2007) *Super Crunchers: Why Thinking-by-Numbers is the New Way to be Smart*, Bantam, New York, NY.
6. Radner, D., Rich, A., Gonzalez, M., Jabine, T. and Muller, H. (1980) 'Report on exact and statistical matching techniques'. Statistical working paper 5, Technical report, Office of Federal Statistical Policy and Standards US DoC.
7. van der Putten, P. (2010) 'On data mining in context: Cases, fusion and evaluation', (PhD thesis), available at http://www.liacs.nl/~putten/library/phdthesisv1_nocropmarks_withISBN.pdf.
8. Harris, P. and O'Brien, J. (1989) 'Data fusion: An appraisal and experimental evaluation', *Journal of Market Research Society*, Vol. 31, No. 2, pp. 153–212.
9. van Pelt, X. (2001) 'The fusion factory: A constrained data fusion approach'. Master's thesis, Leiden Institute of Advanced Computer Science, Leiden University, The Netherlands.
10. Raessler, S. (2002) *Statistical Matching: A Frequentist Theory, Practical Applications, and Alternative Bayesian Approaches*, Springer-Verlag, New York, NY.
11. D'Orazio, M., Zio, M.D. and Scanu, M. (2006) *Statistical Matching: Theory and Practice*, Wiley, Chichester, West Sussex, UK.
12. Leventhal, B. (1997) 'An approach to fusing market research with database marketing', *Journal of the Market Research Society*, Vol. 39, No. 4, pp. 545–558.
13. Baker, K. (2007) 'Development of the touch points planning database: An appraisal'. Prepared for the IPA.
14. Beeftink, B. and Mareck, M. (2010) 'IPA touchpoints', available at <http://www.cimm-us.org/CIMM%20UK%20Touchpoints%20Report%20Final%20Presentation.pdf>.
15. Geller, L. (1996) *Response! The Complete Guide to Profitable Direct Marketing*, Oxford University Press, New York, NY.
16. Bird, D. (2007) *Commonsense Direct & Digital Marketing*, 5th edn., Kogan Page, London, UK.
17. Breur, T. (2007) 'How to evaluate campaign response — The relative contribution of data mining models and marketing execution', *Journal of Targeting, Measurement and Analysis for marketing*, Vol. 15, No. 2, pp. 103–112.
18. Godin, S. (1999) *Permission Marketing: Turning Strangers into Friends and Friends into Customers*, Simon & Schuster, New York, NY.
19. Peppers, D. and Rogers, M. (1999) *Enterprise One to One*, Random House, New York, NY.
20. Lee, D. (2000) *The Customer Relationship Management Survival Guide*, HYM Press, St Paul, MN.
21. Cohen, J. (1988) *Statistical Power Analysis for the Behavioral Sciences*, 2nd edn., Lawrence Erlbaum Associates, Hillsdale, NJ.
22. Kohavi, R., Longbotham, R. and Walker, T. (2010) 'Online experiments: Practical lessons', *IEEE Computer*, Vol. 43, No. 9, pp. 82–85.
23. Manning, C. and Schuetze, H. (1999) *Foundations of Statistical Natural Language Processing*, MIT, Cambridge, MA.
24. Kao, A. and Poteet, S. (2010) *Natural Language Processing and Text Mining*, 2nd edn., Springer-Verlag, London, UK.
25. http://en.wikipedia.org/wiki/Leonhard_Euler.
26. Brin, S. and Page, L. (1998) 'The anatomy of a large-scale hypertextual web search engine', *Computer Networks and Isdn Systems*, Vol. 30, No. 2, pp. 107–117.
27. O'Connor, J. (2011) 'An empirical pilot event study of popularity and performance: How social media consumer brand fan count predicts stock price performance'. Available upon request from Famecount.com.
28. Bollen, J., Mao, H. and Pepe, A. (2011) 'Twitter mood predicts the stock market', *Journal of Computational Science*, Vol. 2, No. 1, pp. 1–8.
29. Asur, S. and Huberman, B. (2010) 'Predicting the future with social media', available at <http://arxiv.org/abs/1003.5699>.