Technovation 34 (2014) 203–214

Contents lists available at ScienceDirect

Technovation

journal homepage: www.elsevier.com/locate/technovation

The wisdom of crowds: The potential of online communities as a tool for data analysis

Marian Garcia Martinez [a,*], Bryn Walton [b]

[a] Kent Business School, University of Kent, UK
[b] Kent Business School, University of Kent, UK

ARTICLE INFO

Available online 20 February 2014

Keywords:
Crowdsourcing
Open innovation
Online communities
Creativity
Predictive modelling competition
Knowledge communities
Data analytics
Shopper behaviour

ABSTRACT

Online communities have become an important source for knowledge and new ideas. This paper considers the potential of crowdsourcing as a tool for data analysis to address the increasing problems faced by companies in trying to deal with "Big Data". By exposing the problem to a large number of participants proficient in different analytical techniques, crowd competitions can very quickly advance the technical frontier of what is possible using a given dataset. The empirical setting of the research is Kaggle, the world's leading online platform for data analytics, which operates as a knowledge broker between companies aiming to outsource predictive modelling competitions and a network of over 100,000 data scientists that compete to produce the best solutions. The paper follows an exploratory case study design and focuses on the efforts by Dunnhumby, the consumer insight company behind the success of the Tesco Clubcard, to find and lever the enormous potential of the collective brain to predict shopper behaviour. By adopting a crowdsourcing approach to data analysis, Dunnhumby were able to extract information from their own data that was previously unavailable to them. Significantly, crowdsourcing effectively enabled Dunnhumby to experiment with over 2000 modelling approaches to their data rather than relying on the traditional internal biases within their R&D units.

1. Introduction

Today's fast paced business environment requires firms to explore the use of external sources of ideas and technology to augment in-house R&D (for a review of the open literature refer to Dahlander and Gann, 2010; Huizingh, 2011; Lichtenthaler, 2011). Advances in information and communication technology have made companies more aware of externally generated scientific knowledge and they are starting to recognise that external ideas and R&D can create significant value (Gassmann and Enkel, 2004). As a result, the ratio of large firms' research to total research has shifted over the last 20 years (Slowinski et al., 2009). Start-up and small and medium-sized enterprises (SMEs) and even individuals now play a far more relevant role in knowledge and technology creation while large firms seek to collaborate with these smaller knowledge creators (Chesbrough, 2006).

Research on open innovation models has largely focused on innovation collaboration through organisational linkages and ecosystems with contributions from across a network of partners ranging from suppliers of raw materials, equipment and research institutes to consumers and customers that create value for the end consumer (West and Lakhani, 2008; Bianchi et al., 2011; Mortara and Minshall, 2011). More recently, interest has moved to consider co-creation interactions with consumer and unidentified individuals on the Internet in the process of product development and value creation (Howe, 2008; Surowiecki, 2005; Tseng and Piller, 2003; Poetz and Schreier, 2012; Marchi et al., 2011). High-quality interactions that enable consumers to co-create unique experiences with companies are the key to unlocking new sources of competitive advantage (Prahalad and Ramaswamy, 2004, p. 7). Co-creation through crowdsourcing in particular, a new form of inbound open innovation, (Huizingh, 2011) where companies or institutions take an idea or solution seeking process, traditionally performed by internal employees, and outsource it to an undefined, generally large group of people, referred to as the 'crowd', in the Internet is gaining attention (Arora and Gambardella, 2010; Natalicchio et al., 2014; Afuah and Tucci, 2012). Advances in Internet, collaboration tools, and new Web2.0 technologies have considerably reduced the costs of accessing large number of geographically-dispersed individuals with rather diverse expertise (Morgan and Wang, 2010; Terwiesch and Ulrich, 2009; Zwass, 2010).

Crowdsourcing is built on the concept that large groups of people are smarter than an elite few, no matter how brilliant the

* Corresponding author. Tel.: +44 1227 824113.
E-mail addresses: m.garcia@kent.ac.uk (M. Garcia Martinez), bcmw@kent.ac.uk (B. Walton).

http://dx.doi.org/10.1016/j.technovation.2014.01.011
0166-4972

elite few may be (Surowiecki, 2005, p. 1). The community character promotes the creativity and quality of contributions since participants from different backgrounds, with different areas of expertise, and different skills, experiences and perspectives can work together (Hargadon, 2003; Majchrzak et al., 2004). Howe (2006, p. 1) describes this new paradigm 'as everyday people using their spare cycles to create content, solve problems, even do corporate R&D'. These 'crowds' form around particular projects (e.g., solutions or ideas contests) and thus have common objectives (Hagel and Armstrong, 1997). Harnessing the wisdom of the crowd promises many advantages, such as eliminating biases that lead to unjustified preference for certain contributors and totally integrating market research and innovation by getting customers to submit own ideas and voting on each other's ideas (Surowiecki, 2005; Howe, 2006). Crowd contests are becoming a popular mechanism to obtain external knowledge with different levels of maturity and plugged into different points in the innovation funnel (Mortara et al., 2013). As Pike et al. (2005) highlight in their study, relational resources may constitute the core value drivers of organisations' R&D departments.

This paper aims to contribute to the emerging crowdsourcing literature by examining the wisdom of the crowd at solving problems, fostering innovation, proposing better – more innovative, more relevant – solutions, even predicting the future. Specifically, the paper considers the potential of crowdsourcing to solve predictive modelling problems, a particular form of crowdsourcing competitions featuring distinct personality traits in terms of participants' knowledge and motivation that until now, have received limited attention in the literature despite their potential to address the increasing problems faced by companies in trying to deal with "Big Data" (Manyika et al., 2011). Organisations are increasingly resorting to statistical/analytical outsourcing as a strategy to extract value from an increasingly turbulent, unstructured digital data environment. By exposing the problem to a large number of participants proficient in different techniques, crowdsourcing competitions can very quickly advance the technical frontier of what is possible using a given dataset (Boudreau and Lakhani, 2013).

Our empirical setting is the data analytics start-up Kaggle (www.kaggle.com), which operates as a knowledge broker between companies aiming to outsource predictive modelling competitions and a network of over 100,000 data scientists that compete to produce the best solutions. As the research object of predictive modelling competitions is new and the extant research limited, this paper follows the example of other studies in the open innovation and crowdsourcing literature (e.g. Chiaroni et al., 2011; Ebner et al., 2009; Leimeister et al., 2009) by adopting an exploratory case study design to provide a rich illustration of the phenomenon under analysis (Eisenhardt and Graebner, 2007) that requires the capability to answer the 'how' and 'why' questions (Yin, 2003). The study focuses on the efforts by Dunnhumby (www.dunnhumby.com), the consumer insight company behind the success of the Tesco Clubcard, to find and lever the enormous potential of the collective brain to predict shopper behaviour. In 2011, Dunnhumby ran an innovation contest through Kaggle, to predict the date and spend of the next visit for each customer in the modelling test set (www.kaggle.com/c/dunnhumbychallenge). By the closing date 537 different players had submitted 2029 different models, making the Dunnhumby shopper challenge at the time the most successful Kaggle competition ever launched. The winning solution improved on Dunnhumby's benchmark model by more than 100%.

Empirical studies on the benefits of open innovation have been largely based on high-technology industries (Bianchi et al., 2011; Todtling et al., 2006; Christensen et al., 2005; Ferrary, 2011), raising questions about the underlying openness decision process in mature and low-technology industries (Santamaria et al., 2009) and service businesses (Love et al., 2011). By observing a company whose primary operation is market analysis, this paper contributes to our understanding of open innovation and crowdsourcing as a strategic business tool outside of the commonly researched areas.

The paper proceeds as follows. Following the introduction, Section 2 introduces the idea of predictive modelling competitions and their attributes. Section 3 looks into the effective design and management of predictive modelling competitions that encourage sustained participation and submission of novel solutions. Crowd contests should be promoted, with rewards and opportunities to increase stature and recognition from peers, so they appeal to skilled data scientists and receive adequate attention from the crowd. In contrast to ideas competitions, the security and privacy implications of data mining competitions pose a challenge to the effective design of crowdsourcing contests. Section 4 details the empirical context and data used in the analysis. Section 5 presents the implementation and main characteristics of the shopper challenge competition. Section 6 discusses the key design and management aspects of data mining competitions and their potential as a tool for data analysis and Section 7 summarises our intended contribution, practical implications of our findings, and a future research agenda that takes into account the study's limitations.

2. Predictive modelling competitions

Predictive modelling is a process used in predictive analytics to create a statistical model based on future behaviour (Davenport and Harris, 2006). A predictive model is made up of a number of predictors, which are variable factors that are likely to influence future behaviour or results (Hair, 2007). Marketing is among the most frequent applications of the technique. Advances in information technology, data gathering and analytics are enabling companies to manage all phases of the customer life cycle, including acquiring new customers, increasing revenue from existing customers and retaining new customers. Retailing in particular is an area where predictive analytics shows the most promise (Harris and Lowitt, 2007). Using increasingly granular data from detailed demographics and psychographics to consumers' clickstreams on the web, retailers are creating highly customised offers that steer consumers to the right merchandise or service, the so-called 'next big offer' – at the right time, at the right price and in the right channel (Davenport et al., 2011). For instance, through its loyalty card programme, Tesco, the leading UK retailer, tracks which stores consumers visit, what they buy and how they pay. As a result, Tesco and Dunnhumby have been able to define 267 lifestyle clusters (Humby et al., 2003) allowing them to match offers to specific demographics and target promotional activity based on purchasing history (Rowley, 2007). This demographic and psychographic information about their customers has also allowed Tesco successfully to target coupon offers resulting in the achievement of redemption rates ranging from 8% to 10%, far higher than the 1–2% seen elsewhere in the grocery industry (Davenport et al., 2011).

Predictive modelling competitions extract 'core' predictive modelling challenges from their application environment which are then presented to participants. They are usually expected to represent real-life problems important enough to warrant dedicated experimentation (Boudreau and Lakhani, 2013). For instance, the Netflix Prize challenged the computer science communities to develop predictive systems for movies preferences (Bennett and Lanning, 2007). The Knowledge Discovery and Data Mining (KDD) Challenge Cup, the oldest data mining competition, running since 1997, focuses on analysing different datasets in search, bioinformatics, web mining