Automated competitor analysis using big data analytics

Evidence from the fitness mobile app business

Liang Guo
Neoma Business School, Mont Saint Aignan, France
Ruchi Sharma
School of Strategy and Leadership, Neoma Business School, Mont Saint Aignan, France
Lei Yin
Neoma Business School, Mont Saint Aignan, France
Ruodan Lu
University of Cambridge, Cambridge, UK, and
Ke Rong
Business School, Bournemouth University, Bournemouth, UK

Received 14 May 2015
Revised 30 June 2015
Accepted 17 August 2015

Abstract

Purpose – Competitor analysis is a key component in operations management. Most business decisions are rooted in the analysis of rival products inferred from market structure. Relative to more traditional competitor analysis methods, the purpose of this paper is to provide operations managers with an innovative tool to monitor a firm's market position and competitors in real time at higher resolution and lower cost than more traditional competitor analysis methods.

Design/methodology/approach – The authors combine the techniques of Web Crawler, Natural Language Processing and Machine Learning algorithms with data visualization to develop a big data competitor-analysis system that informs operations managers about competitors and meaningful relationships among them. The authors illustrate the approach using the fitness mobile app business.

Findings – The study shows that the system supports operational decision making both descriptively and prescriptively. In particular, the innovative probabilistic topic modeling algorithm combined with conventional multidimensional scaling, product feature comparison and market structure analyses reveal an app's position in relation to its peers. The authors also develop a user segment overlapping index based on user's social media data. The authors combine this new index with the product functionality similarity index to map indirect and direct competitors with and without user lock-in.

Originality/value – The approach improves on previous approaches by fully automating information extraction from multiple online sources. The authors believe this is the first system of its kind. With limited human intervention, the methodology can easily be adapted to different settings, giving quicker, more reliable real-time results. The approach is also cost effective for market analysis projects covering different data sources.

Keywords Mobile apps, Big data, Naïve Bayes, Operational strategy, Probabilistic topic modelling, User segment overlapping

Paper type Research paper

Introduction

Firms today face constant pressure to maintain sustainable growth, stay ahead of their competitors, and present superior customer-centric products. It is impossible for any firm to adequately survive, without developing a thorough market perspective. One of the tools for gaining the market insight is by developing the right competitive intelligence that can have a

This research is supported by "the 'Innovation Method Fund of China' from Ministry of Science and Technology of China (MOST) under Project No. 2016IM010200", the National Natural Science Foundation of China (Grant No. 71402051), and "Cyrus Tang Young Scholar Awards". All errors are of the authors.

Business Process Management Journal
Vol. 23 No. 3, 2017
pp. 735-762
© Emerald Publishing Limited
1463-7154
DOI 10.1108/BPMJ-05-2015-0065

far reaching strategic impact on a firm's operations strategy and business process management. Following Amoako-Gyampah and Boye (2001, p. 59) in this paper, we examine the role of competitive acumen in designing operational-level strategy and business processes for business sustainability. Competitor analysis, a set of methods to assess the strengths and weaknesses of current and potential competitors, is a key task for operations managers as they scan their competitive terrain, attempt to understand their market structure, shore up their defenses against likely competitive incursions, improve their business process of core activities, and plan competitive attack and response strategies (Aho and Uden, 2013; Allen and Helms, 2006; Bergen and Peteraf, 2002; Espino-Rodríguez and Rodríguez-Díaz, 2014). These methods to collect competitor information and draw inferences have been the lifeblood of operations managers and the focus of much academic research in the management literature (e.g. Grossler and Grübner, 2006; Hamel and Prahalad, 2005; Porac and Thomas, 1990; Porter, 1980, 1985; Zajac and Bazerman, 1991). It informs operations managers with product design strategies for the price sensitive audience, rich or simple product assortments or the differentiated mix of both. These have direct implications performance on a firm's operations in terms of better quality, lower cost and flexibility in adapting to changing market trends (de Waal and Batenburg, 2014; Shamsuzzoha, 2011). Finally, operations managers will have to constantly refer to the updated competitive intelligence to re-engineer their product strategies according to changing market perspectives (Wieland *et al.*, 2015). Previous research indicates that competitor analysis helps firms appreciate interactive market behavior, understand firm rivalry, strategize for superior competitive gains (Caves, 1984; Porter, 1980; Scherer and Ross, 1990), and improve their assessment of competitors' competencies and the threats these represent (Zajac and Bazerman, 1991).

Competitor analysis is a multi-disciplinary function affecting sales, marketing, product development, operations strategy, and product re-engineering. It requires diverse information spanning these departments. In the past, with fewer available sources of information, conducting competitor analysis was a difficult activity, limiting the power of advanced business analytics, such as conjoint analysis (Green and Srinivasan, 1978), multidimensional scaling (MDS) mapping (Elrod, 1988, 1991; Elrod *et al.*, 2002), market clustering (DeSarbo *et al.*, 1991), and "voice of consumer" analysis (Griffin and Hauser, 1993).

The advent of the internet has led to superior methods of information collection and analysis (Lee and Bradlow, 2011). However, conducting such analyses to monitor competitors can be a time-consuming process, with a vast amount of information available. Even though the content on the web is enormously helpful, it does present difficulties. With overwhelmingly large quantities of data, the task of constantly tracking and detecting new sources of information and then assimilating the knowledge from various online sources would be a big challenge. Another more severe constraint is imposed by the unstructured set of online data that is primarily qualitative in nature and full of noise. Godes *et al.* (2005) note that one of the challenges in using online content is the impossibility of automatically analyzing textual information. Very often, expert systems, based on the human interpretation of knowledge, are used to extract wisdom (Jayaraman and Srivastava, 1996). However, these methods are expensive, and due to a lack of automated tools, such data sources often remain unused.

Recently, with the increasing capabilities of big data analytics, managers are looking to collect market information from vast pools of data and automatically analyze it to search for meaningful knowledge (Feldman *et al.*, 2010). This is especially useful for the extraction and coding of vast amounts of unstructured data in the form of text content, such as product descriptions, expert reviews, blogs, customer reviews, employee testimonials, investor reports, and media news (Doan *et al.*, 2011), that have previously needed a great deal of human intervention, such as hand-coded rules and keyword-based searches, to extract information from web-based text (Shi and Yu, 2013). In this scenario, modern machine learning methods like statistical natural language processing (NLP) with fewer context-specific rules tailored to specific domains can be useful to improve knowledge relationships (Netzer *et al.*, 2012). Some terminologies similar to NLP include content analytics, text mining, ontology induction, concept hierarchy, and so on, all of which use the same methodology of grouping terms into concepts and of identifying different types of relationships between concepts.

Our objective is to harness the growing body of free online content for automated competitor analysis that does not rely on a predefined set of language rules or ex-post interpretation of derived dimensions from consumer surveys. We intend to provide operations managers with a cost-efficient tool to monitor a firm's market position in real time with higher resolution and at lower cost than traditional methods. We combine the techniques of Web Crawler, Naïve Bayes, Latent Dirichlet Allocation (LDA) Topic Modeling, MDS, K-Nearest Neighbors (k-NN) Clustering with data visualization to develop a big data competitor analysis system that can inform operations managers about competitors and the meaningful relationships among them. We illustrate our approach, using the fitness mobile app business. We develop a big data system to collect structured as well as unstructured data from multiple sources, to analyze online content automatically, and to discover competitor knowledge.

We contribute to the operations management literature by presenting a novel big data methodology with management theories to expand the traditional scope of competitor analysis. Our work complements previous studies in at least four ways. First, we present a simpler, automatic and easily replicable method when compared to existing methods. Our method analyses producers' self-provided product descriptions and users' social media data while requiring minimal human intervention. Second, our method is truly "big data" based – large volume (i.e. 1,381 mobile apps; 100,892 user comments; and 95,705 Google Plus user profiles), various formats (i.e. numerical and textual data) and from various sources, which provides more information than traditional market structure analysis and advances existing competitor analysis methods. Third, our approach can automatically map out market positions of similar products, visualize product attributes and reveal customer segmentation. Thus, our method not only facilitates direct competitor analysis, but also identifies indirect as well potential competitors. Finally, our method not only captures the attributes that a producer should emphasize in a product's description, but also tracks subtle differences in vocabulary that may separate brands or identify unique submarkets. These include attributes that are not highlighted by the use of more traditional text mining methods to elicit attributes and dimensions (Lee and Bradlow, 2011; West *et al.*, 1996).

In the following section, we describe the current state of research with respect to competitor analysis, big data analytics and NLP. Thereafter, we describe our methodology and apply it to the fitness mobile app business. We conclude with a discussion of the potential of our big data competitive analysis system, its limitations, and directions for future research.

Literature review

Competitor analysis

In order to remain competitive, it is essential for operations managers to have a clear understanding of their firm's competitors (Calori *et al.*, 1994; Hodgkinson and Johnson, 1994; Porter, 1980) and make right operations strategies that reflect the planning, design and implementation of strategic decisions that span across business processes of a firm (Barnes, 2001; Englyst, 2003; Minarro-Viseras *et al.*, 2005; Miller and Roth, 1994; Paiva *et al.*, 2008; Riis *et al.*, 2006; Slack and Lewis, 2002; Rytter *et al.*, 2007; Ward *et al.*, 1996) and strengthen a firm's competitiveness in the market through improved quality, better delivery, lower cost and enhanced market adaption flexibility (Alegre-Vidal *et al.*, 2004; Amoako-Gyampah and Meredith, 2007; Boyer, 1998; Boyer and Pagell, 2000; Boyer and McDermott, 1999; Christiansen *et al.*, 2003; Dangayach and Deshmukh, 2001; Diaz *et al.*, 2005; Flynn and Flynn, 2004;

Grobler and Grubner, 2006; Hayes and Upton, 1998; Hayes and Wheelwright, 1984; Lawrence, 2008; Martin-Pena and Diaz-Garrido, 2008; Miller and Roth, 1994; Rosenzweig and Easton, 2010; Sum *et al.*, 2004; Ward *et al.*, 1998; Yu and Ramanathan, 2009), quicker design of products (Martin-Pena and Diaz-Garrido, 2008).

In essence, competition between firms refers to the rivalry between their respective business lines at both product- and firm-level. Competitor analysis, also called "competitors' acumen" (Tsai *et al.*, 2011), provides valuable competitive intelligence by creating an accurate, strategic method of understanding competitor's operations (Bloodgood and Bauerschmidt, 2002) and thus becomes a driver of competitive success (Lamb, 1984). Porter (1980) argues that competition between the firms can be classified based on customer offerings that differ in terms of specific functions and associated ease of use, technology, the raw materials used in producing the product and the market segment being catered for. Czepiel and Kerin (2012) further suggest that competitors can be classified into three categories, namely, direct, indirect, and potential. Knowledge about direct competition is a "must have" for any firm when building its competitive intelligence (Ulrich and Eppinger, 2003). It is also the fiercest form of competition that exists between firms with hardly any differentiation between their product offerings. Firms can become indirect competitors in a given business domain, if they serve the same customer needs but with different resources. Finally, there are the potential competitors who do not serve the same customer base but use the same resource base or have equivalent capability (Czepiel and Kerin, 2012). In short, operations managers need to clearly understand the extent of competition that exists in their domain and benchmark the various types of competitors in the market.

Big data analytics and NLP

Following a data-centric approach to the development of business intelligence, big data analytics can significantly facilitate competitor analysis. From a business-function perspective, big data analytics can be described as a set of analytical techniques used on relatively large data sets and involving complicated digital data-gathering sources (Chen *et al.*, 2012). From an information systems perspective, big data analytics can be viewed as a system with extraordinary data integration and warehousing capabilities, used for information extraction, online analytical processing, and reporting, and based on intuitive data mining, statistical analysis and predictive analytics (Goes, 2014; Schlegel, 2014). Recent related research and application topics in big data analytics include text analytics, web analytics, network analytics, mobile analytics, social media analytics, and sentiment analysis on qualitative data (Chen *et al.*, 2012; Pang and Lee, 2008). All of these are based on NLP techniques to process semi-structured text data, to extract meaningful, relevant and non-trivial information and to discover business knowledge from huge amounts of online content (Dörre *et al.*, 1999; Feldman and Sanger, 2006; Lee, 2007).

NLP is a computational approach to text analysis that originated in the 1950s with the convergence of artificial intelligence and linguistics (Jurafsky and Martin, 2009). Creating fundamental algorithms and mathematical models for human language processing, NLP can be considered as an advanced type of text mining technique used for content analysis, that relies more on complex language processing and less on hand-coded parsing rules, hence requiring minimal human intervention only in the analysis phase (Hu and Liu, 2004; Lee, 2005). NLP forms part of the computational linguistics stream of text analytics that can perform lexical acquisition, word sense disambiguation, part-of-speech-tagging, and probabilistic context-free grammars in a given unstructured text context (Manning and Schutze, 1999). Current approaches to NLP are based on probabilistic machine learning and knowledge engineering methods that examine and use patterns in data to improve a program's own understanding (Krauthammer and Nenadic, 2004).

Therefore, many statistical NLP models are able to deliver relatively accurate analytics results and close to human interpretation (Jurafsky and Martin, 2009).

Some applications of text mining and NLP in the management literature include analysis of the relationship between product attributes and sales (Archak *et al.*, 2011); hotel-room demand estimation based on text mining (Ghose *et al.*, 2012); corporate stock performance, by mining the text of sentiment and star ratings of product reviews (Seshadri and Tellis, 2012); ascertaining consumer preferences for products through user comments (Decker and Trusov, 2010); the determination of relationships and predictive analytics of demand in response to price changes (Archak *et al.*, 2007; Ghose and Ipeirotis, 2008; Kamakura and Russell, 1989); marketing strategy (Erdem and Keane, 1996); and new product design (Srivastava *et al.*, 1984). In the area of competitor analysis, Pant and Sheng (2009) mine corporate-level text data, using network linkages between web pages and online news to identify market competitors. Feldman *et al.* (2007, 2008) apply NLP to extract and visualize relationships between product brands from online blogs. Lusch *et al.* (2010) extract market information by analyzing the "conversations" between firms and customers. More recently, Lee and Bradlow (2011) and Netzer *et al.* (2012) improve understanding of market structure by text mining semi-structured product attributes. However, most previous studies still use manual rules and human tagging to understand complex linguistic patterns. Also it is noteworthy that most prior studies consider only "the voice of the customer", which may not be reliable as the sole reference point for product- or firm-level competitor analysis. In our study, we improve on prior work by fully automating information extraction from multiple online sources. We believe this is the first system of its kind. Our methodology can easily be adapted to multiple settings and scenarios, while yielding quicker, reliable, real-time results. By limiting the need for human intervention, our system is cost effective for market analysis projects covering different data sources. In addition, the availability of various types of free online content, such as the continuous stream of expert reviews, product descriptions, user comments, and social media user profiles, provides a practical reason to augment traditional methods (such as surveys and focus groups) of conducting competitor analysis, which can be used continuously, automatically, inexpensively, and in real time. Thus, this study intends to introduce a fast, inexpensive big data analytics system that can have a significant impact on operational decisions. In the remainder of this paper, we describe our methodology, apply our approach to 1,381 health and fitness mobile Apps, and demonstrate how our approach helps operation mangers in conducting competitor analysis.

Methodology

In this study, we classify competitors by applying binary Naïve Bayes, LDA topic modeling and k-NN Clustering, and conduct competitor analysis by applying various data tabulation and visualization techniques. To do so, we develop a big data competitor analysis system (see Figure 1) that can specifically deal with the difficulties involved in collecting data from multiple sources and mining unstructured textual information. The system is divided into three parts: data collection, data classification and data analyses. We first developed three web crawler programs that collected the data from three websites – Google Play app store for the producer self-provided descriptions of the applications within the health-fitness category and the user comments of each application; Google Plus for the demographic information of each user that has commented at least one application within the health-fitness category; and Factiva for the expert reviews of each application in our sample. Then we conducted NLP-based classification analyses to eliminate the non-fitness applications from our sample and extracted ten categories of application functionalities as well as four clusters of applications. Finally, a series of competitor analysis metrics were developed and applied to examine the fitness mobile application market.

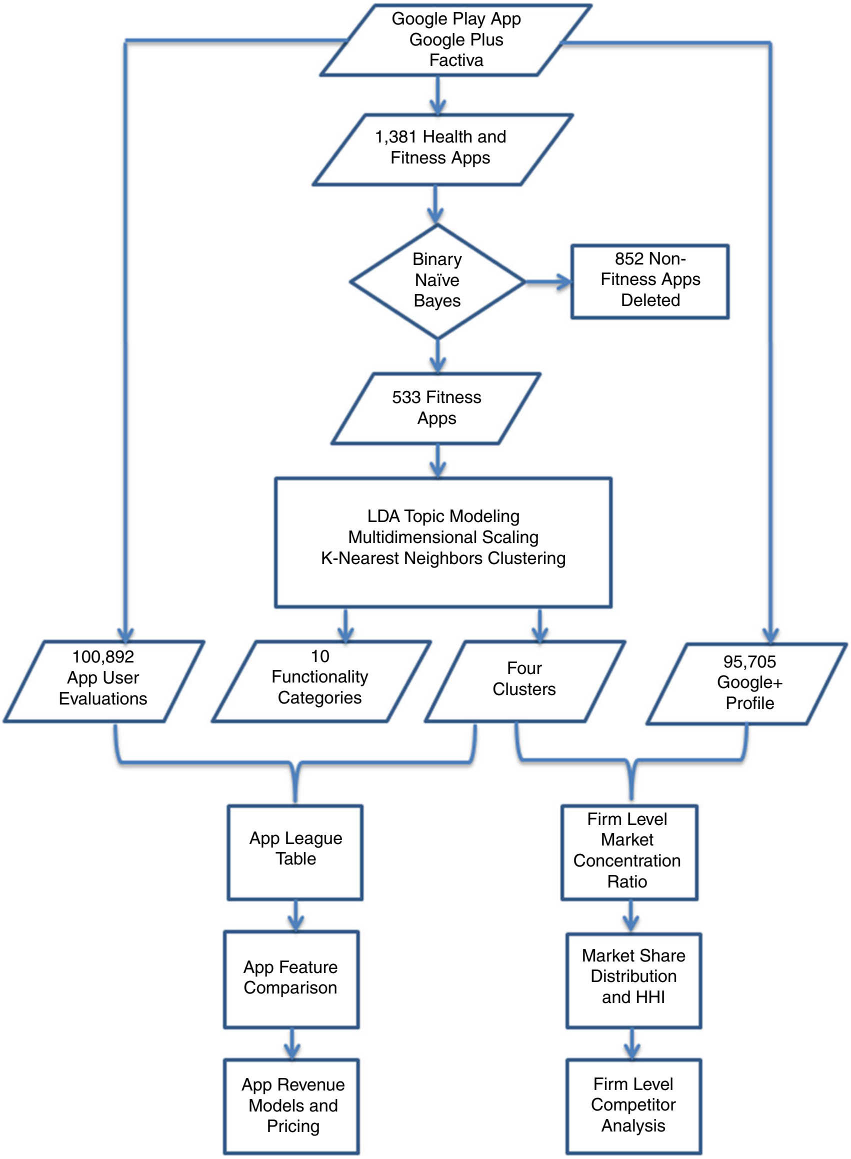

Google Play App
Google Plus
Factiva
1,381 Health and Fitness Apps
Binary Naïve Bayes
852 Non-Fitness Apps Deleted
533 Fitness Apps
LDA Topic Modeling
Multidimensional Scaling
K-Nearest Neighbors Clustering
100,892 App User Evaluations
10 Functionality Categories
Four Clusters
95,705 Google+ Profile
App League Table
App Feature Comparison
App Revenue Models and Pricing
Firm Level Market Concentration Ratio
Market Share Distribution and HHI
Firm Level Competitor Analysis

Figure 1.
The big data competitor analysis apparatus

We do not focus on the fitness mobile app business by chance, but rather due to its popularity and its fit with big data analytics of freely available online content. The "anything, anytime, anywhere" mobile sports app gold rush is on, with an unprecedented number of firms engaged in methods to best monetize the high-value touch points between their apps and users (Flurry, 2014). However, the fitness app's pick-and-shovel business is not thriving. Although a few perceptive mobile application companies are reaping significant gains, the vast majority of firms find themselves with little revenue. Therefore, competitor analysis is a key step in designing and developing new mobile apps, as well as in repositioning existing ones in the mobile app business. It can help managers understand the substitution and complementary relationships between the brands and alternatives that define the market, predict marketplace responses to changes, and make appropriate operational decisions.

Data collection

We used a set of web crawler programs to collect data from the following websites: Google Play app store and developer console (play.google.com, under the category "health and fitness") from which we collect each app's name, developer firm's name, number of downloads, user evaluation (from 1 to 5), self-provided description, price, and in-app purchase items; Google Plus from which we collect each user's recommendation and/or comments for each app and profile information; and The Factiva database (www.factiva.com, a database by Dow Jones and Company), that aggregates content from more than 32,000 sources such as newspapers, journals, magazines, television and radio transcripts, and so on from which we collect fitness app expert reports. Up to December 31, 2014, we collected data about 1,381 health and fitness apps available on Google's Play Store (US market) and 26 fitness app expert reports. We then deleted HTML tags and non-textual information such as photos, icons, and videos. We also removed punctuation marks and stop words based on NLTK's Stop words Corpus (Bird *et al.*, 2009). Finally, all capitalized words were converted to lowercase.

Data classification

Naïve bayes classification to eliminate non-fitness apps. The first task of competitor analysis is to identify rivals that offer similar products/services and compete head-on in the marketplace. However, Google Play's "health and fitness" category lists apps for nutrition, sleep, and healthy lifestyle along with those for sports, fitness and workout. Therefore, we selected only fitness related apps from our sample, using a binary naïve Bayes algorithm.

Naive Bayes is based on probabilistic models using Bayes' theorem to create classification tasks, which makes it possible to predict for uncertain situations (probabilities) based on prior knowledge (probability values) (Wang *et al.*, 2011). Being the simplest of the Bayesian classifiers, Naïve Bayes follows a structural model assuming conditional independence for all instance attributes in a given class. It is applied to the learning of linear functions alone using binary values (Duda and Hart, 1973). It assumes that every independent attribute in the given class is equally important. The advantage of the above assumption is that the model has the flexibility to estimate each attribute separately. Prior research has found that this simplified assumption works well for textual information classification, having "text" as the unit of analysis (Chen *et al.*, 2009; Koller and Sahami, 1997; Li, 2010; McCallum and Nigam, 1998; Rennie *et al.*, 2003; Sahami, 1996; Yu *et al.*, 2013). Friedman (1997) and Domingos and Pazzani (1997) argue that Naive Bayes is an excellent classification method because it has high-classification accuracy, is easy to implement and is relatively effective in text classification tasks. Therefore, we chose the binary Naïve Bayes as our classifier, as our sample data have a large number of attributes proportional to large vocabularies. The classifier works by transforming an app's self-provided description into a list of strings and then to a feature (i.e. word) vector level. The classifier then calculates the prior probability of each class, which is determined by checking the frequency of each class in the training set. Every feature helps determine which class should be assigned to a given input value. The contribution from each feature is then combined with this prior probability to arrive at a likelihood estimate for each class and to choose the class that has the highest value.

In our binary classification tasks, the binary set of classes was defined in advance – 1 for fitness apps and 0 for other types. The objective of this binary classification was to choose the correct class label (1 or 0) for a given app. The success of this machine learning method lies in selecting relevant features and encoding them. Typically, feature extractors are built through a trial-and-error process, guided by intuition based on relevant information (Bird *et al.*, 2009). We established a panel of five IT and mobile business experts to develop a binary feature classifier. First, the panel carefully read the self-provided descriptions of five popular fitness apps featured on the app intelligence website AppAnnie (www.appannie.com), including Runkeeper, Nike+ Runner, Endomondo Sports Tracker, Noom Coach, and Runtastic Six Pack Abs Workout. The panel then used these five apps for three weeks. They selected 92 keywords from the descriptions of these five apps. Next, the panel read 26 expert reports on the fitness apps business from leading online publications specializing in IT and e-Business such as CNET, PCMag, TechRadar, ZDNet, and TechCrunch. They identified a set of 179 keywords (121 unigrams, 39 bigrams and 19 trigrams) and merged them with the previous 92 keywords to form a set of relevant product features of a typical fitness app.

Once an initial set of features was chosen, the panel manually annotated and classified 100 randomly selected apps from the 1,381 sample apps into two categories – fitness app (i.e. class 1) or non-fitness app (i.e. class 0). These 100 apps were randomly split into three sets: the training set (20 apps) that was used to train the classifier model with the trial-and-error approach by classifying the training sample into the two classes with the initial set of features; the development test set (30 apps) to identify errors and rebuild the set of features; and the test set (50 apps) to validate the classifier model. Our binary Naïve Bayes classifier began by calculating the prior probability of each category (i.e. the frequency of each class in the training set). The contribution from each feature was then combined with this prior probability in order to arrive at a likelihood estimate for each class. An app was then classified as a fitness app if the class 1likelihood estimate was higher than that of the class 0.

We compared the results of the automatic classifier model with those of the manual classification performed by the panel. We then revised and retested the set of features with the development test set. This generated a list of the errors that the classifier made when predicting whether an app focuses on fitness or not. We examined individual error cases where the model predicted the wrong category to determine what additional pieces of information would enable the classifier to make the right decision, or which existing pieces of information lead to a wrong decision. We then adjusted the product feature set accordingly.

We used the following metrics (Jurafsky and Martin, 2009; Salton and McGill, 1983) to evaluate the binary Naïve Bayes classifier in the test set by comparing the classes that it generated with the correct classes manually classified by the panel:

(1) Accuracy, the overall correctness of the classifier, was calculated as the sum of correct classifications divided by the total number of classifications. The accuracy value of our classifier was 0.801.

(2) Precision, the number of correct classifications made for each class divided by the total number of classifications predicted by the specific class was estimated as 0.863.

(3) Recall (also called sensitivity), a measure of the ability of a prediction model to select instances of a certain class from a data set, was calculated as the number of correct classifications made for each class divided by the total number of test examples of the considered class. The recall measure of our classifier was 0.791.

(4) The *F*-measure (or *F*-score), which combines precision and recall to give a single score, is defined as the harmonic mean of precision and recall (i.e. $2 \times \text{Precision} \times \text{Recall}/(\text{Precision} + \text{Recall})$). The *F*-value of our classifier was 0.825.

These four metrics indicated that the performance of our binary Naïve Bayes classifier was quite satisfactory. We then employed this classifier to select fitness apps from the whole sample, resulting in 846 non-fitness apps being eliminated. Finally, the panel revised the list of keywords and followed the approach of Ulrich and Eppinger (2003) and Lee and Bradlow (2011) to group them into ten categories of product functionality representing the constituent dimensions of a fitness app (keywords in brackets) : walking tracker (e.g. walk, pedometer, step counter), running tracker (e.g. run, jogging, hill run), biking tracker (e.g. ride, off-road, speedometer), gym workout and aerobic dance (e.g. coaching, abdominals, Pilates), health metrics (e.g. heart rate, blood pressure, calorie counter), challenger and motivator (e.g. cheering, chase, leader board), navigation (e.g. GPS, maps, turn-by-turn), artificial intelligence algorithms (e.g. auto stop, automatic calibration, voiceover feedback), wearable accessory (e.g. smartwatch, wristband, compatibilities), and yoga and weight control (e.g. yoga, fat loss, weight recorder). These ten functional categories of the product will be used to examine within-cluster rivalry in the subsequent sections.

LDA topic modeling, MDS and k-NN clustering to identify potential competitors. For the remaining 535 fitness apps, the goal of competitor analysis was to identify three types of competitors (Czepiel and Kerin, 2012): direct (i.e. directly competing with other firms to serve the same customer needs using the same resources), indirect (i.e. serve similar customer needs but with different resources) and potential (i.e. serve the similar or same customer base but have different resources or capability). To do so, we plotted the market positions of the apps based on the scales of differing purchase needs and place them in different clusters based on their functionalities.

We first applied probabilistic topic modeling (PCM) and MDS to conduct a market position mapping analysis. PCM is a type of machine learning method that makes it possible to explore documents automatically based on themes that run through a collection or corpus (Blei, 2012). PCM makes use of probability-based algorithms to detect "thematic structure" based on the topic distribution in large pools of online documents and depicts documents as a bundle of these topics (Blei *et al.*, 2003; Griffiths and Steyvers, 2002, 2003, 2004; Hofmann, 1999, 2001). A topic is derived from a probability distribution over a fixed vocabulary from a particular subject (Blei, 2012). PCM is a powerful method for automatically organizing, understanding, searching, and summarizing extensive electronic textual data. It offers an alternative method of conducting content analysis and retrieval that is often based on the keyword-based search method. It can also detect the relationship between different documents in a given collection based on the topics or themes that run through these documents (Vulić *et al.*, 2015).

We adopted the LDA PCM algorithm, which uses a statistical approach based on probabilities to detect the themes that run through a collection of documents, assuming that one document can have multiple topics (Blei, 2012). LDA is similar to principal component analysis for discrete data (Buntine, 2002; Buntine and Jakulin, 2005) based on statistical assumptions about the corpus and topics. LDA relaxes the sequence of words condition by considering a document as only a "bag of words" (Blei, 2012) and uses the order of words to generate topics (Wallach, 2006). In addition, LDA also uses a generative probabilistic model based on the hidden structures for generating words and thus does not require pre-annotated documents (Blei, 2012).

We followed the commonly used "kitchen sink" approach (Bird *et al.*, 2009) of treating each app's self-provided description as a topic to create a wide assortment of product features. The distribution of words in the description forms a topic's vocabulary of words (Blei, 2012; Steyvers and Griffiths, 2007). Each app's description was considered as an observable element. The topic distribution per description and the distribution of words per topic per distribution were considered as hidden structures. Using probability distributions, we constructed the hidden thematic structures that were assumed to resemble the thematic structure of the collection and hence, we calculated a bivariate-topic (i.e. between two apps) functionality similarity index (i.e. an indicator measuring how similar two apps were) using the algorithm of Blei *et al.* (2003). Assuming that each app represents a topic, a similarity matrix of 535 × 535 was generated.

We then employed a traditional market structure analysis and visualization tool – MDS to generate product-level market position mapping of the 535 fitness apps. MDS is a means of visualizing the level of similarity of individual products, by plotting the products on a 2D map. To do so, we followed Netzer *et al.* (2012), and applied principal component analysis to reduce the dimensions from 535 to 2. The result of eigenvalue analysis indicated that 31.69 percent of total variance of the similarity matrix was explained by the two factors. We plotted the market positions of the 535 fitness apps in a MDS in Figure 2, which mapped their relative positioning differentiated on their features. This figure was instrumental to our understanding of the classification of these apps in terms of various physical activities and smartphone sensors. The *y*-axis measured physical activity intensity ranging from light to vigorous exercise-oriented workouts. Physical activity intensity refers to the amount of physical power that the body uses to perform a physical activity. There are different ways to measure physical activity intensity using various electronic sensors. With the specialized electronic sensors available, mobile apps can now monitor body movement, heart rate, respiratory rate, blood pressure, blood sugar, and so on. Hence, the *x*-axis represents the ability and means adopted by an app to record vital data using built-in or external sensors such as wristband and smartwatch.

We observed a gradual increment in the intensity of physical activity as we move in an anti-clockwise direction starting from the third quadrant. The apps in this quadrant act as training guides for low intensity yoga exercises. These can also be synchronized with external wearable sensors (e.g. heart rate monitors and wristbands), to collect health metrics and track body positions. The apps in the lower-left portion of the fourth quadrant deal with moderately intensive aerobic gymnastics and aerobic dancing activities for women users. The apps in the upper-right portion of the fourth quadrant and the lower-right portion of the first quadrant focus on unisex intensive indoor physical workout activities. Finally, the apps in the second quadrant and the upper-left portion of the first quadrant are the outdoor sports trackers (e.g. walking, running, hiking, and biking) with built-in smartphone sensors or wearable widgets or both.

To group these 535 fitness apps accurately in such a way that apps in the same cluster are more similar to each other than to those in other clusters, we followed Netzer *et al.* (2012) and used the two-dimension coordinates of the MDS to run a k-NN cluster analysis. k-NN is one of the most popular supervised machine learning algorithms (Adeniyi *et al.*, 2014;

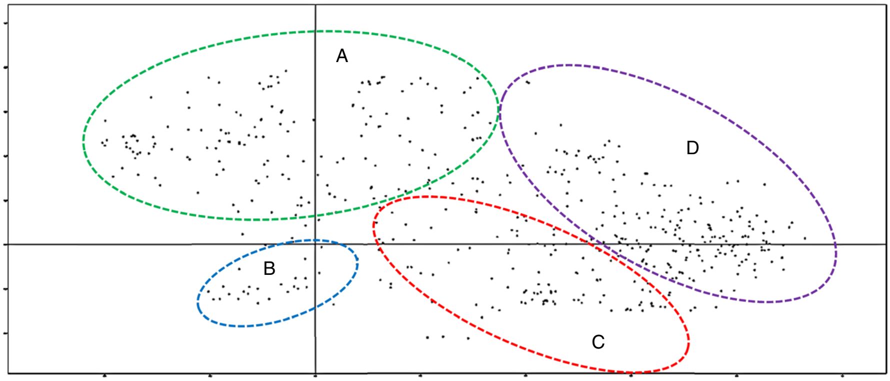


Figure 2.
The market positioning mapping of the 535 fitness apps

Chen *et al.*, 2012; Przewozniczek *et al.*, 2010;Weiss *et al.*, 2010; Wu *et al.*, 2007). It is based on the assumption that things that look alike must be alike (Cover and Hart, 1982). In general, k-NN offers high-accuracy classification with no prior assumptions about the data and is not sensitive to outliers (Cheng *et al.*, 2008; Han *et al.*, 2001; Kacur *et al.*, 2011; Ogbonaya, 2008; Xiao-peng and Xiao-gao, 2007). We found that four clusters could best fit the data (see the dashed ovals in Figure 2):

(1) Cluster A (156 apps): the apps in this cluster are used for tracking sport activities such as running, biking, and walking, and for providing users with vital data such as heart rate, steps, and distance. These apps predominantly target users who engage in physical activities (especially outdoor) on a regular basis and rely not only on internal smartphone sensors (e.g. GPS antenna, accelerometers and gyroscope), but also on external widgets (e.g. wristband, smartwatch, heart rate monitor) for performance measurement. Many apps in this cluster are interoperable with a wide variety of external third-party widgets and permit activity performance metrics data sharing with social media websites.

(2) Cluster B (36 apps): the apps in this cluster act more like personal yoga training guides. The uniqueness of this cluster lies in the way the apps offer an alternative fitness paradigm based on very low intensity yoga exercises. These apps can be connected with external wearable widgets and help users to adjust their body movement. Most of these apps provide animation or video guidance modules.

(3) Cluster C (73 apps): the apps in this cluster target the health and fitness of women through aerobics, gymnastics, and dancing (e.g. Belly, Pilates, Salsa, and Zimba) training tutorials. Their main purpose is to assist with exercises through videos, with or without the aid of exercise equipment based on varying level of workout intensities. Unlike those in cluster B, these apps are self-supported fitness assistants needing no connection with external wearable widgets and do not require any geo-positional data. They are specifically suited for women's indoor physical training programs.

(4) Cluster D (270 apps): the apps in this cluster target the health and fitness segment by providing challenging and motivational exercise video tutorials and a timer/logger for general unisex vigorous physical activities (e.g. calisthenics, strength, rowing machine, and basketball). The apps act as video tutorials and usually track a user's body movement using the smartphone's built-in gyroscope, accelerometer, and magnetometer. These apps usually do not use GPS sensors to record geo-location information as they mainly deal with indoor physical training programs. Finally, most apps in this cluster synchronize fitness performance data with social media websites.

Following Czepiel and Kerin's (2012) classification of competitors, we regard that apps within the same cluster are direct and indirect competitors as they address the same or similar customer needs with the same or similar resources. The apps in other clusters are usually potential competitors as they address the same customer base (i.e. those who engage in fitness activities) but with different resources or capabilities.

Direct and indirect competitor analyses

Product-level competitor analyses. In order to identify market leaders who usually are the most powerful rivals for many products in the same cluster, we first conducted within-cluster ranking with three different metrics:

(1) Number of downloads, which reflects the total of user downloads as obtained from Google Play app store (US market). However, this number does not comprehensively reflect the actual or perceived (by users) performance of the app, especially for free

apps that can be easily discarded by users. This metric can be used to calculate market share (i.e. number of downloads of an app divided by the total number of downloads of all apps in the same cluster).

(2) Google Plus recommendation, derived from the number of users recommending an app on their Google Plus webpage. This metric infers the overall popularity of an app, but cannot be considered to reflect the general user experience, as it only indicates satisfied users.

(3) User satisfaction, which factors in both the positive and negative feedback of an app from users on a scale of 1 (very unsatisfied) to 5 (very satisfied). We weighted the average score by the number of downloads of each app to overcome the disparity in the number of downloads across popular/old and unpopular/new apps. This measure provides a better, unbiased performance indicator based on users' positive and negative feedback.

Table I summarizes the rankings using these three metrics. In cluster A, we observed that even though the free app ANT + Plugins, which offers a robust system integrator using WiFi, NFC, and Bluetooth connection to access external widgets from a smartphone, scored highly on the number of downloads, it was not among the top-five apps in terms of Google Plus recommendations and user satisfaction. Perhaps this was because the app was far too generic, without dedicated features for fitness activities. The other top-five apps (both free and paid) had similar scores for the three ranking metrics. The ranking in cluster B followed a predictable pattern as the same apps scored similarly on the three metrics. The app Daily Yoga – Fitness On-the-Go emerged as the top free app with a wide variety of training videos and live video assistance. Similarly in the paid segment of this cluster, Pocket Yoga was the best seller. Likewise, the top-five free and paid apps in cluster C were almost identical across the three metrics. We observed that stretch exercises were among the top-five downloads but it failed to expand its reach through Google Plus recommendations and user satisfaction. Finally, in cluster D, the rankings of both free and paid apps were consistent across the three metrics. In the paid segment of this cluster, Runtastic Sit-Ups PRO was among the top-five most downloaded apps but it failed to register among the top recommended or satisfactory apps.

Product functionality comparison through radar charts. The top app league table calls for a more detailed study of these apps to identify and compare their specific features. In this section, we map the functionality dimensions of these top-five apps (in terms of user satisfaction) on radar charts by calculating each app's semantic similarity against the ten functionality components as in the previous section of the naïve Bayes classification using the Log-Hyponym algorithm (see Miao *et al.*, 2010; Pirró and Seco, 2008). The higher score of one functionality component represents a higher proportion of words in the app's self-provided description relevant to this functionality component. Even though, we chose these top-five apps for illustration purpose, the radar chart analysis can be applied to other apps in our sample.

From Figure 3(a)-(d), we find that in cluster A, RunKeeper emerged as the best overall performer, excelling on functional aspects such as tracking and logging for biking, gymnastic workouts, yoga, and weight loss. Runkeeper also outperformed the other four apps in terms of interoperability with a wide variety of wearable widgets. Runtastic running and fitness, however, excelled on navigation. It is worth noting that all apps, except cardiograph, performed poorly in their ability to measure internal body statistics such as heartbeat beyond physical activity workout statistics. Focusing on the health metrics dimension, cardiograph differentiated itself in cluster A by offering scope for complementarities with other apps. In cluster B, all top-five apps focused on a specific market niche with very similar functionalities on the yoga/weight control feature. Daily Yoga – Fitness On-the-Go emerged as

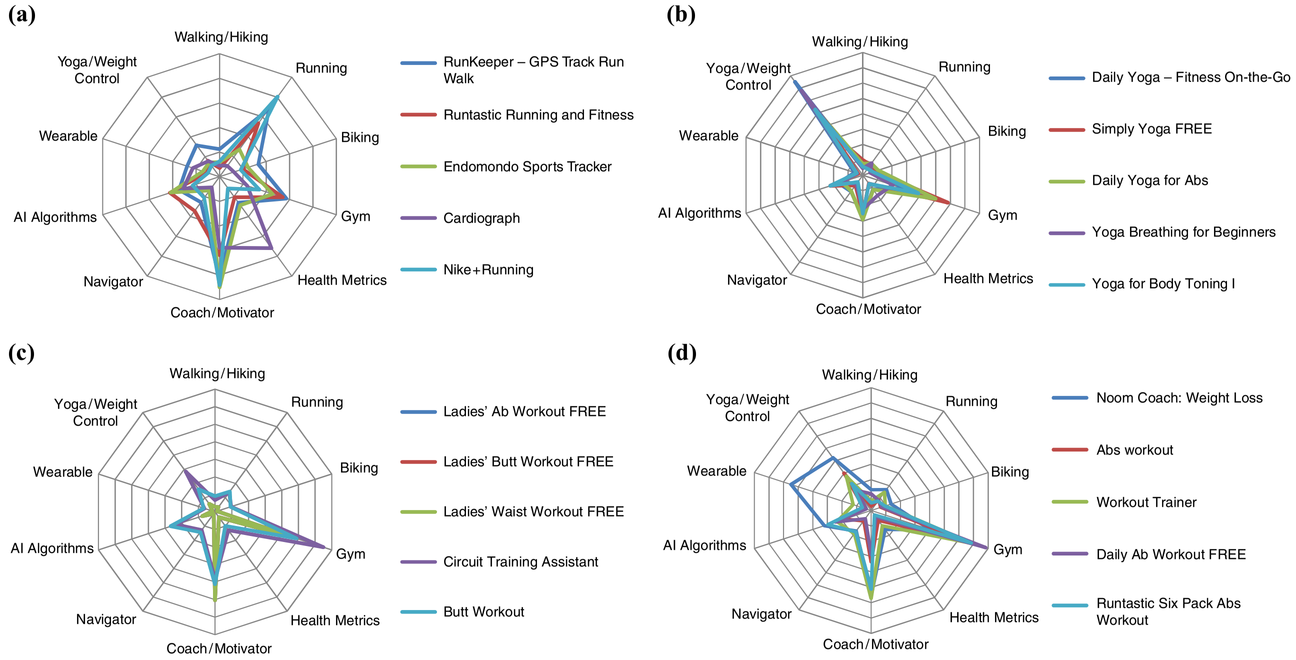
Table I. App-level league table by cluster

	Download	G+Recommendation	User satisfaction	Download	G+Recommendation	User satisfaction
		Cluster A			Cluster B	
Free app						
No. 1	ANT + Plugins	RunKeeper – GPS Track Run Walk	RunKeeper – GPS Track Run Walk	Daily Yoga – Fitness On-the-Go	Daily Yoga – Fitness On-the-Go	Daily Yoga – Fitness On-the-Go
No. 2	RunKeeper – GPS Track Run Walk	Runtastic Running & Fitness	Runtastic Running & Fitness	Simply Yoga FREE	Simply Yoga FREE	Simply Yoga FREE
No. 3	Runtastic Running & Fitness	Cardiograph	Endomondo Sports Tracker	Daily Yoga for Abs	Daily Yoga for Abs	Daily Yoga for Abs
No. 4	Cardiograph	Endomondo Sports Tracker	Cardiograph	Yoga Breathing for Beginners	Yoga Breathing for Beginners	Yoga Breathing for Beginners
No. 5	Endomondo Sports Tracker	Nike + Running	Nike + Running	Yoga for Body Toning	Yoga for Body Toning	Yoga for Body Toning
Paid app						
No. 1	Runtastic PRO	Runtastic PRO	Runtastic PRO	Pocket Yoga	Pocket Yoga	Pocket Yoga
No. 2	Endomondo Sports Tracker PRO	Endomondo Sports Tracker PRO	Endomondo Sports Tracker PRO	Simply Yoga	Simply Yoga	Simply Yoga
No. 3	Pedometer	Zombies, Run!	Zombies, Run!	A Facial Yoga & Facelift	A Facial Yoga & Facelift	A Facial Yoga & Facelift
No. 4	Zombies, Run!	Runtastic Pedometer PRO	Runtastic Pedometer PRO	Complete Yoga For Beginners	Complete Yoga For Beginners	Complete Yoga For Beginners
No. 5	Runtastic Pedometer PRO	Instant Heart Rate – Pro	Instant Heart Rate – Pro	Yoga Training	Yoga for Weight Loss Pro	Yoga for Weight Loss Pro

(*continued*)

Table I.

	Download	G+Recommendation	User satisfaction	Download	G+Recommendation	User satisfaction
		Cluster C			Cluster D	
Free app						
No. 1	Ladies' Ab Workout FREE	Ladies' Ab Workout FREE	Ladies' Ab Workout FREE	Noom Coach: Weight Loss	Noom Coach: Weight Loss	Noom Coach: Weight Loss
No. 2	Ladies' Butt Workout FREE	Ladies' Butt Workout FREE	Ladies' Butt Workout FREE	Abs workout	Abs workout	Abs workout
No. 3	Ladies' Waist Workout FREE	Ladies' Waist Workout FREE	Ladies' Waist Workout FREE	Workout Trainer	Workout Trainer	Workout Trainer
No. 4	Butt Workout	Circuit Training Assistant	Circuit Training Assistant	Daily Ab Workout FREE	Daily Ab Workout FREE	Daily Ab Workout FREE
No. 5	Stretch Exercises	Butt Workout	Butt Workout	8 Minutes Abs Workout	Runtastic Six Pack Abs Workout	Runtastic Six Pack Abs Workout
Paid app						
No. 1	Gymrat: Workout Tracker & Log	Gym Book: training notebook	Gymrat: Workout Tracker & Log	Just 6 Weeks	Just 6 Weeks	Just 6 Weeks
No. 2	Zumba Dance	Circuit Training Assistant Pro	Zumba Dance	Runtastic Push-Ups PRO	Runtastic Push-Ups PRO	Runtastic Push-Ups PRO
No. 3	Circuit Training Assistant Pro	Tabata Pro – Tabata Timer	Gym Book: training notebook	Fitness Buddy: 1,700 Exercises	HIIT – interval workout PRO	HIIT – interval workout PRO
No. 4	Tabata Pro – Tabata Timer	Ladies' Butt Workout	Circuit Training Assistant Pro	JEFIT Pro – Workout & Fitness	Fitness Buddy: 1,700 Exercises	JEFIT Pro – Workout & Fitness
No. 5	Gym Book: training notebook	Ladies' Chest Workout	Tabata Pro – Tabata Timer	Runtastic Sit-Ups PRO	JEFIT Pro – Workout & Fitness	Fitness Buddy: 1,700 Exercises



Notes: (a) The functionality component comparison of cluster A; (b) the functionality component comparison of cluster B; (c) the functionality component comparison of cluster C; (d) the functionality component comparison of cluster D

Figure 3.
The functionality component comparisons by cluster

the best overall performer as it provided a wide variety of yoga exercises and live voice guide. Simply Yoga FREE scored better than the other four apps on the gym dimension because it specializes in simple yoga tutorials. In cluster C, although circuit training assistant did not score highly in any of the three rankings, it managed to distinguish itself by providing a huge variety of workout videos targeted at women and by allowing a high level of customization in the workout regimes. Finally, unlike most apps in cluster D, Noom Coach: Weight Loss emerged as the top performer by integrating built-in fitness programs, diet tracking and workout coaching in one app. It relies heavily on wearable widgets and AI algorithms to control weight. All top-five apps in this cluster are comparable on the coach/motivator and gymnastics/dance dimension with a wide variety of workouts and exercise tutorials.

Revenue model and pricing. Table II shows the results on the revenue models followed by the fitness apps. Based on the free or paid and with or without in-app purchases business model components, the numbers of mobile apps are tabulated for each cluster. In cluster A, the free revenue model dominated the cluster with approximately 54 percent apps 68 percent of free apps and 86 percent of paid apps opted for without in-app purchases. In cluster B, rather more than two-thirds (69 percent) of the apps followed the free revenue model and only 3 out of 36 chose the within-app purchase model. In cluster C, the free revenue model dominated the cluster with 59 percent free apps. Just 4 out of 73 apps adopted for the within-app purchase model. Finally, 54 percent of the apps in cluster D followed the free revenue model whereas just 4 percent of the paid apps chose the within-app purchase model. In general, free (55.7 percent) was the dominant revenue model in the fitness app business and firms seemed to be reluctant to choose the within-app purchase model (15.14 percent).

Table III summarizes the pricing trend for paid, free and in-app purchases of apps in general. The price of a paid app ranges from $0.50 to $19.90 while the prices of the paid in-app purchase items vary from as low as $0.50 to as high as $199. However, prices for the free in-app purchase items varied less, from $0.71 to $55.

Firm-level competitor analyses

Market structure analysis. As a mobile app firm can own multiple apps in different clusters, in this section, we aggregate app-level to firm-level and conduct firm-level competitor analysis. Table IV shows the trend in the number of apps of which a particular mobile app

	Cluster A	Cluster B	Cluster C	Cluster D	Total
Paid					
Without in app purchase	63	10	29	118	220
With in app purchase	10	1	1	5	17
Paid total	73	11	30	123	237
Free					
Without in app purchase	57	23	40	114	234
With in app purchase	26	2	3	33	64
Free total	83	25	43	147	298

Table II. App-level revenue model

Price	Maximum	Minimum
Paid app	19.9	0.5
Paid with in app purchase	199	0.5
Free with in app purchase	54.99	0.71

Table III. App-level pricing model

firm may own. We identified a total of 269 mobile app firms from the sample of 535 mobile apps. The majority of the firms (68.02 percent) owned just one app, while nine firms had between five and nine mobile apps, seven firms had 10 to 20 mobile apps and two firms had more than 20 mobile apps. Table II shows the distribution of firms across the four clusters. The majority of the firms (92.94 percent) operated in a single cluster – 104 and 123 firms in clusters A and D, respectively, whereas clusters B (19 firms) and C (44 firms) were less crowded. A small number of firms offered apps across two clusters and three firms offered apps across three clusters.

We then calculated a firm's market share by cluster as the cumulative sum of the number of downloads of all its apps in a cluster divided by the total number of downloads for that cluster. We analyzed market structure with the concentration ratio, which is the measure of the percentage market share in an industry held by the largest firms within that industry (Bain, 1951). We adopted the most common concentration ratio, CR4, namely the market share of the four largest firms. The results in Table V show that the four largest firms share more than 50 percent of the market in clusters A and D, and more than 80 percent in clusters B and C, suggesting an oligopoly in all four clusters. However, the concentration ratio does not use the market share of all the firms in each cluster. In order to understand market share distribution, we estimated the location and scale parameters for each cluster. The results in Table V suggest that all four location parameters were close to zero and all four scale parameters were much larger than 1, indicating that the market share in all four clusters was not normally distributed. Finally, to gain a complete picture of industry concentration, we calculated the Herfindahl index (HHI). The results of HHI indicated that clusters A and D were less concentrated than clusters B and C.

Firm-level direct and indirect competitor identification. We visualized firm-level within-cluster competition with a brand-new 2D chart. The point of origin represents the reference firm.

	No. of firms	Cluster	No. of firms
> 20	2	A	104
20-10	7	B	19
9-5	9	C	44
4-2	68	D	123
1	183	Cross 2 clusters	16
Total	269	Cross 3 clusters	3

Table IV. Number of firms by cluster and number of app

Cluster A	Market share	Cluster B	Market share	Cluster C	Market share	Cluster D	Market share
ANT+	21.37%	Daily Yoga Inc.	40.51%	DoMobile Health	53.49%	Caynax	18.54%
Runtastic	13.05%	IMOBLIFE Co. Ltd	38.51%	Passion4profession apps	11.70%	Noom Inc.	15.96%
Fitness Keeper, Inc.	10.15%	Daily workout apps, LLC	7.39%	IMOBLIFE INC.	11.69%	Skimble Inc.	15.22%
Endomondo.com	9.30%	Sally Tam	5.76%	PocketFitness	6.04%	Daily workout apps, LLC	13.00%
CR4	53.86%	CR4	92.18%	CR4	82.93%	CR4	62.71%
HHI	11.10%	HHI	32.29%	HHI	31.98%	HHI	11.11%
Location	–8.61	Location	–6.37	Location	–6.88	Location	–8.08
Scale	3.23	Scale	3.59	Scale	2.90	Scale	3.00

Table V. Market structure comparison

The *x*-axis measures the relative functionality similarity of other firms in the cluster to the reference firm using the LDA functionality similarity index matrix. A greater distance from the origin to a particular firm along the *x*-axis denotes a greater similarity in functionality between that firm and the reference firm. The *y*-axis measures the level at which a particular firm competes with the reference firm in terms of user segments. To do so, we first identified users who have recommended or commented on the apps of the reference firm and those of the other firms in the same cluster by mining data from every user's Google Plus webpage. Then, we calculated a User Segment Overlap (USO) Index using the following equation:

$$\text{User Segment Overlap Index}_{i,j} = \frac{\text{User}_{\text{Firm}_i} \cap \text{User}_{\text{Firm}_j}}{\text{User}_{\text{Firm}_i} \cup \text{User}_{\text{Firm}_j} - \text{User}_{\text{Firm}_i} \cap \text{User}_{\text{Firm}_j}}$$

where $\text{User}_{\text{Firm}_{i/j}}$ is the number of users who have recommended or commented on the apps of Firm_i (the reference firm) or those of Firm_j (any particular firm in the same cluster) on Google Plus.

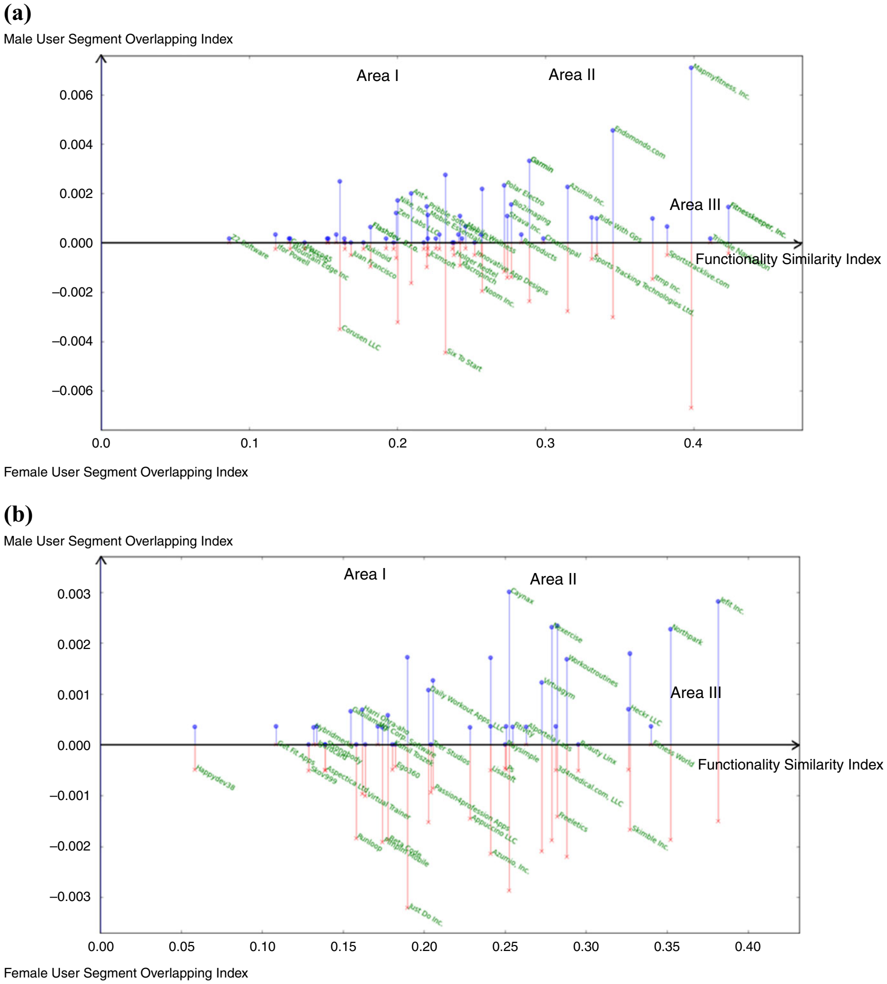
For the firms that have multiple apps in a cluster, we took the averages of LDA similarity indices and USO indices as measures. To provide to better understand the customer influence on product needs and thus help an operations manager with more detailed information, we split users into two groups based on their gender information on Google Plus and calculated the male/female USO indices, respectively. The values along the negative *y*-axis indicated the female USO index (i.e. the red line in Figure 4) and the positive *y*-axis represented male USO index (i.e. the blue line in Figure 4). We illustrate our 2D firm-level competition graph with Runtastic GmbH (hereafter, Runtastic) as the reference firm. We chose this firm because its apps scored highly in the rankings of clusters A and D. Figure 4(a)-(b) map Runtastic's competing firms in clusters A and D, respectively.

We divided each chart into three segments:

(1) Area of indirect competitors, which includes a set of firms whose apps are similar to Runtastic in terms of functionality (i.e. close to the origin along the *x*-axis). For example, firms like Z2 Software in cluster A and Get Fit in cluster D, with some degree of product differentiation, are indirect competitors of Runtastic.

(2) Area of direct competition with low user lock-in, which refers to a set of firms whose apps are very similar to Runtastic in terms of their product functionalities and with highly overlapping user segments. Hence, firms in this area can be regarded as direct or immediate competitors of Runtastic. Given that these firms have a high proportion of the same user segments, this means that a large number of their users have used both their apps and those of Runtastic, indicating a high possibility that users will switch to an alternate app. Consequently, changes in Runtastic's pricing or product features will have direct implications for the other firms. In cluster D, NorthPark is one such firm. Because it operates with only four free apps, user switching to or from Runtastic is highly probable. The trend in cluster A is similar, where firms like MapMyFitness, Endomondo and Garmin are direct competitors of Runtastic and run the risk of user switch over in the free app market.

(3) Area of direct competition with high user lock-in. This is a set of firms whose apps are very similar to those of Runtastic. However, very few users have recommended or commented on both Runtastic's apps and those of the other firm in this set. That is to say, users are thus more loyal to their firm without trying Runtastic's apps. We infer that the firms in this area will not be easily defeated by Runtastic. For example, in cluster D, Fitness World provides highly similar apps and hence is a



Notes: (a) Firm-level comparison in cluster A; (b) firm-level comparison in cluster D

Figure 4.
Firm level comparison in clusters A and D

direct competitor of Runtastic. Yet the low value on the *y*-axis suggests that Fitness World successfully locks its users via its paid apps. We may infer that any incremental change in Runtastic's pricing or product offers will not generate any substantial impact on Fitness World. A similar trend is also visible in cluster A, where a firm like Sportstracklive operates with two paid apps and exhibits high user lock-in. FitnessKeeper, in contrast also maintains high user lock-in but provides only one free app (Runkeeper). This might well be due to Runkeeper's significantly superior quality.

In short, our analyses show that market structure analysis can assist managers in designing sustainable operations strategies based on the mapping of their products *vis-à-vis* competitors offerings. Different product and process re-engineering strategies can be adapted depending on the competitive segment that an app belongs to. Our analyses suggest that one operational competitiveness indicator that firms need to include is that based on service innovation in product design. This can be well illustrated for apps in segment 2 that exhibit poor user lock-in in comparison to other segments. In this segment, customers are constantly trying new products, and hence operations managers may have to be extra vigilant to competitors' offerings. Competitor intelligence information through our automated market structure analyses can be new tool in business process re-engineering of mobile apps. Product functionality competitor analyses, revenue model and pricing analyses can help in faster product redesign to adapt to the constantly changing market needs. Moving from the traditional performance metrics of an operational strategy like quality, flexibility, delivery, and costs (Aboelmaged, 2012), we propose innovation in product design as a valuable indicator to create new winning strategies in the marketplace, as explained in the market structure analyses.

Conclusion

With the increasingly complex competitive environment, operational strategists are investing increased time and effort in revising sustainability strategies of their firms. In this paper, we investigate competitor intelligence as means to determine the operational/functional strategy for business sustainability. We present through our analyses the competitive performance metrics that can guide operations managers in better design of the strategic objectives. Operational strategies aimed at various pricing models, product differentiation, faster adaptation to competitor offerings may be investigated by firms depending on the segment that a firm falls into, as depicted in the market structure analysis. Meanwhile the recent revolution in information and communication technologies has created new means to collect and examine data in radically different ways. In this paper, we have presented a big data analytics approach to process online text content automatically and to conduct competitor analysis and thus propose operational strategy recommendations. Notably, our approach extends previous work by developing a fully automatic big data competitor analysis system that integrates state-of-the-art NLP, machine learning algorithms and data visualization techniques. We have incorporated multiple data sources to overcome the bias that may exist in consumer feedback data. As we are aware that there is no way of validating the authenticity of customer feedback, we employ sources such as expert reviews and product descriptions offered by firms along with user evaluations and social media profiles to better understand the market competition. Thus, the credibility of our study is substantially enhanced by employing multiple data sources and large amounts of structured as well as unstructured data, subjected to a fully automated approach, which is able to provide both a predictive and a defensive mechanism to identify opportunities and threats, coalesce all relevant sources of competitor analyses into a single framework, and support efficient and effective strategy formulation, operations implementation and performance monitoring.

Broadly, our work contributes to the operations management literature by providing significant cross-fertilization between market structure theories and information technology to improve the efficiency and effectiveness of extracting competitor information. This big data approach is systematic and rigorous enough to inject a large dose of objectivity into the competitor analysis process. It is versatile and flexible enough to handle unstructured text information, which can create substantial economic gain for firms. In particular, we contribute to the study of competition analysis with our new functionality-similarity – user-segment overlapping chart (i.e. Figure 4(a)-(b), which harnesses the power of social media data and becomes a key basis of competition metrics to classify a firm's rivals into three different categories. Valuable insights can be gained from all three categories, prompting a swift response to those firms that offer highly similar products and are competing for the attention of the same user segments. The product-level competitor and functionality analyses as well as the market structure analysis can be instrumental in guiding operational managers in gaining a wider market perspective based on customer needs and competitor performance. Finally, these can be employed in designing performance metrics like cost efficiency, product detailing, flexibility in business functions renovations/changes, etc.

In terms of managerial implications, our sonar-like big data competitor analysis system promises to support operational decision making both descriptively and prescriptively. The Naïve Bayes classifier and k-NN clustering separate potential competitors from direct and indirect ones. Our novel LDA topic modeling algorithm, combined with conventional MDS, reveals a product's market position in comparison with its peers. The league table enables managers to identify top performers, and the product feature radar chart highlights salient app attributes and value propositions. The analyses of the revenue model, pricing policies and market structure enables managers to detect tomorrow's opportunities and to predict better courses of action regarding all aspects of a firm's operations. Finally, the rise of hyper-competition makes it necessary to emphasize exploration over exploitation. Operations managers can use the overlapping user segments and functionality similarity indices to map indirect and direct competitors with and without user lock-in. Thus, our system offers firms multiple benefits. With better, explicit information on rival firms in terms of their product features, new product development can be enhanced by delivering superior or unique functionality. Our system also reduces the cost of human intervention while leveraging the vast data marts of free data to obtain competitive intelligence. In summary, our analyses of fitness mobile app firms demonstrated the ability of NLP and machine learning-based big data analytics to assess competitive market structure, gain information on competitor perceptions, and gauge the future moves of rivals using different data sources in a timely and cost effective manner. Our big data competitor analysis system would be helpful for firms in emerging industries in which competitor data are not readily available through conventional marketing studies.

Several limitations should be noted. First, our study does not include historical versions of the descriptions of each app, so we cannot track brand position evolution and evaluate the performance of a firm's operations in the past. Future research could perform such dynamic analyses by using historical data points either from mobile app firms or from Google Play app store. In addition, using Google Plus recommendations and user comments may cause a self-selection bias. Those individuals who are highly motivated to recommend or comment on an app, typically individuals who have strong opinions or substantial consumer knowledge, are over-represented. Thus, individuals who are indifferent or apathetic are less likely to respond and are therefore being under-represented. This often leads to a polarization of responses, with extreme perspectives receiving disproportionate weight in our app rankings. Future studies should aim at developing sophisticated algorithms to overcome this sample bias issue. Finally, we demonstrated the value of the proposed system in the fitness mobile app business setting. Future research could also explore applications of our approach in other domains in order to test its external validity. In short, we hope that our big data competitor analytics approach provides a first step in exploring the enormous, rich, useful body of online data readily and freely available on the internet. This is just the beginning of our big data journey underpinning new waves of productivity growth, innovation, and consumer surplus as long as the right policies and enablers are in place (McKinsey, 2015).

References

Aboelmaged, M.G. (2012), "Harvesting organizational knowledge and innovation practices: an empirical examination of their effects on operations strategy", *Business Process Management Journal*, Vol. 18 No. 5, pp. 712-734.

Adeniyi, D.A., Wai, Z. and Yongquan, Y. (2014), "Automated web usage data mining and recommendation system using k-nearest neighbor (KNN) classification method", *Applied Computing and Informatics*, Vol. 12 No. 1, pp. 90-108.

Aho, A.-M. and Uden, L. (2013), "Strategic management for product development", *Business Process Management Journal*, Vol. 19 No. 4, pp. 680-697.

Alegre-Vidal, J., Lapiedra-Alcamí, R. and Chiva-Gomez, R. (2004), "Linking operations strategy and product innovation: an empirical study of Spanish ceramic tile producers", *Research Policy*, Vol. 33 No. 5, pp. 829-839.

Allen, R.S. and Helms, M.M. (2006), "Linking strategic practices and organizational performance to porter's generic strategies", *Business Process Management Journal*, Vol. 12 No. 4, pp. 433-454.

Amoako-Gyampah, K. and Boye, S. (2001), "Operations strategy in an emerging economy: the case of the Ghanaian manufacturing industry", *Journal of Operations Management*, Vol. 19 No. 1, pp. 59-80.

Amoako-Gyampah, K. and Meredith, J. (2007), "Examining cumulative capabilities in a developing economy", *International Journal of Operations & Production Management*, Vol. 27 No. 9, pp. 928-944.

Archak, N., Ghose, A. and Ipeirotis, P.G. (2007), "Show me the money! Deriving the pricing power of product features by mining consumer reviews", *Proceedings of the ACM KDD 2007*, *San Jose, CA*, *August 12-15*, Association for Computing Machinery, New York, NY.

Archak, N., Ghose, A. and Ipeirotis, P.G. (2011), "Deriving the pricing power of product features by mining consumer reviews", *Management Science*, Vol. 57 No. 8, pp. 1485-1509.

Bain, J.S. (1951), "Relation of profit rate to industry concentration. American manufacturing 1936-40", *Quarterly Journal of Economics*, Vol. 65 No. 3, pp. 293-324.

Barnes, D. (2001), "Research methods for the empirical investigation of the process of formation of operations strategy", *International Journal of Operations & Production Management*, Vol. 21 No. 8, pp. 1076-1095.

Bergen, M. and Peteraf, M.A. (2002), "Competitor identification and competitor analysis: a broad-based managerial approach", *Managerial and Decision Economics*, Vol. 23 Nos 4-5, pp. 157-169.

Bird, S., Klein, E. and Loper, E. (2009), *Natural Language Processing with Python*, 1st ed., O'Reilly Media, Inc., Newton, MA.

Blei, D.M. (2012), "Probabilistic topic models", *Communications of the ACM*, Vol. 55 No. 4, pp. 77-84.

Blei, D.M., Ng, A.Y. and Jordan, M.I. (2003), "Latent dirichlet allocation", *The Journal of Machine Learning Research*, Vol. 3 No. 1, pp. 993-1022.

Bloodgood, J.M. and Bauerschmidt, A. (2002), "Competitive analysis: do managers accurately compare their firms to competitors?", *Journal of Managerial Issues*, Vol. 14 No. 4, pp. 418-434.

Boyer, K. (1998), "Longitudinal linkages between intended and realized operations strategies", *International Journal of Operations & Production Management*, Vol. 18 No. 4, pp. 356-373.

Boyer, K. and McDermott, C. (1999), "Strategic consensus in operations strategy", *Journal of Operations Management*, Vol. 17 No. 3, pp. 289-305.

Boyer, K. and Pagell, M. (2000), "Measurement issues in empirical research: improving measures of operations strategy and advanced manufacturing technology", *Journal of Operations Management*, Vol. 18 No. 3, pp. 361-374.

Buntine, W. (2002), "Variational extensions to EM and multinomial PCA", in Elomaa, T. *et al.* (Eds), *European Conference on Machine Learning*, Springer, Berlin and Heidelberg, pp. 23-34.

Buntine, W. and Jakulin, A. (2005), "Discrete principal component analysis", *Proceedings of the Subspace, Latent Structure and Feature Selection Techniques: Statistical and Optimisation Perspectives Workshop*, *Bohinj*, *February 23-25*.

Calori, R., Johnson, G. and Sarnin, P. (1994), "CEOs' cognitive maps and the scope of the organization", *Strategic Management Journal*, Vol. 15 No. 6, pp. 437-457.

Caves, R.E. (1984), "Economic analysis and the quest for competitive advantage", *Papers and Proceedings the 96th Annual Meeting of the American Economic Association*, Vol. 74 No. 2, pp. 127-132.

Chen, H., Chiang, R.H. and Storey, V.C. (2012), "Business intelligence and analytics: from big data to big impact", *MIS Quarterly*, Vol. 36 No. 4, pp. 1165-1188.

Chen, J., Huang, H., Tian, S. and Qu, Y. (2009), "Feature selection for text classification with naïve bayes", *Expert Systems with Applications*, Vol. 36 No. 3, pp. 5432-5435.

Cheng, D., Liu, G. and Qiu, Y. (2008), "Applications of particle swarm optimization and K-nearest neighbors to emotion recognition from physiological signals", *Proceedings of the International Conference on Computational Intelligence and Security (CIS), Vol. 2, Suzhou*, pp. 52-56.

Christiansen, T., Berry, W., Bruun, P. and Ward, P. (2003), "A mapping of competitive priorities, manufacturing practices, and operational performance in groups of danish manufacturing companies", *International Journal of Operations & Production Management*, Vol. 23 No. 10, pp. 1163-1183.

Cover, T. and Hart, P. (1982), "This week's citation classic: nearest neighbor pattern classification", *IEEE Transactions on Information Theory*, Vol. 13 No. 1, pp. 21-27.

Czepiel, J.A. and Kerin, R.A. (2012), "Competitor analysis", in Shankar, V. and Carpenter, G.S. (Eds), *Handbook of Marketing Strategy*, Edward Elgar, Cheltenham and Northampton, MA, pp. 41-57.

Dangayach, G. and Deshmukh, S. (2001), "Manufacturing strategy – literature review and some issues", *International Journal of Operations & Production Management*, Vol. 21 No. 7, pp. 884-932.

de Waal, B. and Batenburg, R. (2014), "The process and structure of user participation: a BPM system implementation case study", *Business Process Management Journal*, Vol. 20 No. 1, pp. 107-128.

Decker, R. and Trusov, M. (2010), "Estimating aggregate consumer preferences from online product reviews", *International Journal of Research in Marketing*, Vol. 27 No. 4, pp. 293-307.

DeSarbo, W.S., Daniel, J.H. and Kamel, J. (1991), "MULTICLUS: a new method for simultaneously performing multidimensional scaling and cluster analysis", *Psychometrika*, Vol. 56 No. 1, pp. 121-136.

Díaz, M., Gil, M. and Machuca, J. (2005), "Performance measurement systems, competitive priorities, and advanced manufacturing technology: some evidence from the aeronautical sector", *International Journal of Operations & Production Management*, Vol. 25 No. 8, pp. 781-799.

Doan, A., Ramakrishnan, R. and Halevy, A.Y. (2011), "Crowd-sourcing systems on the world-wide web", *Communications of the ACM*, Vol. 54 No. 4, pp. 86-96.

Domingos, P. and Pazzani, M. (1997), "On the optimality of the simple bayesian classifier under zero-one loss", *Machine Learning*, Vol. 29 Nos 2-3, pp. 103-130.

Dörre, J., Gerstl, P. and Seiffert, R. (1999), "Text mining: finding nuggets in mountains of textual data", *Proceeding of Fifth ACM SIGKDD International Conference on Knowledge Discovery Data Mining*, ACM, New York, NY, pp. 398-401.

Duda, R.O. and Hart, P.E. (1973), *Pattern Classification and Scene Analysis*, Vol. 3, Wiley, New York, NY.

Elrod, T. (1988), "Choice map: inferring a product-market map from panel data", *Marketing Science*, Vol. 7 No. 1, pp. 21-40.

Elrod, T. (1991), "Internal analysis of market structure: recent developments and future prospects", *Marketing Letters*, Vol. 2 No. 3, pp. 253-266.

Elrod, T., Gary, J.R., Allan, D.S., Rick, L.A., Lynd, B., Barry, L.B., Carroll, J.D., Johnson, R.M., Kamakura, W.A., Peter, L., Mazanec, J.A., Rao, V.R. and Shankar, V. (2002), "Inferring market structure from customer response to competing and complementary products", *Marketing Letters*, Vol. 13 No. 3, pp. 221-232.

Englyst, L.N. (2003), "Operations strategy formation – a continuous process", *Integrated Manufacturing Systems*, Vol. 14 No. 8, pp. 677-685.

Erdem, T. and Keane, M.P. (1996), "Decision-making under uncertainty: capturing dynamic choice processes in turbulent consumer good markets", *Marketing Science*, Vol. 15 No. 1, pp. 1-20.

Espino-Rodríguez, T.F. and Rodríguez-Díaz, M. (2014), "Determining the core activities in the order fulfillment process: an empirical application", *Business Process Management Journal*, Vol. 20 No. 1, pp. 2-24, doi: 10.1108/BPMJ-01-2013-0012.

Feldman, R. and Sanger, J. (2006), *The Text Mining Handbook*, Cambridge University Press, New York, NY.

Feldman, R., Govindaraj, S., Livnat, J. and Segal, B. (2010), "Management's tone change, post earnings announcement drift and accruals", *Revenue Accounting Studies Journal*, Vol. 15 No. 4, pp. 915-953.

Feldman, R., Fresko, M., Goldenberg, J., Netzer, O. and Ungar, L. (2007), "Extracting product comparisons from discussion boards", *Proceedings of the 2007 Seventh IEEE International Conference on Data Mining*, Association for Computing Machinery, New York, NY, pp. 469-474.

Feldman, R., Fresko, M., Goldenberg, J., Netzer, O. and Ungar, L. (2008), "Using text mining to analyze user forums", *Proceeding of Service Systems Service Management 2008 International Conference (IEEE Systems, Man, and Cybernetics Society), Melbourne*, pp. 1-5.

Flurry (2014), "Health and fitness apps finally take off", available at: www.flurry.com/blog/flurry-insights/health-and-fitness-apps-finally-take-fueled-fitness-fanatics#.VPsPU_nF-kQ (accessed January 21, 2015).

Flynn, B. and Flynn, E. (2004), "An exploratory study of the nature of cumulative capabilities", *Journal of Operations Management*, Vol. 22 No. 5, pp. 439-457.

Friedman, J.H. (1997), "On bias, variance, 0/1 – loss, and the curse-of-dimensionality", *Data Mining and Knowledge Discovery*, Vol. 1 No. 1, pp. 55-77.

Ghose, A. and Ipeirotis, P.G. (2008), "Estimating the socio-economic impact of product reviews: mining text and reviewer characteristics", *IEEE Transactions on Knowledge & Data Engineering*, Vol. 23 No. 10, pp. 1498-1512.

Ghose, A., Ipeirotis, P.G. and Li, B. (2012), "Designing ranking systems for hotels on travel search engines by mining user-generated and crowdsourced content", *Marketing Science*, Vol. 31 No. 3, pp. 493-520.

Godes, D., Mayzlin, D., Chen, Y., Das, S., Dellarocas, C., Pfeiffer, B., Libai, B., Sen, S., Shi, M. and Verlegh, P. (2005), "The firm's management of social interactions", *Marketing Letters*, Vol. 16 No. 3, pp. 415-428.

Goes, P.B. (2014), "Big data and IS research", *MIS Quarterly*, Vol. 38 No. 3, pp. iii-viii.

Green, P.E. and Srinivasan, V. (1978), "Conjoint analysis in consumer research: issues and outlook", *Journal of Consumer Research*, Vol. 5 No. 2, pp. 103-123.

Griffin, A. and Hauser, J.R. (1993), "The voice of the customer", *Marketing Science*, Vol. 12 No. 1, pp. 1-27.

Griffiths, T. and Steyvers, M. (2002), "A probabilistic approach to semantic representation", *Proceedings of the 24th Annual Conference of the Cognitive Science Society*, pp. 381-386.

Griffiths, T. and Steyvers, M. (2003), "Prediction and semantic association", *Advances in Neural Information Processing Systems*, pp. 11-18.

Griffiths, T.L. and Steyvers, M. (2004), "Finding scientific topics", *Proceedings of the National Academy of Sciences*, Vol. 101 No. S1, pp. 5228-5235.

Grobler, A. and Grubner, A. (2006), "An empirical model of the relationships between manufacturing capabilities", *International Journal of Operations & Production Management*, Vol. 26 No. 5, pp. 458-485.

Grossler, A. and Grübner, A. (2006), "An empirical model of the relationships between manufacturing capabilities", *International Journal of Operations and Production Management*, Vol. 26 No. 5, pp. 458-485.

Hamel, G. and Prahalad, C.K. (2005), "Strategic intent", *Harvard Business Review*, Vol. 83 No. 7, pp. 148-161.

Han, E., Karypis, G. and Kumar, V. (2001), "Text categorization using weight adjusted K-nearest neighbor classification", *Proceedings of the Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD), Hong Kong*, pp. 53-65.

Hayes, R. and Upton, D. (1998), "Operations-based strategy", *California Management Review*, Vol. 40 No. 4, pp. 8-25.

Hayes, R. and Wheelwright, S. (1984), *Restoring Our Competitive Edge Competing through Manufacturing*, Wiley, New York, NY.

Hodgkinson, G.P. and Johnson, G. (1994), "Exploring the mental models of competitive strategists: the case for a processual approach", *Journal of Management Studies*, Vol. 31 No. 4, pp. 525-551.

Hofmann, T. (1999), "Probabilistic latent semantic indexing", *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, ACM, pp. 50-57.

Hofmann, T. (2001), "Unsupervised learning by probabilistic latent semantic analysis", *Machine Learning*, Vol. 42 Nos 1-2, pp. 177-196.

Hu, M. and Liu, B. (2004), "Mining and summarizing customer reviews", *Proceeding of Tenth ACM SIGKDD International Conference of Knowledge Discovery Data Mining*, ACM, New York, NY, pp. 22-25.

Jayaraman, V. and Srivastava, R. (1996), "Expert systems in production and operations management – current applications and future prospects", *International Journal of Operations and Production Management*, Vol. 16 No. 12, p. 27.

Jurafsky, D. and Martin, J.H. (2009), *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition*, Prentice-Hall, Englewood Cliffs, NJ.

Kacur, J., Vargic, R. and Mulinka, P. (2011), "Speaker identification by K-nearest neighbors: application of PCA and LDA prior to K-NN", *Proceedings of the International Conference on Systems, Signals and Image Processing (IWSSIP), Bucharest*, pp. 1-4.

Kamakura, W.A. and Russell, G.J. (1989), "A probabilistic choice model for market segmentation and elasticity structure", *Journal of Marketing Research*, Vol. 26 No. 4, pp. 87-96.

Koller, D. and Sahami, M. (1997), "Hierarchically classifying documents using very few words", *Proceedings of the Fourteenth International Conference on Machine Learning, San Mateo, CA*, Morgan Kaufmann, pp. 170-178.

Krauthammer, M. and Nenadic, G. (2004), "Term identification in the biomedical literature", *Journal of Biomedical Informatics*, Vol. 37 No. 6, pp. 512-526.

Lamb, R. (1984), *Competitive Strategic Management*, Prentice Hall, Englewood Cliffs, NJ.

Lawrence, W. (2008), "Small business operations strategy: aligning priorities and resources", *Journal of Small Business Strategy*, Vol. 18 No. 2, pp. 89-105.

Lee, T. (2007), "Constraint-based ontology induction from online customer reviews", *Group Decision and Negotiation*, Vol. 16 No. 3, pp. 255-281.

Lee, T.Y. (2005), "Ontology induction for mining experiential knowledge from customer reviews", presentation at Utah Winter Information Systems Conference, Salt Lake City, UT, March 11-12.

Lee, T.Y. and Bradlow, E.T. (2011), "Automated marketing research using online customer reviews", *Journal of Marketing Research*, Vol. 48 No. 5, pp. 881-894.

Li, F. (2010), "The information content of forward-looking statements in corporate filings – a naïve bayesian machine learning approach", *Journal of Accounting Research*, Vol. 48 No. 5, pp. 1049-1102.

Lusch, R.F., Liu, Y. and Chen, Y. (2010), "The phase transition of markets and organizations: the new intelligence and entrepreneurial frontier", *IEEE Intelligent Systems*, Vol. 25 No. 1, pp. 71-75.

McCallum, A. and Nigam, K. (1998), "A comparison of event models for naive bayes text classification", *In AAAI-98 Workshop on Learning for Text Categorization*, Vol. 752 No. 1, pp. 41-48.

McKinsey (2015), "Big data: the next frontier for competition", available at: www.mckinsey.com/features/big_data (accessed February 9, 2015).

Manning, C.D. and Schutze, H. (1999), *Foundations of Statistical Natural Language Processing*, The MIT Press, Cambridge, MA.

Martin-Pena, M. and Diaz-Garrido, E. (2008), "Typologies and taxonomies of operations strategy: a literature review", *Management Research News*, Vol. 31 No. 3, pp. 200-218.

Miao, Q., Li, Q. and Zeng, D. (2010), "Fine-grained opinion mining by integrating multiple review sources", *Journal of the American Society for Information Science and Technology*, Vol. 61 No. 11, pp. 2288-2299.

Miller, J.G. and Roth, A.V. (1994), "A taxonomy of manufacturing strategies", *Management Science*, Vol. 40 No. 3, pp. 285-304.

Minarro-Viseras, E., Baines, T. and Sweeney, M. (2005), "Key success factors when implementing strategic manufacturing initiatives", *International Journal of Operations & Production Management*, Vol. 25 No. 2, pp. 151-179.

Netzer, O., Feldman, R., Goldenberg, J. and Fresko, M. (2012), "Mine your own business: market-structure surveillance through text mining", *Marketing Science*, Vol. 31 No. 3, pp. 521-543.

Ogbonaya, I.O. (2008), *Introduction to Matlab/Simulink, for Engineers and Scientist*, 2nd ed., John Jacob's Classic Publishers Ltd, Enugu.

Paiva, E., Roth, A. and Fensterseifer, J. (2008), "Organizational knowledge and the manufacturing strategy process: a resource-based view analysis", *Journal of Operations Management*, Vol. 26 No. 1, pp. 115-132.

Pang, B. and Lee, L. (2008), "Opinion mining and sentiment analysis", *Foundations and Trends in Information Retrieval*, Vol. 2 Nos 1-2, pp. 1-135.

Pant, G. and Sheng, O. (2009), "Avoiding the blind spots: competitor identification using web text and linkage structure", *ICIS 2009 Proceedings*, p. 57.

Pirró, G. and Seco, N. (2008), "Design, implementation and evaluation of a new semantic similarity metric combining features and intrinsic information content", in Meersman, R. and Tari, Z. (Eds), *On the Move to Meaningful Internet Systems: OTM 2008*, Springer, Berlin and Heidelberg, pp. 1271-1288.

Porac, J.F. and Thomas, H. (1990), "Taxonomic mental models of competitor definition", *Academy of Management Review*, Vol. 15 No. 2, pp. 224-240.

Porter, M.E. (1980), *Competitive Strategy: Techniques for Analyzing Industries and Competitors*, Free Press, New York, NY.

Porter, M.E. (1985), *Competitive Advantage*, Free Press, New York, NY.

Przewozniczek, M., Walkowiak, K. and Wozniak, M. (2010), "Optimizing distributed computing systems for K-nearest neighbours classifiers – evolutionary approach", *Logic Journal of IGPL*, Vol. 19 No. 2, pp. 357-372.

Rennie, J.D., Shih, L., Teevan, J. and Karger, D.R. (2003), "Tackling the poor assumptions of naive bayes text classifiers", *Proceedings of the Twentieth International Conference on Machine Learning ICML*, Vol. 3, pp. 616-623.

Riis, J., Dukovska-Popovska, I. and Johansen, J. (2006), "Participation and dialogue in strategic manufacturing development", *Production Planning & Control*, Vol. 17 No. 2, pp. 176-188.

Rosenzweig, E. and Easton, G. (2010), “Tradeoffs in manufacturing? A meta-analysis and critique of the literature”, *Production and Operations Management*, Vol. 19 No. 2, pp. 127-143.

Rytter, N., Boer, H. and Koch, C. (2007), “Conceptualizing operations strategy processes”, *International Journal of Operations & Production Management*, Vol. 27 No. 10, pp. 1093-1114.

Sahami, M. (1996), “Learning limited dependence bayesian classifiers”, *KDD*, Vol. 96 No. 1, pp. 335-338.

Salton, G. and McGill, M.J. (1983), *Introduction to Modern Information Retrieval*, McGraw-Hill, New York, NY.

Scherer, F.M. and Ross, D. (1990), *Industrial Market Structure and Economic Performance*, Houghton Mifflin, Boston, MA.

Schlegel, G.L. (2014), “Utilizing big data and predictive analytics to manage supply chain risk”, *Journal of Business Forecasting*, Vol. 33 No. 4, pp. 11-17.

Seshadri, T. and Tellis, G.J. (2012), “Does chatter really matter? Dynamics of user-generated content and stock performance”, *Marketing Science*, Vol. 31 No. 2, pp. 198-215.

Shamsuzzoha, A.H.M. (2011), “Modular product architecture for productivity enhancement”, *Business Process Management Journal*, Vol. 17 No. 1, pp. 21-41.

Shi, M. and Yu, W. (2013), “Supply chain management and financial performance: literature review and future directions”, *International Journal of Operations and Production Management*, Vol. 33 No. 10, pp. 1283-1317.

Slack, N. and Lewis, M. (2002), *Operations Strategy*, Prentice-Hall, Edinburgh.

Srivastava, R.K., Mark, I.A. and Allan, D.S. (1984), “A customer-oriented approach for determining market structures”, *Journal of Marketing*, Vol. 48, Spring, pp. 32-45.

Steyvers, M. and Griffiths, T. (2007), “Probabilistic topic models”, *Handbook of Latent Semantic Analysis*, Vol. 427 No. 7, pp. 424-440.

Sum, C., Kow, S. and Chen, C. (2004), “A taxonomy of operations strategies of high performing small and medium enterprises in Singapore”, *International Journal of Operations & Production Management*, Vol. 24 Nos 3-4, pp. 321-348.

Tsai, W., Su, K.H. and Chen, M.J. (2011), “Seeing through the eyes of a rival: competitor acumen based on rival-centric perceptions”, *Academy of Management Journal*, Vol. 54 No. 4, pp. 761-778.

Ulrich, K.T. and Eppinger, S.D. (2003), *Product Design and Development*, McGraw-Hill, New York, NY.

Vulić, I., De Smet, W., Tang, J. and Moens, M.F. (2015), “Probabilistic topic modeling in multilingual settings: an overview of its methodology and applications”, *Information Processing & Management*, Vol. 51 No. 1, pp. 111-147.

Wallach, H.M. (2006), “Topic modeling: beyond bag-of-words”, *Proceedings of the 23rd International Conference on Machine Learning*, ACM, pp. 977-984.

Wang, G.A., Atabakhsh, H. and Chen, H. (2011), “A hierarchical naïve bayes model for approximate identity matching”, *Decision Support Systems*, Vol. 51 No. 3, pp. 413-423.

Ward, P., Brickford, D. and Leong, G. (1996), “Configurations of manufacturing strategy, business strategy, environment, and structure”, *Journal of Management*, Vol. 22 No. 4, pp. 597-626.

Ward, P., McCreery, J., Ritzman, L. and Sharma, D. (1998), “Competitive priorities in operations management”, *Decision Sciences*, Vol. 29 No. 4, pp. 1035-1046.

Weiss, S.M., Indurkhya, N., Zhang, T. and Damerau, F. (2010), *Text Mining: Predictive Methods for Analysing Unstructured Information*, Springer, New York, NY.

West, P.M., Christina, L.B. and Stephen, J.H. (1996), “Consumption vocabulary and preference formation”, *Journal of Consumer Research*, Vol. 23 No. 2, pp. 120-135.

Wieland, U., Fischer, M., Pfitzner, M. and Hilbert, A. (2015), “Process performance measurement system – towards a customer-oriented solution”, *Business Process Management Journal*, Vol. 21 No. 2, pp. 312-331.

Wu, X., Kumar, V., Quinlan, J.R., Ghosh, J., Yang, Q., Motoda, H., McLachlan, G.J., Ng, A., Liu, B., Yu, P.S., Zhou, Z.-H., Steinbach, M., Hand, D.J. and Steinberg, D. (2007), "Top 10 algorithms in data mining", *Knowledge and Information Systems*, Vol. 14 No. 1, pp. 1-37.

Xiao-peng, Y. and Xiao-gao, Y. (2007), "Novel text classification based on K-nearest neighbour", *Proceedings of the International Conference on Machine Learning and Cybernetics*, *Hong Kong*, Vol. 6, pp. 3425-3430.

Yu, W. and Ramanathan, R. (2009), "Effects of business environment on operations strategy: an empirical study of retail firms in China", NUBS Research Paper Series No. 2009-11, Nottingham University Business School, Nottingham, pp. 1-28.

Yu, Y., Duan, W. and Cao, Q. (2013), "The impact of social and conventional media on firm equity value: a sentiment analysis approach", *Decision Support Systems*, Vol. 55 No. 4, pp. 919-926.

Zajac, E. and Bazerman, M. (1991), "Blind spots in industry and competitor analysis: implications of interfirm (mis)perception to strategic decisions", *Academy of Management Review*, Vol. 16 No. 1, pp. 37-46.

About the authors

Dr Liang Guo is an Associate Professor of big data and the Deputy Head of the BNP-KPMG Endowed Centre of Entrepreneurship & Innovation at NEOMA Business School, France. He completed his PhD at the Cambridge University.

Dr Ruchi Sharma works as a Lecturer in e-business at the School of Strategy and Leadership, Coventry University, UK. She completed her PhD in Management from ECRICOME PhD Universa in 2016 and worked as a Research Assistant at BNP-KPMG Endowed Centre of Business Models & Entrepreneurial Innovation, at NEOMA Business School, Rouen. Her research interests are business model business model innovation and its impact on value creation and value appropriation for firms operating in a coopetitive ecosystem. She holds a MBA and a BTech Degree in Mechanical Engineering.

Lei Yin is a PhD Candidate in Big Data Analytics at NEOMA Business School, France. He is under the supervision of Dr Liang Guo.

Ruodan Lu is a PhD Candidate in Applied Machine Learning at the Cambridge University, UK. Ruodan Lu is the corresponding author and can be contacted at: rl508@cam.ac.uk

Dr Ke Rong is an Associate Professor in the Institute of Economics, School of Social Science at Tsinghua University in China. Ke has received the PhD Degree from the University of Cambridge and obtained Bachelor Degree in Tsinghua University. Before joining Tsinghua, he was a Senior Lecturer in the University of Exeter and Bournemouth University in UK and a Visiting Scholar in the Harvard Business School. His research interests include business/innovation ecosystems, data driven ecosystem, and platform market and network effects in sharing economy. His research has been published in *International Journal of Production Economics, Journal of International Management, Group and Organization Management and Technological Forecasting and Social Change*.