Decision Support Systems 70 (2015) 60–72

Contents lists available at ScienceDirect

Decision Support Systems

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

Online profiling and clustering of Facebook users

Jan-Willem van Dam, Michel van de Velden *

Econometric Institute, Erasmus University Rotterdam, P.O. Box 1738, 3000 DR Rotterdam, The Netherlands

ARTICLE INFO

Article history:
Received 20 November 2013
Received in revised form 22 September 2014
Accepted 1 December 2014
Available online 9 December 2014

Keywords:
Online profiling
Social networks
Customer relationship management
Correspondence analysis
Cluster analysis
Facebook

ABSTRACT

In a relatively short period of time, social media have acquired a prominent role in media and daily life. Although this development brought about several academic endeavors, the literature concerning the analysis of social media data to investigate one's customer base appears to be limited. In this paper, we show how data from the social network site Facebook can be operationalized to gain insight into the individuals connected to a company's Facebook site. In particular, we propose a data collection framework to obtain individual specific data and propose methodology to explore user profiles and identify segments based on these profiles. The proposed data collection framework can be used as an identification step in an analytical customer relationship management implementation that specifically focuses on potential customers. We illustrate our methodology by applying it to the Facebook page of an internationally well-known professional football (soccer) club. In our analysis, we identify four clusters of users that differ with respect to their indicated "liking" profiles.

1. Introduction

Social networks and their role played in daily life increased considerably over the last few years. As illustrated by the editorials of two recent special issues [3,8], recent academic publications cover a broad spectrum of topics related to social media. Some examples concern the potential of social media and its effect on customer loyalty [4]; how to use Facebook to activate customers in sharing product/service recommendations [7,16,19], the role of social networks, in particular Facebook, on intentional social actions [6]; the relationship between personal networks and patterns of Facebook usage [29]; the effectiveness of user generated content in stimulating sales [10,17,39]. This short list of topics and references is by no means exhaustive. It only serves to illustrate the recent interest and range of applications relating to firms and social networks. One common element among extant literature is that none of these studies build on directly observed individual level social network data. Instead, either aggregate data or focus groups and/or (online) surveys were employed in order to answer the research questions. This limitation was recently also observed by [31] In their study of the effect of social media participation on visit frequency and profitability, survey respondents were linked to their social media (i.e. Facebook) profiles by matching of names. In this paper we add to the existing literature by explicitly considering the retrieval and analysis of profile data directly obtained from social network sites. The proposed methodology can be implemented into an analytical customer relationship management (CRM) framework aimed at the analysis of customer characteristics that may help improve a firm's customer management strategies. Moreover, by focusing on data from public profiles from the social media platform Facebook, we are able to identify potential rather than actual customers. That is, in contrast to typical CRM implementations that rely on data directly obtained from customers, we consider a much broader group of individuals that indicated an interest in a firm even when an actual purchase has not yet been recorded.

The contribution of this paper is threefold: First, we show how Facebook users that "like" a firm can be identified. As also observed by [31] this is not a trivial task. Second, using the information volunteered by such Facebook users through their publicly available pages, we show how segments of Facebook users can be identified through data visualization and cluster analysis methods. Clustering of a firms' Facebook fans, may improve understanding of strategic segmentation of social media users connected to a firm [28] Moreover, the cluster results and visualizations can be used to improve targeting of marketing efforts. For example, a company may consider seeking cooperation with another brand or a popular media figure based on the popularity of such a brand or person with the (potential) customers. Moreover, such efforts could be targeted directly at specific groups of (potential) customers rather than at all (potential) customers. Third, we apply our methodology to a (large) data set of Facebook users that indicated liking a popular and successful international football club. This football club granted us administrator rights, under provision of not revealing the name of the club and any results indicative of the club's name. The results of our analysis show that, based on the Facebook users' liking behavior, clusters can be obtained. Given the differences between liking patterns

* Corresponding author at: Michel van de Velden, Econometric Institute, Erasmus University Rotterdam, P.O. Box 1738, 3000 DR Rotterdam, The Netherlands.
E-mail addresses: jwvdam@gmail.com (J.-W. van Dam), vandevelden@ese.eur.nl (M. van de Velden).

http://dx.doi.org/10.1016/j.dss.2014.12.001
0167-9236/

in these clusters, differentiated marketing strategies for the different clusters can be developed.

The remainder of this paper is structured as follows. First, we briefly review previous research on analytical CRM and online profiling. Next, we briefly discuss specific data considerations for Facebook. We introduce some terminology and review Facebook's data analysis and programming facilities. In Section 4, we show how specific Facebook users can be identified. Next, we analyze the individual level data using a combined multiple correspondence analysis and k-means cluster analysis method. We show how results can be visualized and interpreted. The paper concludes with a discussion of our results, implications for research and practice, and future directions.

2. Customer relationship management and online profiling

Customer relationship management (CRM) has become widely recognized as an important business approach [27,31] defines CRM as an "enterprise approach to understanding and influencing customer behavior through meaningful communications in order to improve customer acquisition, customer retention, customer loyalty, and customer profitability". Hence, customer acquisition (or: identification) can be seen as the first step in a customer relationship management cycle that, together with retention and customer development form a complete cycle geared at creating a better understanding of (potential) customers in order to increase long term customer value to the firm [22,33,30].

Customer identification is typically based on information directly available to a firm [38] For example, customers may be required to provide certain background information upon purchasing a product. In addition, companies may ask customers to volunteer information by completing a survey or persuading them to join a loyalty program. Based on the available data, a customer profile, that is, a model of the customer, can be constructed. Based on such a customer profile, a marketeer decides on appropriate strategies and tactics to meet the specific needs of the consumer [32]. Hence, possessing accurate information about preferences and background characteristics of your (potential) customers makes it possible to improve targeting of, possibly individual specific, marketing efforts [25].

Obtaining direct customer information requires an existing relationship with the firm. That is, customers need to have either purchased a product or made contact with the firm in such a way that identification is possible so that additional information can be collected. In the case of yet unidentified potential customers, it is not possible to acquire data in this fashion. Moreover, except for the observable transactional data (i.e., purchase time and amounts, etc.) customers may decline to provide additional information.

Social media offer a new source of customer profile information. In particular, social media offer opportunities to identify *potential* customers. Through social media, individuals often express preferences for brands, products, services, persons, political parties, etc., in a free unsolicited way. Thus, if one is able to collect such information from potential customers of a certain firm, for example by focusing on users that indicated an interest in that firm, online profiles can be created that allow for better, individualized, targeting.

Although it has been suggested that the rise of new media requires novel approaches to successfully manage customer relationships [18], applications in which customer background data from social network sources are used to gain insight into customer backgrounds, appear to be under represented in the academic literature. There are some studies [24,23,15,5] in which social network data were used that contained personal information of the users in the data set. However, the goal of these studies was to study network ties [24], privacy issues [5,15], or relating the number of friends to the amount of information available on a person's Facebook page [23]. None of the studies used the data for online profiling: The collection of information from the Internet for the purpose of formulating a profile of users' habits and interests [37]. In this paper, we fill this gap by proposing a data collection framework for the purpose of online profiling.

Online profiling can be divided into two categories: reactive and non-reactive data collection [37]. *Non-reactive data collection* focuses on the collection of data concerning Web usage behavior, e.g., IP addresses of visitors, time spent on certain Web pages, and clicking behavior information. These data are used to gain insight into Web user behavior, and thus, characteristics of individual visitors or visitor groups. Non-reactive data form a large and potentially interesting source of online profile information. However, for the construction of online user profiles, the observed usage behavior must first be transformed into meaningful variables. The construction and definition of such variables is not always a easy task. In our study, we therefore primarily focus on the retrieval and analysis of online profiles based on reactive rather than non-reactive data.

Reactive data collection zooms in on visitor characteristics which cannot be collected through tracking Web usage behavior of a visitor. Instead, reactive information is collected by using forms and selection menus, which have to be filled in by visitors themselves. Reactive data requires little to no recoding of the original variables and they are immediately collected at the user level. Moreover, in the case of Facebook, providing reactive data requires very little effort from the users. For example, when joining Facebook, users are asked to provide certain personal background information (e.g., name, gender, date of birth). Users provide this information by selecting the appropriate options. This basic background information can be supplemented by more personal information concerning, for example, hobbies, relationship status etc. Finally, by "liking" other pages, personal preferences for persons or objects can be indicated.

The resulting online profiles can be of great value for marketeers, as they can be used to identify different (segments of) users (customers) that require different marketing approaches. Moreover, it enables the company to know its potential customers, that is, individuals that indicated a preference towards the product/brand by "liking" it on Facebook.

3. Facebook data

Facebook users put personal information on their Facebook page. Some examples are someone's name, gender, date of birth, e-mail address, sexual orientation, marital status, interests, hobbies, favorite sports team(s), favorite athlete(s), or favorite music. Furthermore, it is possible to specify your Facebook friends, post messages, publish pictures or other content. Consequently, the potential value to marketeers and researchers of the information available through Facebook is substantial. However, extracting the information is no trivial task as:

1. Facebook users are able to make certain information not publicly available and therefore not visible to non friends.
2. Facebook users are not obliged to fill in fields and therefore, many users do not specify all possible information about themselves.
3. The default statistics that Facebook offers for Facebook page administrators are limited.
4. It is not obvious how Facebook users who "like" your page can be identified.

The first two points are a result of the design and policy of Facebook.com and therefore we take these points as these are. Instead, we focus on the extraction of available data from Facebook and consider a user profile data collection framework taking into account the abovementioned issues.

The Facebook data collection framework that we propose consists of three steps: 1) identification of "fans" of the Facebook page, 2) retrieval of relevant data for the identified fans, and 3) preparation of the data. The first step of this framework requires administrator rights to the page, in the other steps public information from the relevant pages needs to be collected. Before we show how to implement the data

collection framework, we briefly summarize some important aspects concerning Facebook pages and the available data.

3.1. Facebook Insights

The owner of a Facebook page is in principle the administrator of the page. Personal pages are typically managed only by the page owner, however, in the case of a company's Facebook pages, the page administrator can also give other Facebook users these administrator rights. It is possible to have multiple administrators for one Facebook page, e.g., multiple marketing and CRM employees may be page administrators. As Facebook page administrator, one has certain privileges in comparison to regular users or visitors of a Facebook page. As administrator, one can edit, publish and withdraw content, target advertisements and install Facebook apps on the page. Also, administrators have access to Facebook Insights, a dashboard which provides statistics on user's growth, demographics, consumption of content, and creation of content. However, the information made available through the dashboard is aggregated over users who "liked" the page. Consequently, the possibilities concerning the analysis of individual specific data using this feature, are limited.

On Facebook, users can indicate whether they "like" another Facebook page. Thus, they are able to express a form of affinity with the company, person or product behind the Facebook page. Through Facebook Insights it is possible to see how many users "like" your page, how this number evolves over time, and whether these users are active on your page or not. (The definition of an active Facebook user is as follows: users who, within a chosen time period, engaged with, viewed, or consumed content generated by a Facebook page). Furthermore, you can see which media on the page are most popular (e.g., watching videos, listening to audio, or viewing photos). It is also possible to see which Facebook tabs (e.g., the wall, information, photos, and events) are most popular and from which external referrers visitors come. Additionally, one is able to see which page posts have been viewed the most and which posts generated most user feedback.

The above-mentioned possibilities of the Facebook Insights dashboard all concern information related to the Facebook page itself. Information about the Facebook users is only present through aggregated breakdowns.

Fig. 1 gives an example of the breakdowns for the gender and age distributions based on the information a user provided on his Facebook page. The home country and home city are determined using the IP address from which users (who indicated "liking" the page) access the Facebook page. The language is based on the users' default language setting when accessing Facebook.

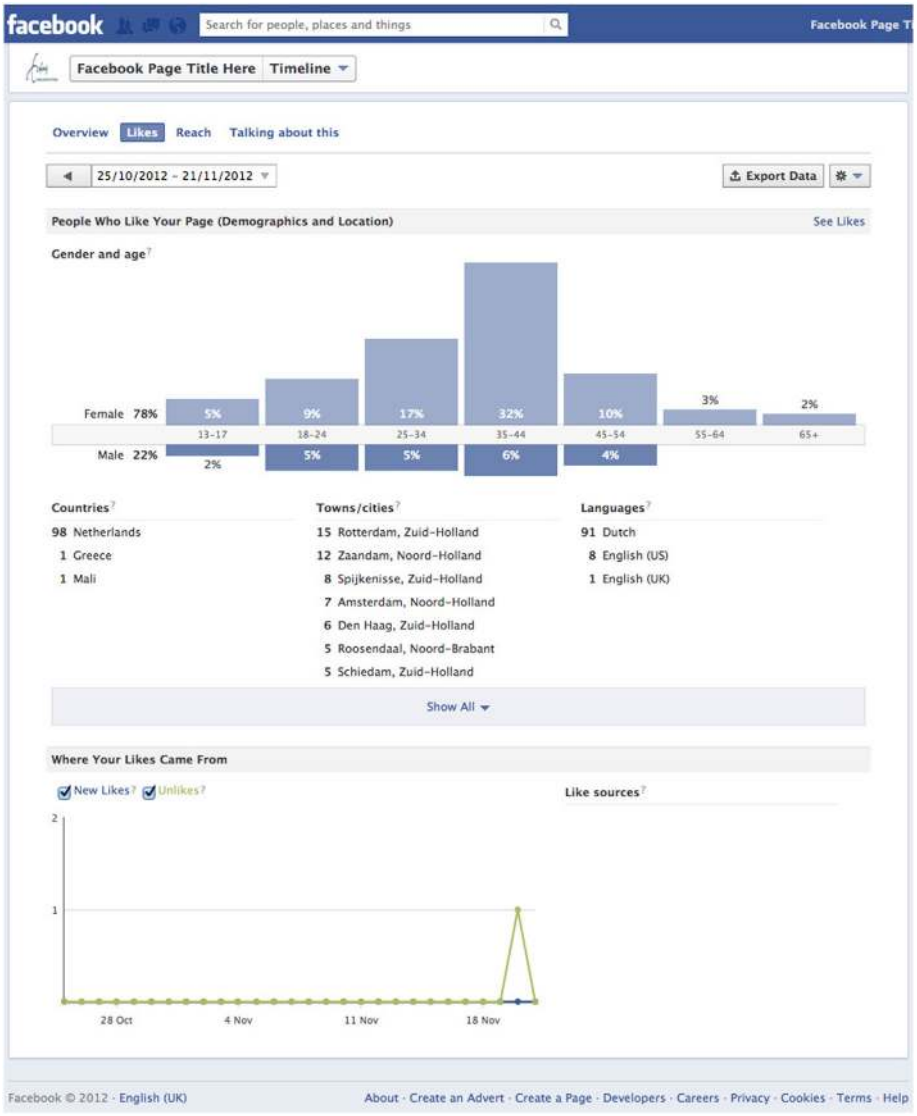

Fig. 1. Screenshot of Facebook Insights.

Other personal information of users, such as, for example, relationship status, sexual orientation, favorite brands, favorite music, liked pages, etc., are not accessible through the Facebook Insights' dashboard.

3.2. Facebook application programming interface (API)

Through Facebook Developers Platform [9] it is possible to develop web applications (or plugins) which make use of the Facebook platform; e.g., mobile applications which makes it possible to connect to your Facebook page to post pictures, applications which integrate Facebook features in a Web site, or applications which make it possible to find friends.

The Facebook Developers Platform consists of multiple application programming interfaces (APIs). The Graph API is the core of Facebook Platform, enabling one to read and write data to Facebook. It provides a simple and consistent view of the social graph, uniformly representing objects (e.g., people, photos, events and pages) and the connections between them (e.g., friendships, likes and photo tags) [9]. In addition to the Graph API, there is an Internationalization API, an Ads API, and a Chat API. For our data collection framework, the Graph API is crucial.

The Graph API makes it possible for developers to integrate Facebook into Web site (Web) applications, and to build Facebook applications. However, even when a Facebook user has a public profile, which is accessible (online) by anyone, the data in his profile are not publicly accessible through the API. In fact, only the following fields are always publicly available through the API: user id; username; full name; first name; last name; gender; locale (i.e. the default language setting); profile picture. For marketing or customer relationship management purposes, this list is not very useful. In addition, if we compare this list with the complete list of fields that Facebook provides, we observe that there is potentially much more relevant information available on the Facebook pages.

Considering the complete list of fields that Facebook provides, we identify as potentially interesting characteristics that are not available through the API: date of birth; place of birth; sexual orientation; political view; relationship status; education; work experience; contact information; activities; interests; likes (i.e. internet pages "liked" by a Facebook user, these could correspond to books, movies, athletes but also friends' Facebook pages. This last field, likes, is of particular interest in this study as we want to see if fans can be clustered according to the preferences indicated in this field. As the API does not allow the retrieval of these data, we need to develop alternative methods. In the next section, we consider how such data can be obtained.

4. Facebook user profile data collection framework

To gather a Facebook user's profile information relevant to customer relationship management and/or for marketing purposes, the information resources described in the previous section, must be combined. Fig. 2 shows the user profile data collection framework. For convenience we introduce the term "fan" for users who "liked" a Facebook page. In fact, Facebook itself originally gave users the option to "become a fan of" other pages and changed this into "like". The data collection framework consists of three parts: 1) identifying the 'fans', 2) gathering the personal information, and 3) preparing and structuring the gathered data.

4.1. Identifying Facebook page fans

Administrators of a Facebook page, can list fans of their page by accessing https://www.facebook.com/browse/?type=page_fans&page_id=1234567890, where 1234567890 should be replaced by the page ID of the page one is interested in (and is administrator of). A screen shot of the URL is given in Fig. 3. When one clicks on the 'See more' button on the bottom of the page, more 'fans' are listed. However, after showing 500 'fans', the button does not show up anymore.

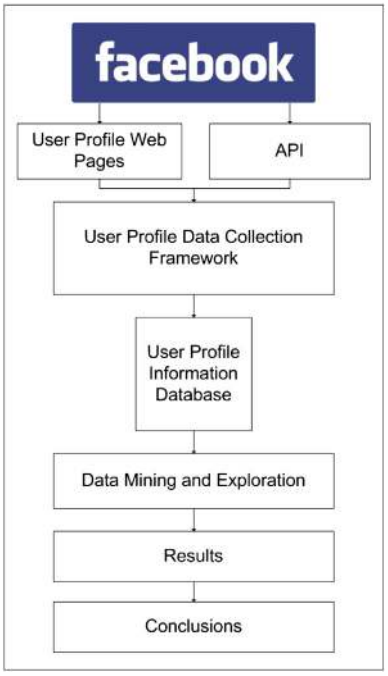

Fig. 2. Facebook data collection framework.

After exploring the HTTP requests resulting from "clicking" the 'See more' button, we conclude that:

- Facebook uses Asynchronous JavaScript (AJAX) for its HTTP requests.
- Facebook uses two parameters (fb_dtsg and post_form_id) to prevent cross-site request forgery (CSRF) in its HTTP requests.
- Only authenticated Facebook administrators have access to the page (cookies are used for authentication).
- The response format of the HTTP request is in JSON format, which contains each fan's picture, name, URL, and ID.

With these observations in mind, a PHP script to store the name, URL and ID of each Facebook user in a (MySQL) database, was written. The pseudo code for this script can be found as Algorithm 1 in Appendix A. Running this algorithm yields, after removing duplicate Facebook IDs, 10,000 unique 'fans'. Facebook does not give information about how these 'fans' are selected. By changing the fb_dtsg, and post_form_id parameters a new set of 10,000 unique 'fans' is obtained. Between these sets there exists some overlap, but the greater part of the 'fans' is different. As the algorithm always results in 10,000 unique Facebook IDs, it will be difficult to obtain a list with all fans if the total number of fans exceeds the 10,000. A sample, however, can be obtained without too much difficulty by using Algorithm 1, and varying the two parameters.

4.2. Gathering fan's public profile information

The second step in the data collection process is gathering information of the Facebook 'fans' identified in the previous step. This can be done by visiting these Facebook pages and storing the relevant, individual level, data. Note that only public information can be obtained. The data we thus obtain, only concerns users that granted public access to their pages.

4.3. Data preparation and storage

The third step in our profile data collection framework, concerns preparation and structuring of the gathered data. When one creates or updates his/her Facebook profile, personal information can (and in

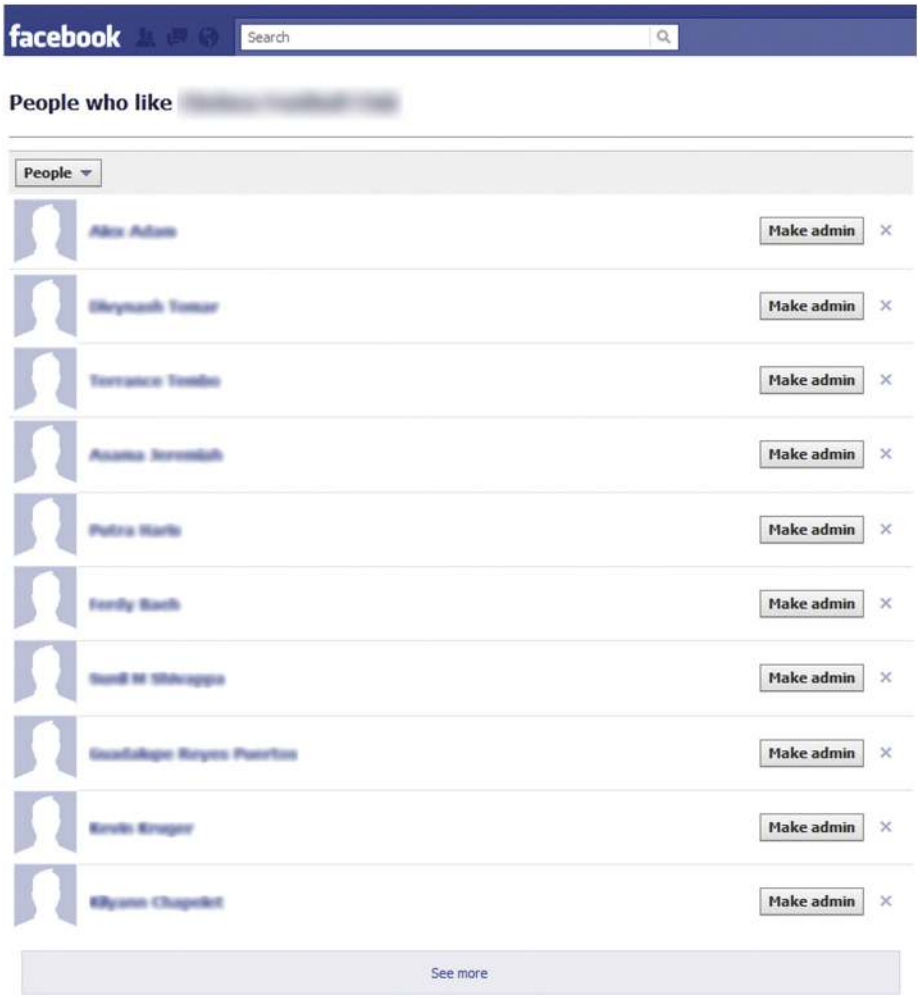

Fig. 3. Facebook administrators: screen shot of Web page which lists people who 'like' a Facebook page.

some cases, such as name and date of birth, must) be provided by completing several fields. There are text fields (e.g., name, language, interests), check boxes (e.g., sexual orientation), or lists (e.g., gender, relationship status). For check boxes and lists, the options which can be selected are limited. For text fields no such limitation exists and users can type in anything they like. We distinguish between two sets of variables: background characteristics and liking data.

4.3.1. Background characteristics

From individual Facebook pages we are able to obtain personal background information of the users. In particular, from the public profiles we can obtain the variables gender, date of birth, location, relationship status and the number of Facebook friends. Gender and date of birth are straightforward background variables. Concerning the other variables we briefly indicate how the data is available on the Facebook pages and how we process these for our analysis.

4.3.2. Location

A user's location is represented by a string with the name of the city or town someone lives in and/or comes from, together with a URL to the Facebook page of that location. Location may be useful when analyzing fans of a page and we may be interested in more details about the location. In particular, for a geographical overview of the fanbase one needs to know the country, continent, and the coordinates (latitude, longitude) of a location. This can be achieved by using the GeoNames geographical database [36]. The GeoNames API has a fuzzy search engine which accepts all kinds of input. For example, the engine accepts both 'Rio de Janeiro' and 'Rio Janeiro' as search terms for the large Brazilian city. As output, GeoNames yields various details such as, city name, country name, latitude, longitude, number of inhabitants, etc.

For 'fans' who don't publicly specify their location we cannot discover their latitude and longitude. However, through Facebook's API it is possible to gather Facebook's language setting. Assuming that the languages correspond to the user's location, one could use the language to determine, at country level, the user's location. That is, a Facebook user who's using Facebook in Japanese is assumed to come from Japan. Although we believe that this assumption is not a very unrealistic one, there are cases in which language is not linkable to one specific country. For example, we cannot infer a country from a user with a language setting such as "Arabic", "English", "French", "German" or "Spanish".

4.3.3. Relationship status

A Facebook user can specify their relationship status by selecting one of the following options: single, in a relationship, engaged, married, it's complicated, widowed, separated or divorced. However, in our data set we also found values as 'in a complicated relationship', 'in an open relationship', or 'civil partnership'. This is probably a result of the fact that Facebook changed the possible values for the field 'relationship status' over the years. At the time of our data gathering process (2011), it was not possible to choose other values than the eight values listed above. Therefore, we convert the values 'in a complicated relationship' and 'in an open relationship' into 'in a relationship', and 'civil partnership' becomes "married".

4.3.4. Number of Facebook friends

The number of Facebook friends can serve as a proxy for Facebook activity or popularity of a user.

4.3.5. Liking data

In addition to the background information, which, with the exception of the number of Facebook friends, is user supplied, we are able to find for each user which other Facebook pages are "liked". Based on this information we would like to see if clusters/segments of users can be identified. That is, is it possible to distinguish groups of Facebook users with similar "liking" patterns. Similar patterns could indicate similar preferences and companies may be able to employ segment specific marketing strategies. For example, if a segment of users tends to like certain artists more often than any other segment, promotions involving such an artist could be specifically targeted at that segment alone. In the next section, we introduce methodology to find and interpret segments based solely on the liking profiles of users.

5. Application: clustering Facebook fans

In the previous section we described in some detail how a Facebook page owner/administrator can obtain data from its fans. A customer relationship manager or a marketing manager would like to make these data operational by, for example, investigating whether fans can be segmented according to their indicated preferences and/or background characteristics. That is, is it possible to identify groups of fans on the basis of individual specific like data. For example, are certain brands or celebrities notably more popular in subgroups. Such information could be useful as it allows better targeting of marketing strategies.

A large internationally successful football (soccer) club granted us administrator rights to its official Facebook site. This enabled us to extract the data using the framework introduced in Section 2. For strategic purposes, the football club requested that its name, and any information that could possibly lead readers to infer the name, be kept from the public. Consequently, in our data analysis only a selection of general, not football related, labels are used.

At the time of the data extraction, February 2011, the total number of Facebook fans of the club was about 4 million. From these, we extracted data from over 40,000 fans. To check representativeness of this sample we compared the gender and age-distribution of our sample to that of the population as obtained through Facebook Insights. The results, presented in Table 1, show only small differences indicating that our data set is representative with respect to gender and age-distribution. Furthermore, we see that the Facebook fans of the football club are, not surprisingly, predominantly young males.

As explained in the previous section, Facebook users' data concerning their location can be enriched with geographical identifiers such as latitude and longitude. In Fig. 4, the resulting concentrations of fans are visualized in a heatmap created using Google maps API [13]. The figure shows a high concentration of Facebook 'fans' in Europe, India, Nigeria, South-east Asia, and Central America. Big parts of Africa and Australia have a low 'fan' concentration and in China there are hardly any 'fans' visible. This is a result of the fact that Facebook's penetration in China and Africa is relatively low, compared to that in other regions [2]. Australia has a relatively high general Facebook penetration (46 %) [2]. However, there are hardly any 'fans' of "our" football club in Australia, according to Fig. 4. Apparently, this football club is not popular in Australia on Facebook.

Table 1
Age and gender distributions: Insights' data versus our sample.

	Overall	13–24	25–34	35–44	45–54	55+
Male population	0.82	0.77	0.17	0.03	0.01	0.02
Male sample	0.79	0.74	0.22	0.01	0.01	0.02
Female population	0.14	0.77	0.15	0.04	0.02	0.02
Female sample	0.19	0.78	0.17	0.03	0.01	0.01

Recall that our data collection framework only allows for the retrieval of publicly published fields. Table 2 shows a breakdown of user's background data available in our initial sample. We see that except for gender and language settings, the percentage of users providing the profile information varies and is generally limited. We therefore exclude these variables when attempting to identify subgroups. Instead, we focus on the "like" data. Our goal is to find clusters of Facebook fans based on their liking data.

Facebook users can specify what/who they like on their profile page. For example, not only famous movie stars, movies, sports, athletes, tv-programs, and actors but also brands, restaurants or personal friends may be liked. For our initial sample of 43,861 fans, we found that 176,381 unique Facebook pages were "liked". However, of these 176,381 pages 77.5% was liked by only 1 user in our sample. Often, these pages are simply personal friends' Facebook pages. For our purposes, such pages are not interesting and a selection must be made.

We consider only the top 150 Facebook pages in terms of "likes" in our sample. Selecting data corresponding to these 150 most popular pages reduces our sample to 16,170 cases. However, the distribution of the number of likes in this sample of 16,170 users is rather skewed as many people have only few likes and only a few have many likes. To allow for discrimination on the basis of the liking profiles, we only consider users that liked at least 5 other pages. The resulting data set consists of 11,712 individuals. Constructing a data matrix with individuals as rows and the top 150 liked pages as columns, yields a large matrix with few observations. The scarcity of data (i.e., the many zeros indicating that a page is not "liked") and dimensionality of the data set, pose a serious problem for "normal" cluster analysis methods. We therefore analyze the large data matrix by using a joint dimension reduction and clustering approach.

5.1. MCA K-means

There exist several methods for clustering high-dimensional data. One popular approach is to use a two-step procedure. In the first step, a dimension reduction technique is used to reduce the dimensionality of the data. In the second step, cluster analysis is applied to the data in the reduced space. This method may be referred to as the tandem approach [1]. As shown by [35] an important drawback of this method is that the dimension reduction may distort or hide the cluster structure. To overcome this problem several methods have been proposed [20, 21,34] here we apply the joint MCA and K-means method proposed by [20].

MCA, also known as homogeneity analysis [12], yields optimal scaling values for the columns (i.e., quantifications for pages) in such a way that pages differently assessed by the individuals receive dissimilar scale values. Furthermore, rows (i.e., individuals) exhibiting dissimilar patterns of liked pages, receive dissimilar scale values. K-means clustering [26], finds clusters by minimizing the sum of squared deviations between the individual observations and their cluster means. [20] proposed a joint method, from here on referred to as MCA–Kmeans, that averages the MCA and K-means objective functions. An important advantage of the MCA–Kmeans approach is that it enables visualization of the data. A more formal formulation as well as an efficient algorithm useful for dealing with large data matrices is given in Appendix B.

5.2. Analysis

We apply MCA–Kmeans to the 11,712 observations with the 150 binary variables indicating whether a page was or was not liked by an individual. To decide upon the number of dimensions and clusters we inspect the changes in fit when more dimensions/clusters are added. In particular, for the dimensionality, we consider the adjusted explained

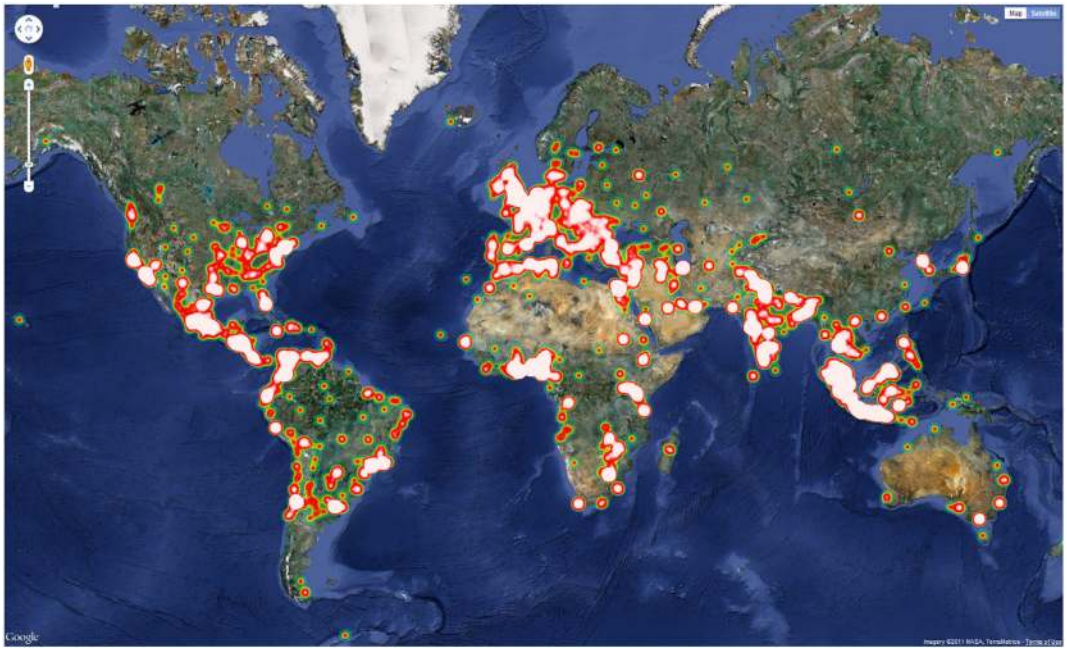

Fig. 4. Where do the football club's Facebook fans come from?

inertia of the MCA solution, as defined by [14] The adjusted explained inertia takes into account the rather specific structure of the data approximated in MCA. In particular, it corrects for the underestimation of typical correspondence analysis fit measures when applied to a (super)indicator matrix. Fig. 5 gives the cumulative explained inertia for the MCA solutions with different dimensionality. Although the effect is small, we can see that after three dimensions the effect of adding more dimensions decreases. We therefore consider only three dimensions in our analysis. An additional benefit of this choice is that it allows for graphical representations.

To select the number of clusters we consider the value of the objective function, using a three dimensional solution, with different numbers of clusters. In Fig. 6 the final objective function values are plotted against the number of clusters. The decrease in objective function value after four clusters is small and we therefore consider three dimensional solutions with four clusters.

5.3. Results

In Fig. 7, the solution using the first two dimensions of the MCA–Kmeans solution is given. Cluster memberships are indicated by using different colors and symbols. We see that in the first two dimensions three clusters appear separated from each other whereas the fourth cluster, situated around the origin, appears to overlap all three other clusters. As can be verified from Fig. 8, this fourth cluster separates itself from the other three clusters in the third dimension. Note that, in MCA, the origin corresponds to the average profile. That is the, average distribution of likes over Facebook pages. The fact that many attribute points are situated close to the origin, is partly due to the relatively large amount of not liked pages. That is, most users did not "like" more than 8 out of the 150 pages. Hence, each row of the data matrix contains many zeros for the "like" columns and, consequently, many ones for the corresponding "not liked" columns. This caused the attribute points corresponding to the "not liked" pages to dominate the mean profile and draw the corresponding points to the origin.

The spread and sizes of the clusters are nicely displayed in Figs. 7 and 8. However, the attributes (i.e. the Facebook pages) are not labeled to avoid further cluttering. Consequently, interpretation of the clusters in terms of the liked/not liked pages is not possible from these figures. For a better interpretation of the cluster with respect to the pages, Figs. 9 and 10 give joint plots of the cluster centers and the attributes. Points close to the origin have not been labeled to avoid clutter and, due to the confidentiality agreement we have relabeled pages corresponding famous football players as FFP and pages corresponding to football clubs as FC.

Table 2
Overview of the available profile information for the 43,861 users in our sample.

	% FB users in our sample
Gender	99.6
Date of birth	2.5
Relationship status	21.5
Sexual orientation	27.8
Location	54.2
Hometown	29.8
FB language setting	97.4
# FB friends	48.6
Education	37.1
Work experience	22.3

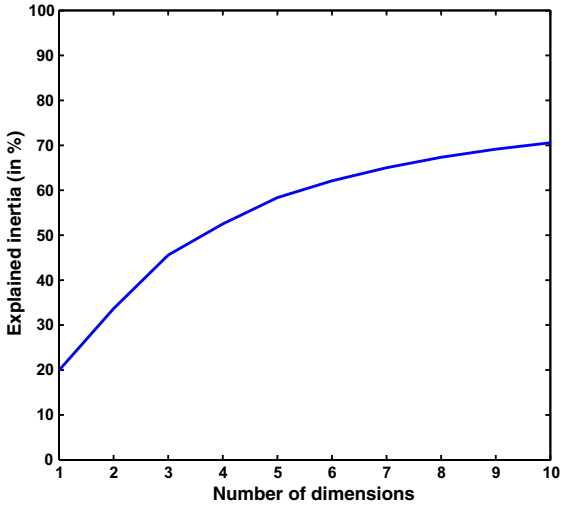

Fig. 5. Explained adjusted inertia as a function of number of MCA dimensions.

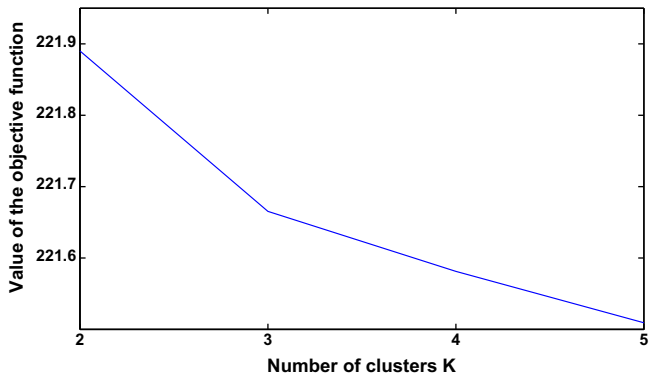

Fig. 6. Value of MCA–Kmeans objective for 3 dimensional solutions for different numbers of clusters.

Looking at the positions of the attributes and the clusters, we see that in cluster 1 (3549 observations) there appears to be a link with Latin America. That is, the Facebook page fútbol and pages corresponding to entertainers particularly popular in Latin America (e.g., Daddy Yankee, Wisin & Yandel) are relatively often liked. Also, the football club near cluster 1 is in fact the only South American club present in our data set. In cluster 2 (1734 observations) we find relatively many likes to Southeast Asia related stars and topics (e.g. SCTV and RCTI are Indonesian television stations, Upin and Ipin is a Malaysian television series, and Timnas Indonesia is the Facebook page of the Indonesian national football team). The three pages farthest removed from the origin and relatively often associated with individuals in cluster 3 (4048 observations) are cricket and India related (e.g., Sachin Tendulkar is a famous Indian cricket batsman). Other pages that are relatively often associated with this cluster are chess, traveling and sleeping. Note that the cluster mean for this cluster is not far from that of cluster 4 (2381 observations) so we should be careful in interpreting the pages close to both cluster centers on the basis of the first two dimensions. Instead, plotting the second and third dimension clarifies some differences as the clusters separate along the third dimension. Fig. 10 gives the corresponding plot, where again, to avoid clutter, we removed some labels and use the general labels for players and clubs. Individuals in the fourth cluster relatively often like pages corresponding to American entertainers (e.g. Vin Diesel, Selena Gomez, John Cena, Megan Fox). Also, Disney, Jackie Chan and Mafia Wars (a multiPlayer social network game) and Facebook are liked more often than average in this cluster.

It is important to note that the four clusters are characterized by liked Facebook pages that predominantly are not immediately football related. In fact, the clutter of football related pages close to the origin indicates that in all clusters, these pages are liked as well. This is not surprising as all individuals in our sample 'liked' the football club which granted us administrator rights thus asserting their interest in football. However, as the clusters differentiate themselves through non-football related pages, opportunities arise for cluster specific marketing efforts.

The MCA–Kmeans approach emphasizes relative rather than absolute differences. This means that if we look at the distribution of likes in each cluster, the attributes closest to the cluster means in the plot, need not be the most often observed in the cluster. In fact, as indicated before, for all clusters, the most often liked pages are predominantly football related pages. Table 3 lists the 10 most often liked pages in the four clusters. To distinguish between different football clubs and players we numbered them. Note that differences among the most popular Facebook pages are limited. The order of the clubs and players varies, but these are small differences that are of no practical significance.

The MCA–Kmeans results suggest that the clustering may be linked to geographical factors. To further study this we consider the Facebook data concerning the locations of the individuals. However, if individuals chose not to publish their locations, we cannot determine the country of origin. The language settings could be used to find plausible country or regions for these data. On the other hand, the fact that the information is missing may also be informative in itself and we choose to leave the missing locations as they are. Table 4 gives, for each cluster, the 10 most frequently occurring countries and the corresponding percentages of occurrences per cluster. We see that, as conjectured earlier, the first cluster has a clear Latin American

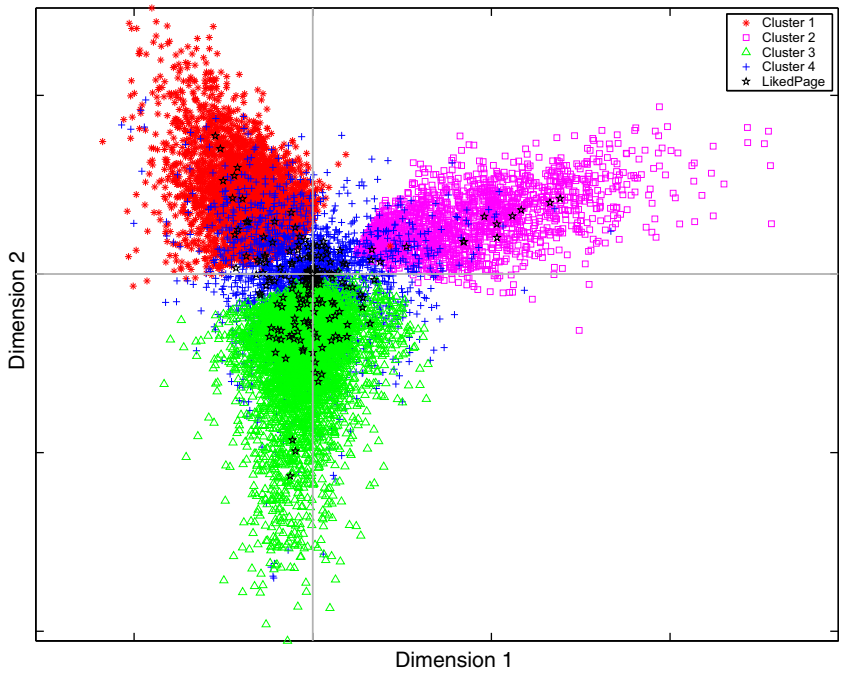

Fig. 7. MCA–Kmeans solution with attributes and subjects.

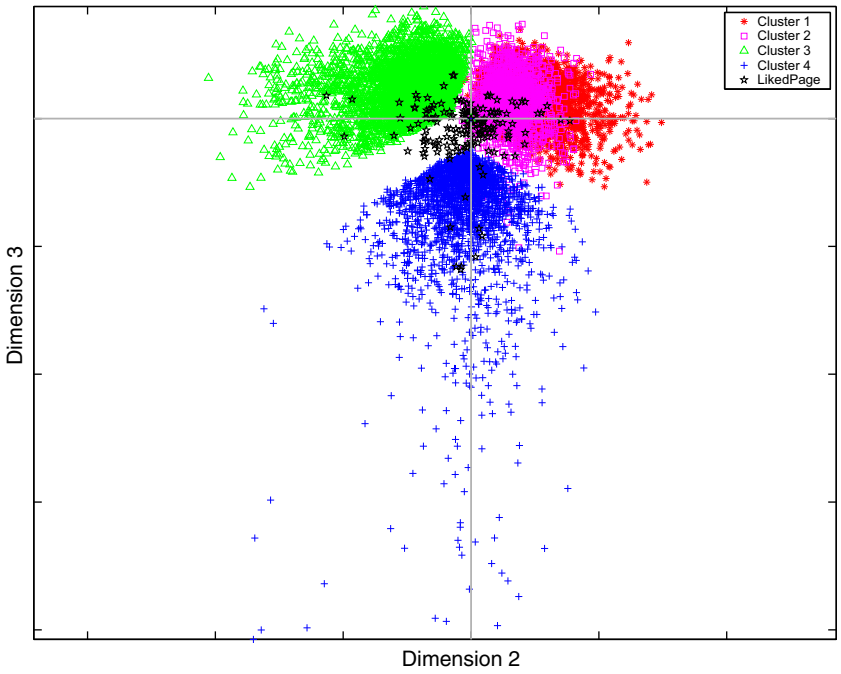

Fig. 8. MCA–Kmeans solution with attributes and subjects.

component. Cluster 2 is heavily dominated by Facebook users from Indonesia. Note that this is the only cluster in which "unknown" is not the most frequently observed country. Also, with over 62% users from Indonesia, it is by far the most homogeneous cluster concerning nationalities. Facebook users from India are over represented in the third cluster. For the fourth cluster there does not appear to be a strong geographical link.

6. Conclusions

In a relatively short time, social network sites have become an important part of daily life for millions of people. Consequently, such sites are considered to be an important marketing tool. Interviews with marketing and customer relationship managers reveal that a clear strategy regarding the social network sites often does not exist

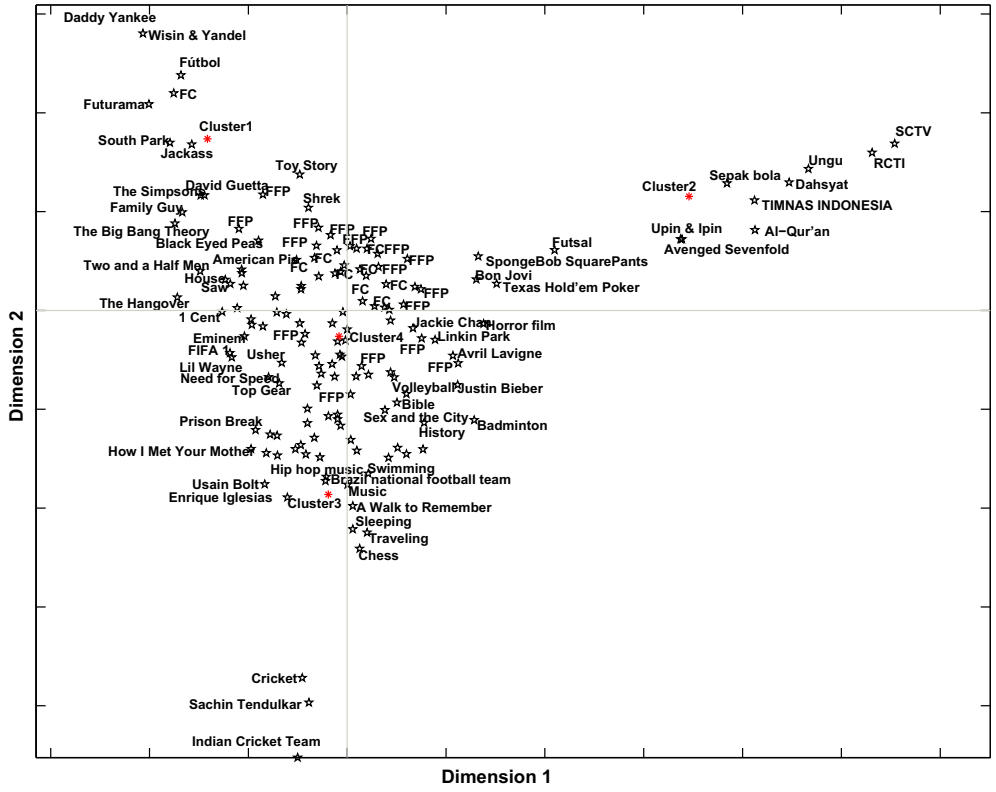

Fig. 9. MC–Kmeans plot with cluster means and liked Facebook pages. FC labels denote football clubs, FFP indicates famous players.

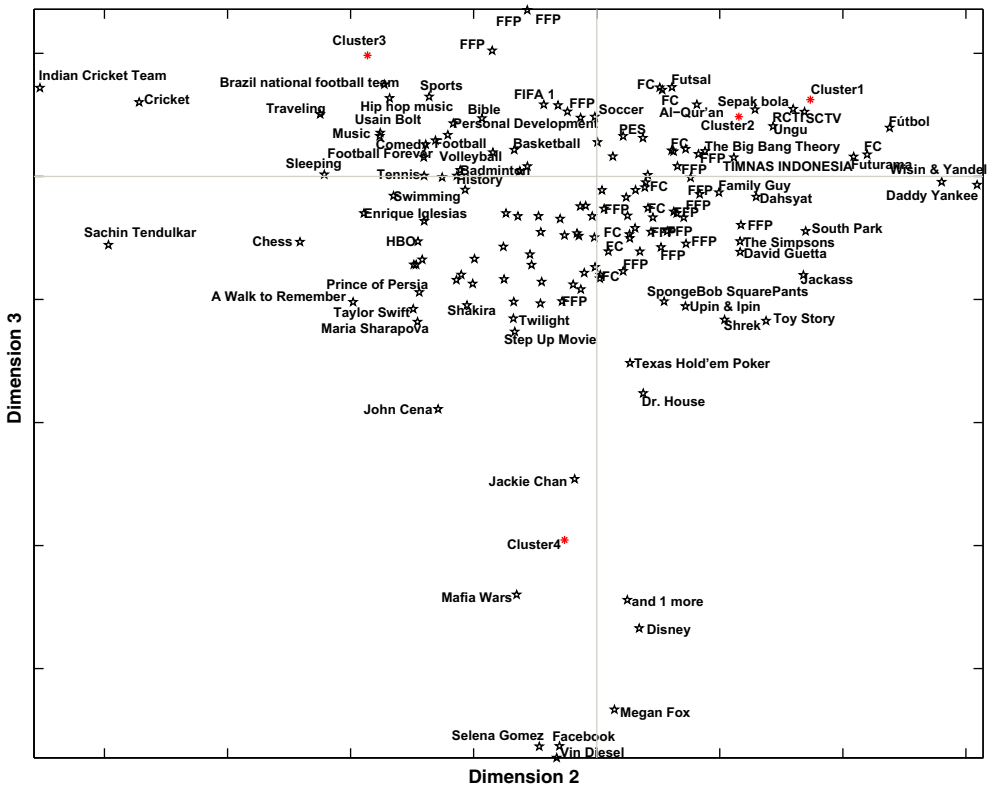

Fig. 10. MCA–Kmeans plot with cluster means and attributes (liked Facebook pages). FC labels correspond to clubs, FFP labels correspond to famous football players.

and managers are often unable to use the social network data in their customer relationship management strategy.

In this paper, we formulated a data collection framework for retrieving online profile data from Facebook users. In particular, we showed how a Facebook page owner, that is, a person or company with administrator rights to the Facebook page, can find other Facebook users that indicated liking their page. Then, by visiting the pages of such users, individual level data can be collected.

We applied the data collection framework to obtain a sample of Facebook users who indicated "liking" a large international football club. Then, using a joint dimension reduction and clustering approach, clusters could be identified on the basis of the users' liking patterns. Four clusters were obtained that differed with respect to the liking patterns. Moreover, the visualizations immediately exposed how the clusters differentiated themselves. In particular, differences in relative popularity of non football related Facebook pages characterize the different clusters. Furthermore, the clusters appear to be separated along geographical lines. That is, although no geographical data were used, the clusters differentiated themselves along Facebook pages of locally popular music/tv/sport stars. The popularity of certain pages in only certain (or one) clusters, could be used to formulate better targeted, differentiated, marketing strategies.

Table 3
Top 10 pages per cluster.

Cluster 1	Cluster 2	Cluster 3	Cluster 4
Club 1	Club 5	Football	Club 3
Club 2	Player 2	Club 1	Club 2
Player 1	Club 2	Club 2	Player 2
Player 2	Club 1	Club 3	Club 1
Club 3	Timnas Indonesia	Player 2	Club 4
Fútbol	Harry Potter	Player 1	Harry Potter
Club 4	Player 3	Harry Potter	And 1 more
South Park	Club 4	Club 4	Player 5
The Simpsons	Player 1	Player 6	Player 1
Club 5	Player 4	AKON	Club 5

6.1. Implications for research

In the CRM literature [33,27,38] customer identification is considered as the first step in a CRM cycle. Typically, the identification concerns directly observable customer generated content (e.g., transaction data). The identification of potential rather than actual customers as implemented in the data collection framework presented in this paper, offers several new research opportunities. It would, for example, be interesting to study the added value and incorporation of the proposed framework into existing CRM systems. Merging the online profile data from the (potential) customers as obtained from social media, with actual transaction data, offers other research opportunities. Moreover, tracking the profiles over time allows researchers to study effects of targeted marketing efforts in a structural fashion.

The data collection framework presented in this paper was designed specifically for Facebook. However, Facebook is not the only social media platform on which individuals provide information about their preferences and personal backgrounds. Similar ideas and methods can perhaps be used to obtain, publicly available data from other social media platforms (e.g., Twitter, Instagram, Google Plus, LinkedIn). It may in fact depend on the firm and its product which social media outlet is the most interesting.

6.2. Implications for practice

Despite the often acknowledged potential of social network data, most Facebook related marketing research relies on (online) questionnaires and/or focus groups rather than directly exploiting social network data. One reason for this situation concerns the limited possibilities for

Table 4
Top 10 countries per cluster (with cluster sizes) and clusterwise relative frequencies per country.

Cluster 1 (3549)		Cluster 2 (1734)		Cluster 3 (4048)		Cluster 4 (2381)		All (11,712)	
Unknown	21.67	Indonesia	62.11	Unknown	16.67	Unknown	21.50	Unknown	18.53
Mexico	9.30	Unknown	12.34	India	16.30	Indonesia	11.68	Indonesia	15.45
USA	6.03	Malaysia	5.94	Indonesia	8.37	Malaysia	6.34	India	7.03
Colombia	5.04	UK	2.48	Malaysia	5.09	India	5.12	Malaysia	4.46
Brazil	3.38	USA	2.13	Nigeria	4.47	UK	4.49	USA	3.94
UK	3.35	Thailand	1.04	UK	4.47	Egypt	3.02	UK	3.84
Indonesia	3.27	Turkey	0.98	USA	3.53	USA	2.86	Mexico	3.63
Argentina	3.07	Spain	0.92	Egypt	2.67	Mexico	2.48	Nigeria	2.23
Chile	2.70	Brazil	0.75	Brazil	2.03	Nigeria	2.10	Brazil	2.09
France	2.51	Mexico	0.58	Iran	1.90	Algeria	1.68	Egypt	2.00
Total	60.33	Total	89.27	Total	65.51	Total	61.28	Total	63.20

directly retrieving data from social network sites. In this paper, we formulated a data collection framework for retrieving online profile data from Facebook users. We showed how a Facebook page owner, that is a person or company with administrator rights to the Facebook page, can find other Facebook users that indicated liking their page. Then, by visiting the pages of such users, individual level data can be collected.

The proposed data collection framework has direct potential for marketing managers as it makes it possible to investigate whether distinct clusters requiring distinct marketing efforts can be identified among potential customers (that is, users that already showed some form of affiliation to the company by "liking" it on Facebook). Hence, the general framework presented in this paper can be used to improve and enhance implementation of the identification phase in a firm's CRM process. In particular, by focusing on potential rather than existing customers, information becomes available that can be used to improve marketing efforts aimed specifically at acquisition of new customers. Ideally, a system should be implemented that merges the online profile data with other available profile data (e.g., profiles based on transaction data).

6.3. Limitations and future research directions

The proposed data collection framework only allows for the retrieval of data from users with public profiles. Moreover, from the sample of users with a public profile, we selected users that "liked" at least five of the most popular 150 Facebook pages. The sample analyzed in this paper therefore does not necessarily represent the population of Facebook users who "liked" the football club. Instead, the sample only represents active main stream users.

Another important issue, inherent to some extent to Facebook and other "new" media, concerns the rapid developments that may overtake current research. In our case, since collecting the data, February 2011, and finalizing this paper, the number of users who "liked" the Facebook page increased from around 4 million to 21 million. More importantly, however, given the steady increase of Facebook users, it may very well be the case that current users differ from previous users in their usage of Facebook. As we only received administrator rights for a short period of time, we did not study such changes. However, the data collection framework makes it possible to easily track such changes and act upon them.

In this paper, we considered a clustering analysis based solely on liking patterns of Facebook users. Although such indicated liking patterns require very little effort from the users, they are considered as so-called reactive data. It would be interesting to see whether the reactive data can be augmented by non-reactive data. For example, considering network data (i.e. by incorporating data concerning the connections between users), and/or by other data available on users' Facebook pages (e.g., posted messages/links/pictures, etc.). Augmenting the data in such a fashion, may yield even richer and more challenging data sets.

Finally, it should be noted that other social network related applications can also benefit from our data collection framework. For example, recently, [11] considered targeting strategies directed towards individuals in a social network using data obtained directly from a large social network site. Their analysis could be extended by using Facebook data obtained after application of our methodology.

Appendix A. Identifying Facebook page fans

Require: The parameters:

- α, string with the fb_dtsg variable
- β, string with the post_form_id variable
- γ, string containing FB page's administrator cookie
- δ, string with the Facebook fan page ID
- ϵ, string with the HTTP request URL without the α, β, γ, and δ parameters

Require: HTTPrequest $(\alpha, \beta, \gamma, \delta, \epsilon)$ performs a HTTP request (being a Mozilla Firefox browser) to a specified Facebook URL (ϵ) with corresponding parameters $(\alpha, \beta, \gamma, \delta)$. Returns: JSON.
Require: extractFanInfo $(rawJSONresponse)$ returns an array containing each JSON records's Facebook ID, Facebook URL and name in $rawJSONresponse$
Require: savetoDatabase (fan) saves fan's Facebook ID, Facebook URL, and name to a MySQL database

```
while rawJSONresponse != NULL do
    rawJSONresponse = HTTPrequest (α, β, γ, δ, ε)
    fans = extractFanInfo (rawJSONresponse)
    for all fans as fan do
        savetoDatabase (fan)
    end for
end while
```

Appendix B. MCA–K-means

In MCA–Kmeans, the objective is to minimize a weighted average of the MCA objective and a K-means objective. The resulting objective function of MCA–Kmeans can be expressed as:

$$\begin{aligned}\min_{\mathbf{Y},\mathbf{B},\mathbf{C},\mathbf{G}} \phi(\mathbf{Y},\mathbf{B},\mathbf{C},\mathbf{G}) &= \alpha_1 \text{MCA} + (1-\alpha_1)\text{K-means} \\ &= \alpha_1 \sum_{j=1}^{q} \left|\left|\mathbf{Y}-\mathbf{Z}_j\mathbf{B}_j\right|\right|^2 + (1-\alpha_1)||\mathbf{Y}-\mathbf{CG}||^2 \\ &\quad s.t. \mathbf{Y}^{\mathrm{T}}\mathbf{Y}=\mathbf{I}_k\end{aligned}$$

where $\mathbf{Y}$ denotes the $n \times k$ group configuration, $\mathbf{Z}_j$ is the $n \times p_j$ (observed) indicator matrix for the jth variable, $\mathbf{B}_j$ is a matrix of category quantifications (attribute weights), $\mathbf{C}$ denotes the $n \times K$ cluster membership matrix and $\mathbf{G}$ gives the $K \times k$ matrix of cluster means. The number of clusters (K) and the dimensionality (k) need to be selected by the user. The α coefficient, which lies between zero and one, allows us to control for the importance of the dimension reduction part versus the clustering part. In our application we fix α to 0.5 so that both parts are equally important. An alternating least-squares algorithm can be used

to solve the minimization problem. For fixed **Y**, the category quantifications become:

$$\mathbf{B}_j = \left(\mathbf{Z}_j^{\mathrm{T}}\mathbf{Z}_j\right)^{-1}\mathbf{Z}_j^{\mathrm{T}}\mathbf{Y},$$

and, similarly, for fixed **Y** and **C**, the cluster group means can be calculated as:

$$\mathbf{G}=\left(\mathbf{C}^{\mathrm{T}}\mathbf{C}\right)^{-1}\mathbf{C}^{\mathrm{T}}\mathbf{Y}.$$

Furthermore, [20] shows that, for fixed **C**, the configuration matrix **Y** can be obtained using the eigenequation

$$\left(\alpha_1\sum_{j=1}^{q}\mathbf{Z}_j\left(\mathbf{Z}_j^{\mathrm{T}}\mathbf{Z}_j\right)^{-1}\mathbf{Z}_j^{\mathrm{T}}+\alpha_2\mathbf{C}\left(\mathbf{C}^{\mathrm{T}}\mathbf{C}\right)^{-1}\mathbf{C}^{\mathrm{T}}\right)\mathbf{Y}=\mathbf{Y}\boldsymbol{\Lambda}. \quad (1)$$

By considering the eigenvectors (i.e., the columns of **Y**) corresponding to the k largest eigenvalues, the optimal group configuration, for fixed **C**, is obtained. After updating **Y** in this fashion, the cluster membership matrix **C** is obtained by considering distances of the k-dimensional points in **Y** to cluster means in **G** and by subsequently assigning observations to the closest cluster.

Starting with some initial values for **C** and **Y** (e.g., random cluster memberships and **Y** the configuration obtained after applying MCA) the approximations are sequentially updated leading the objective to decrease monotonically. If the decrease is below a certain threshold, the algorithm terminates and a solution is obtained. To reduce the chances of obtaining a local minimum, several random starts should be applied.

Note that the eigenEq. (1) is of crucial importance in the proposed algorithm. For large n, the matrix that needs to be considered becomes large. It is therefore useful to reformulate the method in a more efficient way. This can easily be achieved by defining

$$\mathbf{X}=\left(\sqrt{\alpha_1}\mathbf{Z}\mathbf{D}_z^{-\frac{1}{2}}\quad\sqrt{(1-\alpha_1)}\mathbf{C}\mathbf{D}_c^{-\frac{1}{2}}\right),$$

where $\mathbf{Z}=[\mathbf{Z_1}, \mathbf{Z_2}, \ldots, \mathbf{Z_j}, \ldots, \mathbf{Z_q}]$, $\mathbf{D}_z = diag(\mathbf{Z}^{\mathrm{T}}\mathbf{Z})$ and $\mathbf{D}_c = \mathbf{C}^{\mathrm{T}}\mathbf{C}$.

If we consider the singular value decomposition

$$\mathbf{X}=\mathbf{Y}\boldsymbol{\Lambda}^{\frac{1}{2}}\mathbf{V}^{\mathrm{T}},$$

where $\mathbf{Y}^{\mathrm{T}}\mathbf{Y}=\mathbf{V}^{\mathrm{T}}\mathbf{V}=\mathbf{I}$, we get, in accordance with Eq. (1),

$$\mathbf{X}\mathbf{X}^{\mathrm{T}}\mathbf{Y}=\mathbf{Y}\boldsymbol{\Lambda}$$

and

$$\mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{V}=\mathbf{V}\boldsymbol{\Lambda}. \quad (2)$$

The group configuration **Y** can thus be obtained as

$$\mathbf{Y}=\mathbf{X}\mathbf{V}\boldsymbol{\Lambda}^{-\frac{1}{2}}. \quad (3)$$

Finally, although not specifically mentioned in [20], it is important to consider all **Z** matrices in deviation from the mean vector to avoid a so-called trivial solution. Alternatively, the trivial solution, i.e. the eigenvector corresponding to the largest eigenvalue of $\mathbf{X}^{\mathrm{T}}\mathbf{X}$, should be ignored. An important advantage of using Eq. 2 over Eq. 1 is that we only need to find the $k+1$ largest eigenvalues and corresponding eigenvectors for the $Q \times Q$ matrix $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ rather than for the $n \times n$ matrix $\mathbf{X}\mathbf{X}^{\mathrm{T}}$.

References

[1] H. Arabie, L. Hubert, Cluster analysis in marketing research. Factorial k-means analysis for two-way data, in: R. Bagozz (Ed.), Advanced Methods of Marketing Research (160–189), Blackwell, Oxford, 1994.
[2] E.D. Argaez, http://www.internetworldstats.com/facebook.htm2011 (Accessed on 2 June 2011).
[3] S. Ba, H.R. Rao, DSS special issue on the theory and applications of social networks, Decision Support Systems 55 (4) (2013) 939–940 (1. Social Media Research and Applications 2. Theory and Applications of Social Networks).
[4] C.H. Baird, G. Parasnis, From social media to social customer relationship management, Strategy and Leadership 39 (2011) 30–37.
[5] R. Chakraborty, C. Vishik, H.R. Rao, Privacy preserving actions of older adults on social media: exploring the behavior of opting out of information sharing, Decision Support Systems 55 (4) (2013) 948–956 (<ce:title > 1. Social Media Research and Applications 2. Theory and Applications of Social Networks</ce:title>).
[6] C.M. Cheung, M.K. Lee, A theoretical model of intentional social action in online social networks, Decision Support Systems 49 (1) (2010) 24–30.
[7] J. Claussen, T. Kretschmer, P. Mayrhofer, The effect of rewarding user engagement: the case of Facebook apps, Information Systems Research 24 (2013).
[8] W. Duan, Special issue on social media: an editorial introduction, Decision Support Systems 55 (4) (2013) 861–862 861–862. 1. Social Media Research and Applications 2. Theory and Applications of Social Networks.
[9] Facebook.com., http://developers.facebook.com/2011 (Accessed on 16 March 2011).
[10] C. Forman, A. Ghose, B. Wiesenfeld, Examining the relationship between reviews and sales: the role of reviewer identity discloser in electronic markets, Information Systems Research 19 (2008) 291–313.
[11] Gelper, S., Lans, R. Van der Van Bruggen, G. 2014. Competition for attention in online social networks: implications for viral marketing (unpublished manuscript).
[12] A. Gifi, Nonlinear Multivariate Analysis, Wiley, Chichester, 1990.
[13] Google, http://code.google.com/apis/maps/index.html2004 (Accessed on 3 May 2011).
[14] M.J. Greenacre, Correspondence Analysis in Practice, Academic Press, London, 1993.
[15] R. Gross, A. Acquisti, Information revelation and privacy in online social networks (the Facebook case), Proceedings of the 2005 ACM Workshop on Privacy in the Electronic Society2005. 71–80.
[16] L. Harris, C. Dennis, Engaging customers on Facebook: challenges for e-tailers, Journal of Consumer Behaviour 10 (2011) 338–346.
[17] T. Hennig-Thurau, K.P. Gwinner, G. Walsh, D.D. Gremier, Electronic word-of-mouth via consumer-opinion platforms: what motivates consumers to articulate themselves on the Internet, Journal of Interactive Marketing 18 (2004) 38–52.
[18] T. Hennig-Thurau, E.C. Malthouse, C. Friege, Gensler S. Lobschat, A. Rangaswamy, B. Skiera, The impact of new media on customer relationships, Journal of Service Research 13 (2010) 311–330.
[19] S. Ho, D. Bodoff, K. Tam, Timing of adaptive web personalization and its effects on online consumer behavior, Information Systems Research 22 (2011) 660–679.
[20] H. Hwang, W.R. Dillon, Y. Takane, An extension of multiple correspondence analysis for identifying heterogenous subgroups of respondents, Psychometrika 71 (2006) 161–171.
[21] A. Iodice D' Enza, F. Palumbo, Iterative factor clustering of binary data, Computational Statistics 1–19 (2012), http://dx.doi.org/10.1007/s00180-012-0329-x.
[22] A.H. Kracklauer, D.Q. Mills, D. Seifert, Customer management as the origin of collaborative customer relationship management, Collaborative Customer Relationship Management, Springer, 2004, pp. 3–6.
[23] C. Lampe, N. Ellison, C. Steinfield, A familiar face(book): profile elements as signals in an online social network, Proceedings of Conference on Human Factors in Computing Systems, ACM Press, 2007, pp. 435–444.
[24] K. Lewis, J. Kaufman, M. Gonzalez, A. Wimmer, N. Christakis, Tastes, ties, and time: a new social network dataset using Facebook.com, Social Networks 30 (4) (2008) 330–342.
[25] Y.-M. Li, Y.-L. Shiu, A diffusion mechanism for social advertising over microblogs, Decision Support Systems 54 (1) (2012) 9–22.
[26] J. MacQueen, Some methods for classification and analysis of multivariate observations. In L. Cam & J. Neyman (Eds.), Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability (1, 281–297), University of California Press, California, 1967.
[27] E.W. Ngai, L. Xiu, D.C. Chau, Application of data mining techniques in customer relationship management: a literature review and classification, Expert Systems with Applications 36 (2) (2009) 2592–2602.
[28] S. Okazaki, What do we know about mobile internet adopters? A cluster analysis, Information Management 43 (2006) 127–141.
[29] N. Park, S. Lee, J.H. Kim, Individuals' personal network characteristics and patterns of Facebook use: a social network approach, Computers in Human Behavior 28 (2012) 1700–1707.
[30] A. Parvatiyar, J.N. Sheth, Customer relationship management: emerging practice, process, and discipline, Journal of Economic and Social Research 3 (2) (2001) 1–34.
[31] R. Rishika, A. Kumar, R. Janakiraman, R. Bezawada, The effect of customers' social media participation on customer visit frequency and profitability: an empirical investigation, Information Systems Research 24 (2013).
[32] M.J. Shaw, C. Subramaniam, G.W. Tan, M. Welge, Knowledge management and data mining for marketing, Decision Support Systems 31 (2001) 127–137.
[33] R.S. Swift, Accelerating Customer Relationships: Using CRM and Relationship Technologies, Prentice Hall Professional, 2001.

[34] S. Van Buuren, W. Heiser, Clustering n objects into k groups under optimal scaling of variables, Psychometrika 54 (1989) 699–706.
[35] M. Vichi, H. Kiers, Factorial k-means analysis for two-way data, Computational Statistics and Data Analysis 37 (2001) 49–64.
[36] M. Wick, http://download.geonames.org/export/dump/readme.txt2005 (Accessed on 28 April 2011).
[37] K.P. Wiedmann, H. Buxel, G. Walsh, Customer profiling in e-commerce: methodological aspects and challenges, Journal of Database Marketing 9 (2) (2002) 170–184.
[38] M. Xu, J. Walton, Gaining customer knowledge through analytical CRM, Industrial Management Amp; Data Systems 105 (7) (2005) 955–971.
[39] F. Zhu, X. Zhang, Impact of online consumer reviews on sales: the moderating role of product and consumer characteristics, Journal of Marketing 74 (2010) 133–148.

Jan-Willem van Dam obtained his Master's degree Cum Laude in Economics & Informatics from Erasmus University Rotterdam, the Netherlands, in 2011. The focus of his Master's thesis is on employing data mining techniques for enhancing sport marketing applications. His research interests cover areas such as data mining, Web 2.0, the Semantic Web foundations and applications, and Web information systems.

Michel van de Velden is an assistant professor at the Econometric Institute of the Erasmus University Rotterdam. His research interests concern development and application of visualization methods for multivariate data. His work covers a wide range of research disciplines ranging from linear algebra to transportation science, and has been published in an equally wide range of high standing academic journals including Linear Algebra and its Applications, Psychometrika, Journal of Computational and Graphical Statistics, Journal of Statistical Software and Marketing Letters. For a full CV and list of publications, please visit, http://people.few.eur.nl/vandevelden/