Available online at www.sciencedirect.com

ScienceDirect

Procedia Engineering 131 (2015) 1140 – 1145

Procedia Engineering

www.elsevier.com/locate/procedia

World Conference: TRIZ FUTURE, TF 2011-2014

Fast Lead User Identification Framework

Sanjin Pajo*, Paul-Armand Verhaegen, Dennis Vandevenne, Joost R. Duflou

Center for Industrial Design, KU Leuven, Celestijnlaan 300 bus2422, Heverlee 3001,Belgium

Abstract

Large portion of product innovation and development is accomplished by customers and only a small segment of the customer population engages in such innovation activities. Empirical research has shown that users in this subgroup, called lead users, tend to experience needs before the rest of the marketplace and stand to benefit greatly by finding solutions to those needs. To meet the challenge of quickly and effectively identifying lead users and uncovering their innovation ideas, the authors propose a fast and systematic approach, called Fast Lead User IDentification (FLUID), utilizing data mining techniques to identify lead users on social networking sites. The paper describes the steps taken to build and optimize the FLUID system to effectively identify lead users on the micro-blogging site Twitter. This entails studies using validated lead user questionnaires resulting in clusters of lead and non-lead Twitter users for a single product. The gathered online user metadata and behavior are then used as training data for the automated system. An overview of data processing techniques and relations to the empirically derived lead user characteristics are presented. Finally, classification algorithms that help to separate lead users from non-lead users are discussed, including optimization leading to the validation of the proposed approach. By making use of data-mining techniques on data rich sites like social networking sites, the FLUID approach minimizes the resource and time costs in identifying lead users and this provides a step towards systematizing the fuzzy-front end of the new product development process.

Peer-review under responsibility of the Scientific Committee of TFC 2011, TFC 2012, TFC 2013 and TFC 2014 – GIC

Keywords: lead user, data mining, social media, systematic innovation, fuzzy-front end.

1. Introduction

In a rapidly changing marketplace companies are looking to stay competitive by quickly identifying and meeting the emerging customer needs. A small portion of customers, less than 7%, called lead users, experience new

* Corresponding author. Tel.: +32 (0) 16 37 27 82; fax: +32 (0) 16 32 29 86.
E-mail address: sanjin.pajo@kuleuven.be

Peer-review under responsibility of the Scientific Committee of TFC 2011, TFC 2012, TFC 2013 and TFC 2014 – GIC
doi:10.1016/j.proeng.2015.12.434

needs before the rest of the marketplace and stand to benefit greatly by finding solutions to those needs [1, 2]. They actively engage in innovation activities and bring to fruition novel and commercially attractive solutions [3].

The initial approaches to systematically identifying lead users consist of surveying methods like screening, broadcasting and pyramiding that make use of questionnaire and interview methods, and involve experts [4 - 7]. Recent approaches, like Netnography have looked to the web for finding lead users although these approaches also rely on experts to analyze vast amounts of collected user online data [8, 9].

In applying surveying methods, researchers have identified a set of measurable characteristics to systematically separate lead users from non-lead users. The initial characteristics ascribed to lead users by Von Hippel, who coined the term lead users, are 'ahead of trend' in experiencing needs before others and 'high expected benefit' by finding solutions to meet those needs [2, 10, 11]. Lead users also actively engage in innovation activities, a characteristic known as 'investment' [2, 10]. Other empirical research also shows that lead users have extensive technical expertise, called 'product knowledge' and 'use experience' [12, 13]. Another crucial characteristic is 'opinion leadership', meaning that lead users support information flow between customers and thereby also help diffuse new solutions into the marketplace [13, 14].

Although the uncovered characteristics allow for effective and systematic lead user identification, traditional methods like surveying approaches are time and resource consuming. The identification process can last half a year and requires task intensive analysis by experts [11]. Web based methods like Netnography point to great opportunities in systematic and automated lead user identification online, and growing social networking sites are great sources of publicly available rich data [8, 9]. In the next section, a framework to formulate a new semi-automated approach, called Fast Lead User IDentification (FLUID), that makes use of data mining techniques to systematically identify lead users, is described.

2. Framework

The purpose behind the proposed FLUID approach is to automatically classify online users into two groups, lead and non-lead users based on online posts and metadata. Twitter was selected as the data source due to the large amounts of publicly available data including an easy to access Application Program Interface (API). A Java based tool, named FLUID after the approach was built to test the feasibility of the approach and to evaluate the identification process. For planned industrial cases to build the framework, the procedure consists of keyword formulation, data collection and pre-processing, and classification leading to validation of identified lead users in close collaboration with the involved companies. The following subsections each detail a stage of this procedure.

2.1. Keyword Formulation

During the information gathering stage, a set of keywords is collected to be used as query terms to retrieve data from Twitter. Currently, keywords are provided by the stakeholders, i.e. by the company design team, and are weighted based on the value to the company. The keywords typically represent consumer language, which is constantly and rapidly evolving and changing. This step is the only non- automated part of the process. The keywords are used as query terms in the next step, data collection.

2.2. Data Collection

The second step in the FLUID process is retrieval and storage of relevant user metadata and tweets. The tool makes use of the Twitter search engine with the keyword query to find relevant tweets and potential lead users. The query consists of two or more terms, product name, i.e. lens and relevant keywords, for example product features, i.e. glass. Using only a single term, for example the product name, leads to a poor performance with tweets that are polluted with non- relevant information or information from other domains. The retrieved data for each user includes the user's timeline, i.e. tweets, and user metadata, i.e. number of followers, friends, tweets, list of user connections, etc. Before storage, data are verified for completeness, accuracy and integrity with incomplete or inaccurate records removed and missing values flagged, as users oftentimes fail to provide complete information. Verified and complete data are loaded into a database for processing.

2.3. Data Pre-Processing

The data pre-processing stage consists of filtering of non- relevant data and feature extraction. Before pre-processing starts, users in the database can be filtered out based on stakeholder set preferences (i.e. location, language, type, etc.). Filters like the language filter may reduce the overall number of lead users found and will be reported with the results for each industrial case. Additionally, tweets are tokenized and filtered by removing stop-words [15] and performing a stem operation [16]. For each retained user, automatically generated metadata and user inputted metadata are given values. The agglomerated user values can be categorized into four groups: engagement, sentiment, relevance and influence. Category engagement pertains to user behavior, for example number of hashtags per tweet and number of followers per day. The sentiment category contains percentages of positive, negative and neutral tweets based on emotional disposition analysis of the user timeline [17]. Relevance or expertise category includes values of estimated relevance of tweets and hashtags to the set of keywords collected during the information gathering stage. Relevance is estimated using the Okapi BM25 ranking function [18], which reflects the significance of a term to a tweet in a collection of tweets. The remaining category influence includes values based on network graph analysis, for example betweeness centrality [19], degree centrality [19], closeness centrality [19] and eigenvector centrality [19] to estimate influence of a user in the social network in terms of supporting flow of product related information. Impact scores are standardized (0, 1) and then normalized before the next stage, data classification.

2.4. Classification

After pre-processing, a prediction is made whether a user is a lead or non-lead user by using the statistical classifier with the splitting criterion being the normalized difference in entropy. The next sub-section describes the generation of the training set for the classification process.

2.4.1. Training Set

To identify lead users on Twitter a self-administered questionnaire was used. The questionnaire consisted of 31 items that measure for the previously described characteristics (see Section 1): dissatisfaction with existing product in the marketplace, product knowledge and use experience, the opinion leadership and ahead of trend. The respondents were asked to read each statement and specify their level of agreement on a 5 point Likert scale. The five dimensions were obtained from validated surveys and have been proven reliable and valid in the previous studies [4, 20-23].

The survey population consisted of users that discuss or follow camera lens product related topics on Twitter. Lead users were identified by calculating the mean for each construct and comparing it to mean value of the construct for all the users. Sample mean for a construct is commonly used in lead user studies to separate lead users from other customers [20]. The extracted respondent Twitter metadata and tweets are used as training data.

2.5. Validation

Two step validation is planned for each industrial test case. The first step is the evaluation of Twitter user tweets by the company. Company representatives are given sample data of an equal number FLUID identified lead users and non-lead users, including questionnaire items that measure for characteristics of being a lead user. Analysis of the completed questionnaire would give an early indication about the performance of the approach. The second step of the validation process is contacting and engaging users in idea generation and evaluating the quality of the generated ideas.

3. Training Set Cross-Validation

This section describes the method and results of the cross validation of the statistical classifier C4.5 used to label users based on the online measured data. Cross validation is a manner of evaluating the performance of a machine learning algorithm. With a 10 fold cross validation, the training data set is divided once and into 10 equally large subsets with 9 subsets used for training and one subset for validation. This is repeated 10 times using a different subset for validation each time. Because of a small data set, stratified cross validation is used, ensuring that each fold has the right proportion of each class value. Table 1 shows the resulting confusion matrix using the training set (LU – lead user, NLU – non-lead user). The overall accuracy is 98%, with precision and recall 0.9 and 1, respectively.

Table 1. Confusion matrix.

		Predicted class	
		LU	**NLU**
Actual Class	**LU**	9	0
	NLU	1	48

C4.5 algorithm generates a decision tree where each node shows the attribute that most effectively splits the data into lead user or non-lead user class. The generated decision tree is shown in the Appendix A. The significant attributes are: relevance of user's description (DescrRSsn), relevance of tweets (TweetRsn), relevance and average mention rates (MentionRsn, MentionRRsn) and sentiment (PositiveSsn, NegativeSsn). After cross-validation, the tool outputs an extra model built on the entire set that can be deployed in practice.

4. Conclusion

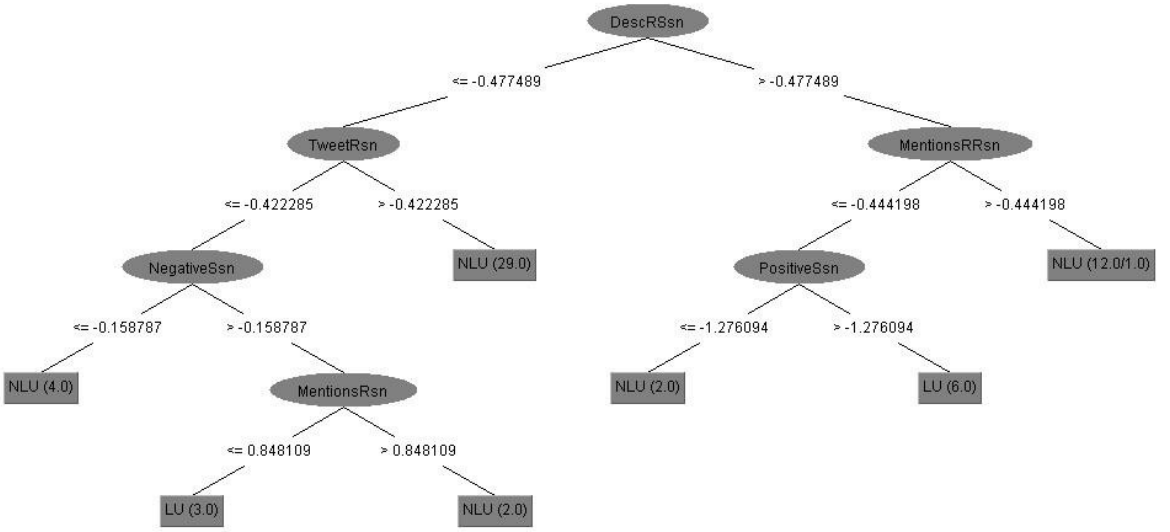
As indicated in the previous section, the accuracy in classifying lead users in the training set is 98%, with precision and recall 0.9 and 1 respectively. The effectiveness of the classifier is high, demonstrating a substantial agreement between the results given by the questionnaire and by the tool classifier. Additionally, the decision tree shows that lead user characteristics like sentiment, relevance and influence play a significant role in the classification process. The engagement attributes on the other hand, were not found to be significant factors in separating lead users from non-lead users.

The results should be interpreted cautiously as the cross- validation results are based on a single study and further studies are needed for a more indicative validation. Restricting the study to one product may restrict the generalizability of the results. As indicated in the Section 2, several industrial empirical studies are planned to show that lead users have a presence on the social networking site Twitter and that they can be effectively identified using the FLUID approach. The classification process can be further refined pending additional studies.

Furthermore, to fully automate the approach, it needs to acquire new and relevant keywords or terminology that can be used as part of the query set and for textual relevance analysis. The nature of language in social media is highly dynamic, with informal and rule breaking terminology that is constantly evolving.

The described research contributes to the innovation management and design field by offering a systematic framework to identifying human resources necessary for the innovation process. It offers a tangible alternative to the existing approaches like the lifestyle and TRIZ trends. The results support further research in systematic identification of users with innovation ideas online. Early cross-validation indicates achievability of systematic selection of lead users from an online customer population. Companies are increasingly looking at customers as a source of new innovation ideas whereby systematic and automated approaches can offer a fast and an effective solution.

Appendix A. Training data decision tree



Decision tree key table:

Item	Meaning
DescRSsn	Relevance of the user description to the set of keywords (standardized and
TweetRsn	Relevance of the user tweet timeline to the set of keywords (standardized and
MentionsRRsn	Rate of relevant mentions, per tweet (standardized and normalized)
NegativeSsn	Percent of tweets with a negative disposition (standardized and normalized)
PositiveSsn	Percent of tweets with a positive disposition (standardized and normalized)
MentionsRsn	Rate of mentions, per tweet (standardized and normalized)
LU	Lead User
NLU	Non-Lead User

References

[1] von Hippel E., Successful Industrial Products from Customer Ideas. Journal of Marketing; 1978; 42, 1:39–49.
[2] von Hippel E. Lead Users: A Source of Novel Product Concepts. Management Science; 1986; 32, 7:791-806.
[3] Schreier M, Prügl R. Extending Lead User Theory: Antecedents and Consequences of Consumer's Lead Userness. Journal of Product Innovation Management; 2008; 25:331-46.
[4] Bilgram V, Brem A, Voigt KI. User-Centric Innovations in New Product Development - Systematic Identification of Lead Users Harnessing Interactive and Collaborative Online Tools. International Journal of Innovation Management; 2008; 12,3:419-458
[5] Von Hippel E, Franke N, Prügl R.. 'Pyramiding': Efficient Identification of Rare Subjects. MIT Sloan Research Paper No. 4720-08;2008.
[6] Lakhani K. Broadcast search in problem sovling: attracting solutions from the periphery. MIT Soan School of Management. Working Paper; 2006.
[7] Hienerth C, Poetz M, von Hippel E. Exploring key characteristics of lead user workshop participants: Who contributes best to the generation of truly novel solutions? DRUID Summer Conference, Copenhagen, Denmark; 2007.
[8] Kozinets R. E-Tribalized Marketing? The Strategic Implications of Virtual Communities of Consumption. European Management Journal; 1999; 17, 3:252-264.

[9] Kozinets R. The Field Behind the Screen: Using Netnography for Marketing Research in Online Communities. Journal of Marketing Research; 2002; 39:61-72.
[10] Urban GL., von Hippel E. Lead user analysis for the development of new industrial products. Management Science; 1988;. 34, 5:569-582.
[11] von Hippel E, Thomke S, Sonnack M. Creating Breakthroughs at 3M. Harvard Business Review; 1999; 77, 5
[12] Lüthje C. Characteristics of Innovating Users in a Consumer Goods Field: An Empirical Study of Sport-Related Product Consumers. MIT Sloan Working Paper No. 4331-02; 2004.
[13] Schreier M, Prügl R. Extending Lead User Theory: Antecedents and Consequences of Consumer's Lead Userness. Journal of Product Innovation Management; 2008, 25:331-46.
[14] Franke N, Shah SK. How Communities Support Innovative Activities: An Exploration of Assistance and Sharing Among End-Users. Research Policy; 2003; 32:157-178.
[15] Wang X. Source Code: Stopwords.java – Java API Examples. University of Waikato, Hamilton, New Zealand; 2012. Web. 14 November 2013.
[16] Wikipedia contributors. Stemming [Internet]. Wikipedia, The Free Encyclopedia; 2014 May 10, 13:07 UTC [cited 2014 May 14]. Available from: http://en.wikipedia.org/w/index.php?title=Stemming&oldid=607906946.
[17] Cui A, Zhang M, Liu Y, Ma S. Emotion tokens: Bridging the gap among multilingual twitter sentiment analysis. Information Retrieval Technology; 2011; 238–249.
[18] Wikipedia contributors. Okapi BM25 [Internet]. Wikipedia, The Free Encyclopedia; 2014 Apr 15, 14:42 UTC [cited 2014 May 15]. Available from: http://en.wikipedia.org/w/index.php?title=Okapi_BM25&oldid=6043087 32.
[19] Wikipedia contributors. Centrality [Internet]. Wikipedia, The Free Encyclopedia; 2014 May 12, 12:21 UTC [cited 2014 May 14]. Available from: http://en.wikipedia.org/w/index.php?title=Centrality&oldid=608213570..
[20] Spann M, Ernst H, Skiera B, Soll JH. Identification of lead users for consumer products via virtual stock markets. Journal of Product Innovation Management; 2009; 26:322–335.
[21] Cate E. Patterns in Lead User Innovation: the Case of Winter Windsurfing, MS Thesis Aarhus School of Business. Aarhus University; 2014; Web.
[22] Franke N, Von Hippel E, Schreier M. Finding commercially attractive user innovations: a test of lead-user theory. Journal of Product Innovation Management; 2006; 23:301–315.
[23] King, CW, Summers JO. Overlap of Opinion Leadership A cross Consumer Product Categories. Journal of Marketing Research; 1970; 7:43-50.
[24] Wikipedia contributors. Accuracy and precision [Internet]. Wikipedia, The Free Encyclopedia. 2014 May 12, 13:22 UTC [cited 2014 May 14]. Available from: http://en.wikipedia.org/w/index.php?title=Accuracy_and_precision&oldi d=578217217