Available online at www.sciencedirect.com
ScienceDirect
Procedia Engineering 131 (2015) 509 – 513
Procedia Engineering
www.elsevier.com/locate/procedia

World Conference: TRIZ FUTURE, TF 2011-2014

Towards Automatic and Accurate Lead User Identification

Sanjin Pajo[a], Paul-Armand Verhaegen[a], Dennis Vandevenne[a], Joost R. Duflou[a]

[a]*Katholieke Universiteit Leuven, Mechanical Engineering Department, Celestijnenlaan 300 bus 2422, Heverlee 3001, Belgium*

Abstract

Identifying product development opportunities is one of the initial phases in the fuzzy front end of a creative innovation process. A number of approaches have been developed to systemize this process, e.g., lifestyle trends, documenting customer complaints, TRIZ trends, and lead user identification and interviewing. Lead users are an essential group of users for companies to involve as part of knowledge management in product design and evaluation. A lead user is an active and experienced individual engaged in modifying and developing products for personal gain, making systematic identification of a lead user the principal basis to precipitate product innovation and development. Currently, some companies, e.g., 3M, employ an interactive screening method, a time intensive activity consisting of semi-structured telephone interviews and written questionnaires by expert users to select lead users [1]. Another more recent approach, netnography, entails systematic analysis of online communities by observing, collecting and laboriously interpreting data by expert users [2]. In contrast to these two time and cost sensitive methods, in this paper a Fast Lead User Identification (FLUID) systematic approach for an automated and accurate discovery of lead users, by exploiting data mining methods and algorithms on rich social networking sites like the micro blogging site Twitter, is proposed. Social sites provide easy access to rich data for extraction and analysis of extended user information and behavior, enabling assessment and classification of lead users with dissimilar expertise as an early algorithmic step of creative innovation. In addition to the discussion of the three aforementioned methods, steps for validating the FLUID method are outlined.

Keywords: Lead user, data mining, netnography, screening, FLUID;

1. Lead Users

End-users that actively engage in modifying, developing, and advancing products or services for which they see personal benefits are called lead users [3][4]. Significant portion of innovation comes from experienced and ardent lead users that can be of great value in product innovation and development to companies [5]. With their extensive knowledge of a product or a service and capacity for substantive solutions or prototypes they supplement the available resources within TRIZ. In a large customer base, finding lead users can be difficult as

doi:10.1016/j.proeng.2015.12.445

they are not limited to the experts in the target product field. The goal of the paper is to formulate a model for a systematic approach for identifying lead users to be able to explore their knowledge and experience and possibly uncover innovative ideas. Most approaches look at a large number of potentially relevant users, a large population with scarcity of information on their activities, which poses a major challenge. Belz [2] and Lüthje [6] [5] advocate using six characteristics for better identification of lead users in consumer goods markets, ahead of trend, dissatisfaction, product- related knowledge, use experience, involvement and opinion leadership. The characteristics emphasize the importance of product knowledge and engagement in product development that is profitable for an end user. In the following sections, a couple approaches, mass screening and netnography, which make use of listed characteristics for finding lead users are discussed.

2. Existing Methodologies

2.1. Mass Screening

Mass screening, a quantitative and qualitative approach conducted using written and online questionnaires and telephone interviews, is the main method of lead user identification [2]. The search for lead users starts with trend analysis through exploration of current literature and expert interviews. Expert researchers gather information about users and the market place through semi-structured phone or in person interviews. The format is loosely structured with only a few topics and exploratory questions. The interviewer explores the interviewee’s subject knowledge about the target product and tries to gather innovation ideas [3] [4]. The sources are used to build networks to find lead users, uncover their breadth of knowledge and identify major trends. The approaches like mass screening tend to have drawbacks in low sample efficiency, high search costs, and reliance on self-assessment of respondents [2]. The method is time consuming, taking at times more than a few weeks to complete and demands labor intensive analysis of user data by groups of experts. In the following section, a more recent systematic approach to finding lead users that tackles the problem of self-assessment, called netnography is discussed.

2.2. Netnography

Netnography is a web-based systematic approach to lead user information retrieval and analysis [7]. It brings together two components, internet and ethnography and it strives to systematically examine communities of online target groups [2]. Observations are limited to monitoring of online interactions between users. The processing step involves intensive reading and parsing of retrieved online community information by experts, including user’s own contributions and comments to other users [7]. One of the main benefits of this approach is that it allows detection of user needs, desires, experiences, motivations, and attitudes [7]. As mentioned above, mass screening netnography also does not rely on self-assessment of respondents. Netnography has a high search cost and the experts also must be trained to have deep knowledge of discussed material and be able to get the feel for the language and its relation to the discussion [7]. In addition, there is no monitoring of users performing real life tasks using target products. Users with low number of posts are frequently discarded because they cannot be coded according to the existing set of user characteristics [2]. Another drawback is the need for a qualitative assessment of explicitly verbalized and implicitly existing needs [7]. To minimize the search cost and labor intensive analysis a Fast Lead User IDentifaction (FLUID) method that utilizes data mining is proposed in the following section.

3. FLUID

Data mining facilitates extraction of models and patterns from data to better systematically determine user characteristics, behaviors and needs. The process allows for codification of online users and usage profiles to interpret and conceivably predict behavior leading to tangible solutions or prototypes. In evolutionary nature of TRIZ, Fast Lead User IDentification (FLUID) brings about a systematic approach using data-mining methods to derive user characteristics and behavior for identification of innovative end-users and innovation opportunities.

The initial step of the approach is to select a suitable and rich online domain for finding lead users. Empirical evidence stipulates that lead users take part in exchanging knowledge and activities on the web [2]. Online communities, groups, social networks and lists are often a platform for enthusiastic and involved users and consumers who exchange knowledge and ideas, but also for businesses trying to adapt to user needs and wants. Twitter, a large social network, is a prominent example of a rich unorganized community, selected as a test domain for the FLUID approach. Unorganized communities are networks where discussions can cover many topics, some that are loosely connected or dissimilar, unlike organized communities that are focused on a particular product or service [8]. Organized communities like forums tend to have a smaller population of users and may not involve experts in a related business and industry. Since its creation Twitter has exponentially increased the number of users and currently has around 300 million users. Users tweet real time thoughts on their daily activities and events by posting status messages, called tweets. The community can be a valuable source of relevant data and at the same time a source of vast amounts of non-related data. The aim is to use data mining methods to separate out discussions and users that can serve as start network for identifying lead users and extracting new and current product trends.

Data mining micro-blogs like the social medium Twitter is a challenge due to its large and complex data set. Through the Twitter's open Application Programming Interface (API) a dataset of user profiles and twitter updates, tweets can be collected. The Twitter API provides two distinct services, REST API and Streaming API, the latter being an unfiltered stream of tweets. The FLUID sample test makes use of the REST API which is restricted at 150 calls per hour to keep the API traffic under control. From the twitter calls we can obtain the following data:

Twitter Data	
User	id, screen name, full name, description, creation date, language, location, favourites, lists, followers, friends, number of statuses, and is a translator, protected, geo enabled, verified and contributors enabled.
Tweet	id, text, creation date, language, location, hashtags, URL(s) and is reply or re-tweet (RT).
List	name, full name, slug, description, members, friends and is public.
Twitter	current rate limit status.

Tweet messages are short in length, 140 characters. Users make use of shorthand and nonstandard vocabulary, often missing proper names and terms. An example of relevant tweet for a bike product would be:

'Have just made a rear light bracket more robust with a little clear silicone sealant. #cycling Smart brackets tend to wobble too much' by D.J. Cook, @Downfader [9].

Tweets, as the one above, have a number of evolving conventions that markup individual words or text. User names are prefixed with a @ sign, for example @Downfader. Hash-tags denote discussion subjects or categories, i.e. #cycling. RT acronym signifies that the text of tweet is of another user and is being echoed, re-tweeted. Due to the short message length, nonstandard vocabulary Twitter data is difficult to process with existing data mining methods. At the same time, real time activity streams and topic categorizations make it a great source for knowledge discovery. Filtering and classification algorithms are applied to the posts and meta-user or tweet information to discern the essential components, presence and significance of a specific category and relevance of a particular end-user.

3.1. Data Processing

First step in data processing is to filter out protected users, verified celebrities, lists, spammers, translators, promoters, and other automated-script style Twitter accounts, see Fig 1. Such information is often irrelevant, deficient and inadequate, i.e. statuses or tweets and user extended information cannot be accessed for users that have the protection setting turned. In some cases, twitter users may be filtered also based on the language and location. Thereafter, as shown in the Fig. 1 the retrieved twitter user data is measured for likelihood of characterizing a lead user by employing five indexes: activity, relevance, trend, polarity and influence index, all derived from the aforementioned user characteristics [2]. The activity index, expressing the user involvement is for example influenced by the number of user statuses and number of lists. Influence index on the other hand, expresses involvement and leadership and is in part determined by user friendship or follower connections. Trend index is weighted by the relevance of the user's lists, hashtags and keywords.

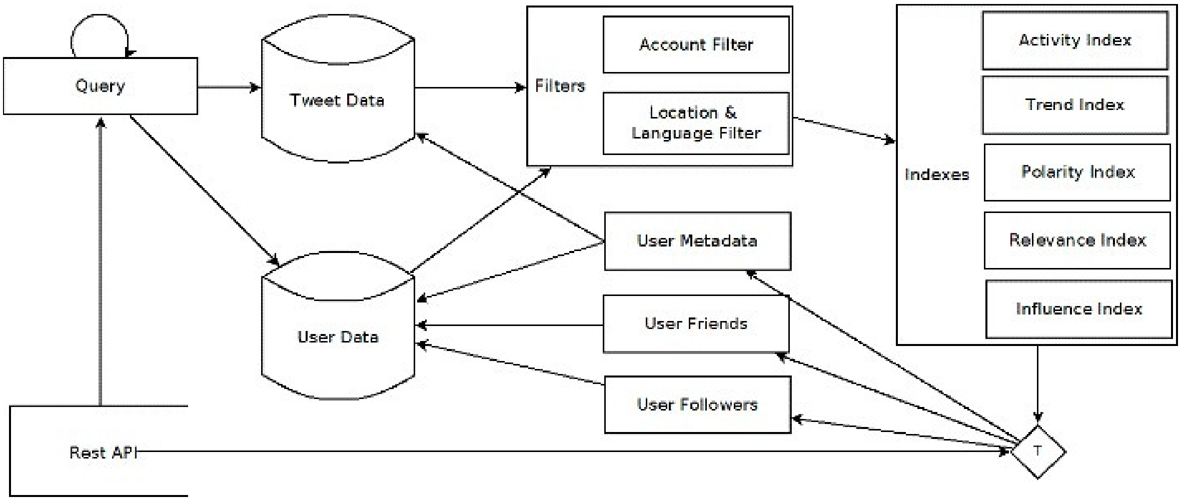

Fig. 1. FLUID model for the Twitter micro blogging community

Tweets are often babble but can also be used to express knowledge, ideas, and opinions on a subject or a product. User tweets tend to encode product related opinions, feelings and knowledge. The relevance scoring tells us how likely it is that the selected tweet or content contains a specific category. The final components of data processing are opinion mining and sentiment analysis that determine the polarity index. Tweets are classified depending if the text conveys positive or negative emotions. The dissatisfaction or positive sentiment can be estimated based upon the content of the user tweet. Emoticons are one feature that can be used to identify positive and negative sentiment in a tweet. Emoticons are composed by resembling facial expressions with symbols available on a standard keyboard [10]. Sentiment algorithms can also be utilized to automatically and accurately classify tweets into positive or negative views on a subject. Some possible machine learning algorithms for classifying tweets are Naïve Bayes, Maximum Entropy, and Support Vector Machines. A potential concern when using classification algorithms apart from low number of word occurrences is that some tweets may contain both positive and negative sentiments in a very short length of text. Other tweets may express sarcasm, irony, cynicism or taunts, all which can be difficult to classify [10]. The total index score is calculated for each user in the database and for those above predetermined threshold extended user information is obtained for further processing and building up a network of users and their tweets. Through extensive iterative data retrieval and processing clusters in the user data enable us to identify end users as lead user.

4. Validation

Validation is imperative in the final analysis of the approach. To verify the authenticity of a lead user, similar to the screening method, focus groups and interviews can be arranged [1] [4]. The selected experts may be in the field or a related field and the subject and the interviewee are researched in advance with primary information objectives decided and a few exploratory questions constructed, such as:

- What are some of the trends in this area?
- Could you give an example from your experience?

The questions allow us to find out what type of advances concerning innovation they expect [3] [4]. It is also possible to have site visits for further verification. During site visits information through observations is gathered that might be hard to collect in the user interviews. The interviews, focus group discussions, and site visits can shed light on the precision of lead user identification, new strategies or existing methods and potential improvements in the FLUID model.

5. Conclusion

The FLUID model is a systematic and cost effective approach to identifying lead users that does not rely on self or expert assessment of end users. Data mining methods can help rapidly evaluate online user data for potential knowledge and extract useful and relevant information about the end user and the target product. Developers, topics, discussions, and tasks within the community are identified and classified in addition to impressions, motivations, and insights. By having TRIZ encompass the FLUID method, there is an additional increase in previous knowledge gathered and resources through utilization of some of the primary stakeholders in innovation process, lead users. With a large and diverse user group, the selected online social network Twitter is a suitable example of a rich source of quantitative and qualitative user data. A potential concern is lack of availability of extended user information. Twitter users rarely list their location or it may lack relevant meaning. To overcome such impediments, it is possible to gather similar metadata that may fill in the gaps, for example, user location can be derived through tweet location or analysis of recent tweets. Additionally, the authors propose investigation of text stream mining of Twitter where data mining algorithms can make use of the real time updates.

References

[1] Von Hippel, Eric A. Democratizing Innovation. *Democratizing Innovation.* MIT Press, Cambridge, MA; 2005.

[2] Belz F., Baumbach W. Netnography as a Method of Lead User Identification. In: *Creativity and Innovation Management*, Vol. 19, Issue 3; 2010, p. 304–313.

[3] Von Hippel E. Lead Users: A Source of Novel Product Concepts. *Management Science*, 32, 791–805; 1986.

[4] Von Hippel E. The Source of Innovation. Oxford University Press, New York; 1988 .

[5] Lüthje C. Characteristics of Innovating Users in a Consumer Goods Field: An Empirical Study of Sport-Related Product Consumers. *Technovation*, 24; 2004, 683–95.

[6] Lüthje C. Kundenorientierung im Innovationsprozess. *Eine Untersuchung der Kunden-Hersteller Interaktion in Konsumgütermärkten*. Gabler, Wiesbaden; 2000.

[7] Bartl M., Netnography: Einblicke in die Welt der Kunden. In: *Planung & Analyse*, 2007, Vol.5, pp. 83-89.

[8] Baumbach Wenke, Schmidle Michael. Identifikation von Leadusern in Online-Portalen am Beispiel von www.seniorenportal.de, *Diskussionsbeitrag Nr. 17.* Februar 2009, ISBN 978-3-938236-67-3.

[9] Downfader. Have just made a rear light bracket more robust with a little clear silicone sealant. #cycling Smart brackets tend to wobble too much [Twitter post]. Retrieved from https://twitter.com/Downfader/status/216939320894947329. June 24, 2012.

[10] Bifet, A., & Frank, E. . Sentiment knowledge discovery in Twitter streaming data. In: *Proceedings of the 13th International Conference on Discovery Science*. (pp. 1–15). Berlin, Germany: Springer; 2010.