Idea Mining – Text Mining Supported Knowledge Management for Innovation Purposes

Completed Research Paper

Paul Kruse
Technische Universität Dresden
paul.kruse@tu-dresden.de

Andreas Schieber
Technische Universität Dresden
andreas.schieber@tu-dresden.de

Eric Schoop
Technische Universität Dresden
eric.schoop@tu-dresden.de

Andreas Hilbert
Technische Universität Dresden
andreas.hilbert@tu-dresden.de

ABSTRACT
Following the emergence of Social Media and the increasing willingness of customers to share thoughts, ideas, and experiences companies are trying to capitalize on such activities. Due to the vast amount of user-generated content, manual analysis and interpretation will not meet the demands of companies in highly competitive environments. Based on an integrative process model, which describes the process of idea generation, we outline a BPMN-based path that allows companies to steer user participation and the application of Text Mining methods to gain valuable ideas for innovative products. Our approach also illustrates the Knowledge Management perspective supporting the customers during idea generation. In order to demonstrate the applicability of our model we finally depict the whole process utilizing Dell’s IdeaStorm.

Keywords
Innovation, Knowledge Management, Text Mining, BPMN.

INTRODUCTION
Innovation is commonly defined as “the outcome of an interactive process between the firm and its environment, as the result of the collaboration between […] actors, located both inside and outside the firm” (Mention, 2011, p. 44). Spanning from idea generation (ideation) to their commercialization (Xu et al. 2010, p. 581) innovation requires social interaction from which knowledge is created, distributed, and adopted. Traditionally driven by internal researchers, innovation nowadays focuses more on customers’ ideas driven by an open innovation approach (Chesbrough, 2003). For companies customers’ communication and knowledge exchange (user-generated content, UGC) – discussing trends, product developments, and individual needs – are highly valuable. Therefore, companies should listen to their customers and integrate them into their innovation process. Beside the discussion about products companies are strongly interested in customer ideas. Ideas can be regarded as images formed in the mind written down as textual information. These images are often the base for technological breakthrough (Thorleuchter et al. 2010, p. 7182), but are mostly hidden in large amounts of data.

Following the emergence of Web 2.0 technologies, the WWW provides many opportunities to share ideas. Such diversity confronts companies with some disadvantages: As potential customers can use several technologies the amount of data from which valuable knowledge (i.e. ideas) can be extracted is vast. And beside potential ideas there is a lot of “noisy” data from these sources.

Fostering the process of ideation (Graham & Bachman, 2004) we suggest that companies should apply methods of Text Mining (TM) on the collected content. TM focuses on large amounts of textual data and its transformation into valuable knowledge. We therefore use TM methods on social media analyzing the provided content. To supply TM with data and to support customers during ideation the methods of Knowledge Management (KM) provide the prerequisites in our approach.

We develop a BPMN-based process model integrating state-of-the-art-methods of KM and TM for efficiently discovering knowledge from web sources to support the innovation process. First, the process model is aimed at motivating people to share their ideas to fuel new product development resulting in a huge text corpus. Second, our model supports the selection of eligible TM methods for an automated extraction of knowledge in the collected data. Hence, the present study helps companies to foster a steady generation of innovative ideas and thereby to sustain competitiveness.
RELATED WORK

The paper is located in three different research areas: Innovation Management, KM and TM. The importance of innovation has a very long history (Rowley et al. 2011). Although there are many efforts in innovation research, a general definition and detailed description of the innovation process itself is still missing. Utterback (1974) describes a simple process to which Desouza et al. (2009), Miles et al. (2000), and Xu et al. (2010) add aspects of KM considering the important role of knowledge in innovation.

Our second research area covers the field of KM: Gibbert et al. (2002) point out that KM enables companies to provide and maintain the requirements and resources for customers to participate in innovation and allows them to contribute ideas and feedback, discuss trends, etc. This includes technology, motivational aspects as well as the involvement of the firm’s employees guaranteeing that enough UGC is available for analysis.

The analysis of that UGC content leads to TM (Felden et al. 2006; Hippner & Rentzmann, 2006; Weiss et al., 2010). TM methods are able to automatically analyze textual content and, e.g., to cluster ideas of similar topics. Many researchers demonstrated the applicability of these methods in several fields. Related to our work are applications in product development and KM (Ur-Rahman and Harding (2012)), patent technology mining (Feng & Fuhai (2012)), and even the extraction of textual information from blogs (Thorleuchter et al. (2010)). Nevertheless, their work does not provide a model to an integrated approach.

These findings illustrate that many scientists are working in this research area. Hence, a complete and integrated description of the whole process from ideation over knowledge and its management to concrete TM methods is still missing.

BACKGROUND

To fill this research gap we identified we propose a process model, which covers the peculiarities of an integrated approach between Innovation Management and KM-supported TM. Figure 1 illustrates the different perspectives and highlights the relations between each sub-step of the model.

Starting from a general process of innovation our approach strongly focuses on the ideation or idea generation phase. This phase integrates two main paths of user participation and results in product ideas and the discussion about them, which we, first, want to foster by applying supportive KM methods and which subsequently shall be analyzed through TM methods. A more detailed description of the single steps and the relation between the identified sub-processes can be found in the following chapters.

INNOVATION MANAGEMENT

It is commonly known that organizations need to innovate responding to changing customer demands as well as capitalizing on opportunities offered, e.g. changes in markets (Rowley et al., 2011). However, the process of innovation lacks a general definition. Many authors highlight several perspectives, which relate to innovation as a process, as an item (e.g., product, service, or program) or innovation as an attribute of organizations. Although we do not waive the possibility that our process model can be applied on services or processes, the present research is limited to product innovation.

Process of Innovation

Merging state-of-the-art-methods of KM and TM for efficiently discovering knowledge from UGC, we first focus on the process of innovation. It can be defined as “the generation, acceptance and implementation of new ideas, processes products or services” (Thompson, 1965). In a more specific manner, Du Plessis (2007, p. 21) describes innovation “as the creation of new knowledge and ideas to facilitate new business outcomes, aimed at improving internal business processes and structures and to create market driven products and services”.

Following Utterback (1974, p. 621) the process of innovation can be divided into three stages: “generation of an idea, problem-solving or development, and implementation and diffusion.” Generation involves a synthesis of diverse information, e.g., about a market or needs and technologies to meet the needs. Problem solving is concerned with “setting specific technical goals and designing alternative solutions to meet them” and leads to an original solution or invention. After that implementation, i.e. “manufacturing-engineering, tooling, and plant and market start-up required to bring an original solution or invention to its first use or market introduction” is followed by diffusion “after the innovation is introduced” (Utterback, 1974, p. 621).
Innovation is extremely dependent on the availability of internal and external knowledge (Du Plessis, 2007). Current definitions of the process of innovation increasingly focus on the knowledge perspective, e.g., on “knowledge creation” (Miles et al., 2000, p. 304) or “knowledge commercialization” (Desouza et al., 2009, p. 23). Hence, the present study draws upon an innovation process which was developed to support the increasing importance of the knowledge perspective (Xu et al., 2010, p. 581).

Figure 1. Overall process model

Figure 2. Process of innovation

Idea generation and research/development can be compared to what Utterback (1974) described with “generation of an idea, problem-solving or development” (p. 621). The subsequent stages separate the single stage of implementation and diffusion and thereby differentiate between early development of, e.g., a new product and its final commercialization.
**APPROACH**

The present paper primarily focuses on the idea generation step of the innovation process (Figure 2). We believe that this stage in particular can be supported by KM and TM methods. In order to apply these methods we divide the ideation process.

![Figure 3. Process of idea generation](image)

From a general customer-centric perspective our approach begins with user participation on suitable social media channels (Figure 3). After that data must be extracted from the channel(s), followed by the extraction of valuable data with TM methods. Subsequently, the aggregated and structured data can be handed over to R&D where the ideas are checked for suitability, reliability, etc. and the process leaves our observation focus.

**Challenges of User Participation in Idea Generation**

Chau & Tam (2000, p. 230) illustrate “two motivations and driving forces” behind ideation: It can be driven by technology push or by market/need pull. While the former suggests that innovation is driven by science, and thus drives technology and diffusion, the latter indicates that ideas are derived from user needs as key drivers of adoption: Figure 4 illustrates this differentiation by representing two possible ways.

On the one hand, we observed that some companies set up own ideation platforms and provide one (or more) social media channels for submitting ideas. On the other hand, companies also extract data from already established channels (e.g., Twitter).

In the first case, companies can exercise more influence on the customer. They not only control the technology and the problem (product) that needs to be innovated, they can also support ideation by offering incentives directly to the customers. The other case gives companies less power. They can only narrow down the unit of analysis in order to limit the amount of data. Nevertheless, they have to browse through vast amounts of blog posts, tweets, etc.

Like Langrish et al. (1972) other researchers have concluded, ideas from a market pull show a higher probability to gain commercial success than technology-push innovation. Plus, recent research on innovation leads to a paradigm shift towards the concept of Open Innovation (Chesbrough, 2003) focusing on tapping the knowledge of the customer. Therefore, our process model allows companies to benefit from customers’ ideas.

![Figure 4. Process of user participation](image)

After covering innovation and its process we now examine suitable KM methods. As mentioned above, we investigate methods that facilitate the exchange of customer knowledge. Due to the fact that customers know best what they need, they can provide most valuable ideas for innovative products. If companies collaborate with such bearers of external knowledge (Kang & Kang, 2009) they will be able to cope with shortened innovation cycles, rising R&D costs, etc. (Gassmann & Enkel, 2004).
Hence, the key questions remain, how companies can tap such knowledge/ideas, where do good ideas come from and what can companies do to push external ideation?

**Knowledge Management in a Web 2.0 Environment**

With the emergence of Web 2.0 customers participate in communities, networks, and other social media activities. After purchasing a product customers review the product providing recommendations for others, name positive and negative characteristics and discuss possible improvements or entirely new ideas.

When companies are willing to harvest such valuable input, they must overcome several obstacles:

- Customers tend to criticize and discuss a product only if it failed to match their expectations. Providing valuable ideas and suggestions to improve a product depends on the individual motivation of the customer.

- Many customers do not address companies directly. Unless the company has its own feedback system through which customers can directly provide their feedback, innovators have to retrieve such data all by themselves.

- Web 2.0 offers a wide range of communication channels that can hardly all be monitored or controlled. Some of these may even be private and not accessible for non-invited users. Therefore, companies must focus on the most important channels in order to gather the best input from them.

Depending on the particular situation, we identified two general approaches and KM supported paths that can push the extraction of textual data (Figure 4): From a KM perspective, a company can, on the one hand, establish an own platform or channel to gain full control over product discussions. On the other hand, the focus can remain on selected public channels, such as blogs, micro-blogs or social networks. Thus, bias caused by the presence of the company during discussion or the need to invest and administrate an own platform can be avoided. In both cases companies rely on methods and techniques, which provide them with the required knowledge. They always must collect data from web sources in form of unstructured textual data.

KM offers a wide range of practices to identify, extract, create, distribute, and adopt external customer knowledge fostering innovation. For example: Companies, such as SAP (SAPiens), Lufthansa (Air Cargo Innovation Challenge), or Dell (Ideastorm) are increasingly drawing in external ideas from customers by providing a central platform. Other companies, such as Subway’s (Subway Fresh Buzz) or McDonald’s (McCafe Your Day) limit their activities to certain social media channels (e.g., own Twitter channels) and thereby allow and encourage customers to provide innovative ideas. Thus, the latter do not limit the customers’ creativity to a specific task (innovation/idea contests) or a certain product (co-design, co-creation, etc.) but are less visible than companies with own platforms. In addition, they cannot offer incentives directly or benefit from community effects. Following a third approach, companies can also refuse to set up an own channel and focus just on existing channels (Kruse, 2012). Thereby, they attract even less direct feedback and lose any control over the communication of customers, but can access a much bigger data pool. Recent studies illustrate that, e.g., the amount of tweets has increased by nearly 700% over the last two years (Blog.twitter.com, 2012). Hence, companies can access large amounts of data even without establishing own channels.

**Knowledge Management supported Idea Generation**

From a KM perspective companies should engage in certain activities supporting the ideation. Cooper & Edgett (2009, p. 94ff.) identify 18 different sources of new product ideas. We consider most of them suitable to feed the phase of ideation with valuable data. Nevertheless, some methods such as patent mapping, open innovation with vendors, and ethnography depend on physical contact with the source, are rather inward-looking or do not involve any Social Media channel. Due to our research limitation we rather focus on those sources and methods that comprise customer knowledge, such as customer brainstorming, communities of enthusiasts, external idea contests, etc. Finished designs from customers or open innovation projects with partners and vendors may also be interesting for our research, but only if they cover at least one of the above-mentioned idea generation and KM paths.
In general, Web 2.0 with its new application classes not only allows companies to tap more channels through KM, it also facilitates the collection of customer knowledge (i.e. ideas). Effective KM may also lead to a better understanding of demands, better product ideas, more innovative products, a shorter time to market and lower product costs. Hence, in order to overcome the above-mentioned obstacles and to provide data for TM a company’s KM must ensure the following points:

Hence, our KM perspective supports customers to provide valuable knowledge, motivates them to participate and ensures data extraction for subsequent TM (Figure 5).

**Analyzing Unstructured Data with Text Mining**

After data extraction we collected a huge textual corpus. In order to reduce labor costs which would incur by reading and classifying the collected texts, it is necessary to analyze this data using appropriate algorithms.

**Description and Process of TM**

Related to the methods of data mining discovering patterns in structured data TM methods reveal information in unstructured textual data (Weiss et al., 2010, p. 1). TM describes the partially automated discovery of new and valuable knowledge from text documents (Feldman & Sanger, 2006, p. 1; Hippner & Rentzmann, 2006, p. 287). Because of the mentioned relationship to data mining the process also shows several pre-processing tasks preparing the data (Figure 6).

**Task Definition and Document Selection**

The first step in any TM project is to define the objectives. In our case the task is to understand customer ideas. Therefore, we group the collected data in homogeneous segments containing similar content. Afterwards, we select the relevant documents, in our case represented by UGC extracted from Web-2.0-channels (Figure 4).

**Pre-processing**

Before we can apply TM methods and identify idea clusters, several pre-processing tasks have to be performed. This step is very important and comprises necessary tasks structuring the otherwise unstructured text data. Therefore, terms (or tokens,
i.e. a single word or a group of words) representing the documents are extracted and set into relation with each document (Weiss et al., 2010, p. 16). For term extraction are used methods of the research area of natural language processing which are separated in three groups: morphological analysis, syntactic analysis, and semantic analysis (Hippner & Rentzmann, 2006, p. 288).

![Figure 6. TM process](image)

The aim of the morphological analysis is to reduce the complexity for analysis methods (Weiss et al., 2010). Complexity in text analysis correlates with the word count: irrelevant words or terms have to be removed. Therefore, we convert terms into a unified expression. This procedure is called stemming or lemmatization (Hippner & Rentzmann, 2006, p. 288; Weiss et al., 2010, p. 18). For example, the words “complexity” and “complexities” are different terms but are forms of the same word. With stemming such terms are identified and normalized. Another possibility for complexity reduction is to remove stop words (Heyer et al., 2006, p. 80). Stop words are words or terms, which appear very often (e.g., articles or pronouns) and have no special meaning within the text.

The aim of the subsequent syntactic analysis is the annotation of the terms with part of speech (POS) tags structuring the raw text data and extracting information selectively, e.g., concentrating on proper nouns or adjectives (Heyer et al., 2006, p. 112; Ur-Rahman & Harding, 2012, p. 238; Weiss et al., 2010, p. 31). For POS-tagging a dictionary showing word-POS correspondence can be useful (Hippner & Rentzmann, 2006, p. 288). Afterwards the terms are analyzed regarding their function in a sentence, e.g., subject, predicate, object, allowing us to select information from specific syntactic units.

The aim of the final semantic analysis is the detection of the context the document deals with (Hippner & Rentzmann, 2006, p. 289). Since specific words have different meanings, this procedure tries to discover the right intent. This task can also be supported by a dictionary or a product database containing terms in the relevant context (Schieber et al., 2012).

After pre-processing the raw text data obtained a kind of a structure: sentences are separated in relevant terms, POS- and sentence-functions are determined, and the context within the idea texts is revealed. Thus, we prepared the data for applying the TM methods.

**Text Mining Methods and Evaluation**

As mentioned before, our task is to divide the corpus in groups with similar content (i.e. ideas). Since our aim is not to provide new methods for TM, we use two established methods: first, we classify the documents into existing groups, and second, we segment the documents regarding their content.

Methods of text (or document) classification are related to traditional data mining methods coping with classification tasks. These methods require an existing catalogue of possible classes by which the documents can be merged (Felden et al., 2006, p. 2; Weiss et al., 2010, p. 6). In context of innovation processes we should elaborate this catalogue with regard to knowledge and innovation management. Suitable approaches are, e.g., decision trees, Naive Bayes, or Support Vector Machines (Felden et al., 2006).

In contrast to text classification as described above, we can segment the documents using text cluster algorithms (Heyer et al., 2006, p. 195; Weiss et al., 2010, p. 91). Therefore, we do not require a predefined catalogue: the method finds the clusters by itself scoring the document similarity. The similarity is evaluated by comparing terms: documents containing similar terms are merged in a cluster. An advantage of this procedure is that we are able to detect classes, which we did not bear in mind previously. As traditional clustering methods we suggest, e.g., k-Means (Weiss et al., 2010, p. 96) or co-occurrence-based approaches like topic models (Blei & Lafferty, 2009; Sommer et al. 2011).
After performing the TM methods we can evaluate the document clusters by browsing in a specific segment for further analysis. In particular, when segmenting the documents without a predefined catalogue, we can gain important insights by evaluating the keywords of the found segments. So, we get a feeling about the problems or ideas our customers have. This clean set of ideas is handed over to R&D, the subsequent step in the innovation process (Figure 2) where the ideas are checked for suitability, etc.

**APPLYING OUR MODEL ON DELL’S IDEASTORM**

To illustrate the applicability of the proposed process we apply the different steps on Dell’s ideation platform IdeaStorm. Since its introduction in 2007 customers submitted 18,500+ ideas, voted 740,000+ times, and thereby contributed to the implementation of 520+ ideas (IdeaStorm.com, 2013). On IdeaStorm, users write articles containing their ideas, vote for them and add comments. With this platform Dell’s main interests are ideas to new products/services. Therefore, we use this platform as an example which can be allocated to our process of innovation (Figure 2) and which illustrates the applicability of our overall process model very well.

The first step of the process of innovation is the idea generation dealing with user participation and KM (Figure 3). Referring to a specific idea\(^1\) and the discussion about it, we explain the applicability of our process model. The idea – dealing with color variations for PCs – was posted on August 1\(^{st}\), 2012. Following the upper path of the process of user participation (Figure 4), the user published his idea (‘submit problem’) on Dell’s platform (‘use resources’), Dell motivates contributors by highlighting their reputation (‘receive incentives’). The idea received 25 votes, one extension (fostering the evolution of the idea through user collaboration) and 29 comments (‘discuss’) by other users and a Dell Partner representing the Dell’s KM (Figure 5). On IdeaStorm (‘select channel’) the Partner looks for new ideas (‘select problem’) and discusses them with the community providing feedback. The last step of the company’s KM is the extraction of discussion data. Regarding this example, we extract the description of the idea itself, the extension, and the comments. In turn, the comments to this idea can be separated in those containing commendations, useful hints, opinions, or off-topic statements.

In this simple case we can easily group and aggregate the information regarding this idea. As this platform contains 18,500+ ideas, some of them can be similar or refer to similar concerns. Besides, there are 97,000+ comments, which have to be analyzed to get an idea about the users’ commitment. Dell also interacts with customers through several other Social Media channels, which have to be considered in an integrated innovation process. Facing this situation, we support the last step of idea generation with TM methods analyzing the discussion data automatically. Following Figure 6 after defining our task – i.e. group ideas/comments – we have to pre-process the textual data (e.g. separating terms, identifying adjectives/nouns/synonyms/misspellings) before we segment the extracted documents obtaining a list of similar ideas. Afterwards, we can evaluate the results and start another analysis looking into the comments of the corresponding ideas. After the successful detection of a valuable idea the subsequent step is to submit this idea to the step of R&D in the process of innovation. For the mentioned idea the Dell Partner set its status to ‘under review’ showing the community Dell’s interest on their concern.

Overall, the example above illustrates how much data can incur and how important useful automatic procedures can be in order to gain advantage from listening to customers. Nevertheless, our proposed model still requires further evaluation.

**CONCLUSION**

With the emergence of Social Media the amount of UGC that covers valuable ideas is too high to be handled. Therefore, if companies aim to capitalize on customer knowledge by identifying such ideas manual analysis and interpretation will not meet their requirements in today’s competitive environment.

In order to facilitate the analysis of UGC and ideas we suggested an integrated process model. As a prerequisite of analysis we illustrated how KM and selected methods allow companies to source data from Social Media and highlighted the importance and implementation of customer support during ideation. After that we showed the potentials of TM methods to identify structures in the extracted data and how to embed them into the process of innovation.

Following our process, we believe that companies can start the subsequent steps of innovation on a more sophisticated level, as they gathered valuable ideas from a range of sources they were not able to handle or even access before. In addition, we

\(^1\) Color variations for PCs, http://dell.to/VlyGa4
illustrated the dependencies between each step and developed a methodology that covers the process of ideation from user participation through supporting KM methods to TM and its results.

The depicted process model makes no claims of being exhaustive. It should be regarded as a road map, which covers the main paths but is open to side roads. Hence, we believe that other TM (e.g., Opinion Mining for prioritization of ideas or Document Warehousing for long-term analyses) or KM methods (e.g., Social Media supported brainstorming or focus groups) can be included into the process. This would also require a proof of concept.

REFERENCES


