Extracting Consumers Needs for New Products

A Web Mining Approach

Dirk Thorleuchter
Fraunhofer INT
Euskirchen, Germany
dirk.thorleuchter@int.fraunhofer.de

Dirk Van den Poel, Anita Prinzie
Faculty of Economics and Business Administration
Ghent University, Department of Marketing
Gent, Belgium
dirk.vandenpoel@ugent.be; anita.prinzie@ugent.be

Here we introduce a web mining approach for automatically identifying new product ideas extracted from web logs. A web log - also known as blog - is a web site that provides commentary, news, and further information on a subject written by individual persons. We can find a large amount of web logs for nearly each topic where consumers present their needs for new products. These new product ideas probably are valuable for producers as well as for researchers and developers. This is because they can lead to a new product development process. Finding these new product ideas is a well-known task in marketing. Therefore, with this automatic approach we support marketing activities by extracting new and useful product ideas from textual information in internet logs. This approach is implemented by a web-based application named Product Idea Web Log Miner where users from the marketing department provide descriptions of existing products. As a result, new product ideas are extracted from the web logs and presented to the users.

Web Mining; Text Classification; New Product Development, Knowledge Discovery

I. INTRODUCTION

Web logs are web pages that consist of textual and non-textual information combined with links to further web logs, web pages etc. Individual persons normally write them [1]. Many web logs offer the possibility for readers to leave comments in an interactive format. We can see an increasing amount of these web logs for nearly each topic [2]. Most web logs consist of textual information. To analyze textual information tools and methods from text mining can be used. Web logs have been analyzed for marketing aspects by these text-mining tools in three different ways. Firstly, we can find out brand reputation, secondly we can measure effects of advertisements, and thirdly we can focus on consumers needs for new products [3].

In this paper, we focus on analyzing consumers' needs for new products. Many web logs deal with existing products. There, enterprises publish descriptions of existing products in the internet and offer the possibility for readers to leave comments. In general, consumers use this possibility to describe their experiences and problems with the product. However, sometimes, suggestions for new product features or even for completely new products are published.

These new product ideas are of particular interest for this approach. This is because they probably can lead to a change in the product development process. We identify product ideas that have two properties. Firstly, they have to be new. Then, an existing product does not yet contain these new ideas and new ideas are not part of product development activities. Secondly, product ideas have to be useful. They should have a relation to existing products or actual product development activities otherwise it is not possible to integrate them in the product development process. An example for this is that a manufacturer who produces coffee machines can start a new product development process for espresso machines. This is because coffee machines are related to espresso machines. However, he normally is not able to produce completely different products in the (near) future (computer, furniture etc.).

Identifying new and useful product ideas is a well-known activity in marketing [4,5]. Therefore, with an automatically process we support these marketing activities by extracting these new product ideas from web logs.

One important aspect in this approach is that consumers write new product ideas in web logs by use of a colloquial language [6]. In contrast to the technical language, we see that terms are not defined exactly and that many homonym and synonym problems occur by evaluating this textual information with text mining methods [7]. However sometimes, consumers also use technical terms if their need for new products refers to a technological product.

II. RATIONALE BEHIND THIS APPROACH

Finding these new product ideas is a well-known task in marketing. Below, we describe how marketing professionals identify these new product ideas in the internet.

A person - usually from marketing department - analyzes the situation of an enterprise so that (s)he recognizes all existing products and all product developments. We assume, that these products and product developments are already described in form of textual information. Then, the person searches the internet for new product ideas that are published by consumers. For this, (s)he specially searches web logs because there, consumers present their needs for new products. If a web log system is installed on the enterprise website then the person checks these web log comments for new product ideas. After this, he searches external web log sites where consumers provide comments concerning these products or concerning similar products e.g. competing products on the enterprise website of competitive firms.

978-0-7695-3923-2/10 $26.00 © 2010 IEEE
DOI 10.1109/WKDD.2010.14

Searching in web logs can be done by using an internet search engine and by limiting the query results to textual information from web logs. Therefore, a search query consists of several domain-specific terms. For this, the person uses terms from the description of the products and the product developments. These search queries are executed by an internet search engine.

Each retrieved document from a query result consists of a title, a short description, and an internet link. The short description contains search terms from the query in bold print and the internet link leads directly to the full text of the retrieved web log site. The person checks the title and the short description for new and useful product ideas. This normally is done in an intuitive way. If query results have promise to the person then (s)he focuses on the full text by using the query result link. There (s)he probably extracts the new product idea.

III. METHODOLOGY

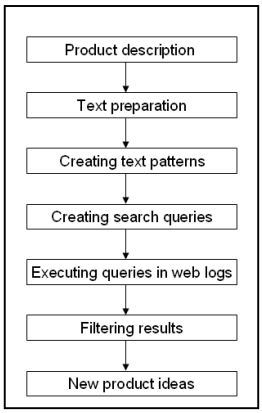

Figure 1. Processing of the web log mining approach in different steps.

We use the methodology in Fig. 1 to realize the rationale as described in Sect. 2. Therefore, this web log mining approach has the aim to support users by finding new product ideas from web logs. Below, we describe how to prepare textual information from a provided product description (see Sect. 4), how to create text patterns, and how to build search queries based on the created text patterns (see Sect. 5). Additionally, we describe how to execute search queries by use of a web log search engine and how to filter the query results to get new product ideas that are presented to the user (see Sect. 6).

IV. TEXT ACQUISITION AND PREPARATION

For this web log mining approach, a user has to provide a product description. Normally descriptions of existing products are available because persons from marketing department of an enterprise use them for marketing/public relations purposes. However sometimes, descriptions of future products that are in a product development process are not existent. Then, to extract new product ideas concerning these future products, the user first has to create these descriptions.

In a pre-processing phase, the provided product description is tokenized [8] by using the term unit as word. Additionally, we use a standard word stop list [9]. Normally in text mining applications, stop words are deleted and all other terms are used for further processing. However, in this approach it is important to know the position of stop words in the texts for creating text patterns.

V. TEXT PATTERN AND SEARCH QUERY CREATING

Around each term in the provided product description, we build a text pattern if the selected term is not a stop word. We compute the length of text patterns by a provided length from the user and by a user-given term weighting schema. The schema distinguishes between stop words and non-stop words because they are not equally important.

Then, we build search queries from the created text patterns. As described in the rationale, a person uses several terms from the product description to build a search query. Heuristically, we estimate the number of terms in a search query at four. This is, because if the number of terms in a search query is too large then the query is too specific and we probably do not find the relevant web log comments. If the number of terms in a search query is too small, then we probably get homonym and synonym problems, because consumers normally use colloquial language. Additionally, we get too much query results. This causes performance problems. Therefore, we think that four terms is a good compromise.

We use stemmed terms [10] to build the queries. This is because searching in a web search engine with a stemmed term leads to query results that contain several different terms, which all have the same stem. Further, we do not use stop words in the search query because normally search engines delete these terms. Therefore, we delete the stop words and all further terms are stemmed using the well-known Porter stemmer [11]. Then, we build search queries that consist of four different stemmed terms from one text pattern.

In Fig. 2 we show an example for this processing. Here we start the processing with a user-given product description. We build text patterns around each term that is not a stop word. We identify these sixteen terms in the following order of appearance: wireless, LAN, coffee, machine, future, wireless, LAN, control, coffee, machine, wireless, LAN, coffee, machine, compatible, and standard. In this example, we set the length of a text pattern to seven. Then, a text pattern consists of the selected term and seven terms from its left context as well as seven terms from its right context.

Therefore, in the first step, we create these sixteen text patterns. Text patterns that are built around the first appearance of the terms wireless, LAN, and coffee as well as text patterns that are built around the last appearance of the terms coffee, machine, compatible, and standard are contained in further text patterns. This is because text patterns at the beginning or at the end of a text are smaller than these further text patterns. They do not contain

additional information. Therefore, these text patterns are discarded.

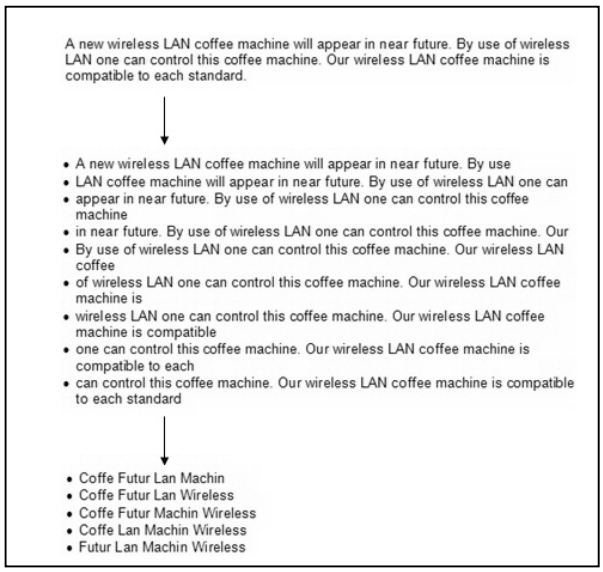

Figure 2. This example shows how text patterns are extracted from a user-provided product description.

Then we build search queries that consist of four stemmed and stop-word filtered terms, which occur together in a text pattern. To reduce the number of created search queries, we only build a search query with terms that occur frequently in the product description. For this, we compute the z % most frequently stemmed and stop-word filtered terms from the product description. In this example, we can identify eight stemmed and stop-word filtered terms. The terms wireless, LAN, coffee, and machine occur three times in the text. The terms future, control, compatible, and standard occur once. We set the parameter z to 70 % that means, we identify the five most frequent terms. For this, we select the four terms: wireless, LAN, coffee, and machine. Additionally, we also select further non-frequent terms in order of appearance. Then, the five most frequent terms are wireless, LAN, coffee, machine, and future.

All these five terms appear together in one text pattern e.g. in the first text pattern in Fig. 2. Therefore, five search queries are built that consist of each combination of four terms.

The parameter z can be selected by the user to reduce or increase the number of created search queries. This is because large texts lead to a large number of search queries, which causes performance problems. Additionally, small texts lead to a small number of queries. Then probably, new product ideas cannot be identified.

VI. Search Query Executing and Result Filtering

To execute the created search queries we use web services. A web service is a software system that is designed to support interoperable machine-to-machine interaction over a network. Frequently web services are just web based advanced programming interfaces. Access to these interfaces is possible over the internet. Then the requested service is executed and resulting data is transferred back to an application that requested the service [12]. A lot of internet search engines offer web services. By use of these web services search queries can be limited on information from web logs [13]. Therefore, we use them in our web-based application to execute queries automatically and to get the query results.

The query results consist of a title, a short description that contains terms from the search query in bold print, and a hyperlink that leads to the full text (see Fig. 3). In this approach, we identify new product ideas from the short description text pattern. However, a short description from a query result probably consists of several text patterns that are separated by several dots. In this case, we discard this query result.

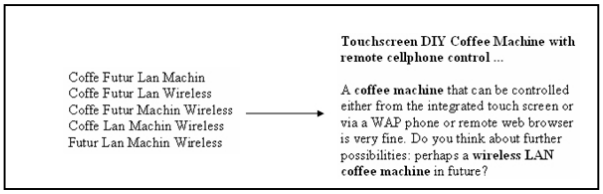

Figure 3. In this example, we identify a consumer's need for a new product: This idea deals about integrating touch screens, WAP phone interfaces, or remote web browser interfaces in the future product (the wireless LAN coffee machine) of the enterprise.

As described in Sect. 1 the identified product ideas should have two properties: novelty and usefulness [14]. Novelty means that the short description consists of information, which must not be in the product description. This means, several stop-word filtered terms from the short description should not occur in the product description. For this, we compare the stemmed and stop-word filtered terms from short descriptions to all extracted text patterns from the product description. This comparison has the aim to identify several terms from a short description that does not occur together in a text pattern from the product description. We compare the short description to a text pattern and not to the complete product description because the product description probably consists of textual information about several existing or future products from an enterprise. Then, product description consists of a large size and probably many domain-specific terms are contained in this description. In this case, a new product idea that combines several features from different products will not be assigned as a new idea.

Usefulness means that the short description consists of information, which must be in the product description because it should have a relation to the problem. This means, several stop-word filtered terms from the short description should occur in the product description. We already have identified terms from the search query as characteristic (most frequent) terms from the product descriptions. Therefore, we check for the appearance of all four terms from the corresponding search query in the short description.

Then we present the selected short description to the user as new and useful product ideas.

VII. EVALUATION

One important aspect in this approach is that consumers write new product ideas in web logs by use of colloquial language. In contrast to the technical language, we see that terms are not defined exactly and that many homonym and synonym problems occur by evaluating this textual information with text mining methods.

We compare this web log mining approach to a baseline model because we are not aware of other approaches for identifying new product ideas from web logs at the present time. For this, we do not use the chance baseline, which assigns a classification randomly because a high percentage of extracted query results do not represent a new product idea. Therefore, we use the frequency baseline. Here we have two classes (A means a query result represents a new product idea, B means a query result does not represent a product idea) in our data, and we classify each instance (query results) with a specific percentage as either A or B.

Additionally, consumers use colloquial language by providing comments in web logs. Therefore, we also compare this approach to an approach that identifies new ideas from the technical language.

For evaluation, we use product descriptions from several randomly selected products that are published in web logs. We create a web log mining application that realizes this approach. It is available at http://www.text-mining.info. There, the web-based application and all texts that are used for evaluation are presented. The application automatically extracts new product ideas from web logs and presents them to the user.

To evaluate the results of our approach, we use precision and recall measures commonly used in information retrieval based on true positives, false positives and false negatives. For this, we have to define the ground truth for our evaluation. Therefore, a human expert uses descriptions of 40 products. For each product, he manually identifies problem solution ideas from the internet. Additionally, he checks the results of this approach to find further new and useful ideas by using the web log mining application.

To compute the percentage for the frequency baseline, we use the average percentage of all 40 products, which is computed by the number of new product ideas - as computed by the human expert - divided by the number of query results.

We use the web mining approach to identify the number of queries from each patent. For each query, we focus on the first ten query results. Therefore, we multiply the number of queries by ten to get the number of query results for each patent. Then, for each patent, we divide the number of new and useful ideas as computed above by the number of query results. After this, we get a percentage x. It says that x % of all query results represent a new and useful problem solution idea. Then we compute the average percentage for all 40 patents. As a result, 3 % of all query results represents a new product idea. Therefore, we set the frequency baseline to 3 %.

For each product, we compute the values of true positives, false positives and false negatives using the web mining application. Then, we compute the precision and recall values. After this, we compute the average precision and recall values for all products.

As a result, we get a precision value of 30 % and a recall value of 50 %. A precision value of 30 % means that if this approach predicts ten new and useful ideas then three of them are new and useful ideas. A recall value of 50 % means that if there are 10 new and useful ideas in the internet then this approach identifies five of them.

To see whether these results are good or bad we compare them to a further approach that has the aim to identify new ideas from the technical language [14]. Here, we get a precision value of 40 % at a recall value of 50 %. This is because in the technical language, terms are defined more exactly and homonym and synonym problems do not occur so often.

Additionally, we compare the results to the frequency baseline. Here, we get a precision value of 3 % at a recall value of 50 %. Therefore, we think that this web mining approach can be used to support persons from marketing department by finding new product ideas from web logs.

[1] C. Zsunyi, "Weblogs/ Blogs: Stand der Technik und Zukunftspotentiale," GRIN Verlag, pp. 4-6, 2007.

[2] S.C. Herring, L.A. Scheidt, S. Bonus, E. Wright, "Bridging the gap: a genre analysis of Weblogs," Proc. 37th Annual Hawaii International Conference on System Sciences, Hawaii, 2004.

[3] J.H. Soll, S. Strauch, "Ideengenerierung mit Konsumenten im Internet," Springer, Berlin, Heidelberg, 2006.

[4] L. Lawton, A. Parasuraman, "The impact of the marketing concept on new product planning," Journal of Marketing, vol. 44:19, 1980

[5] A. Kuss "Marketing-einfuehrung: Grundlagen- Ueberblick-Beispiele," Springer, p. 189, 2006.

[6] L. Hoffmann, H. Kalverkaemper, H.E. Wiegand, "Languages for Special purposes," Walter de Gruyter, p. 1602, 1998.

[7] M.J. Martin-Bautista, D. Sanches, J.M. Serrano, M.A. Vila "Text Mining using Fuzzy Association Rules," V. Loia, M. Nikravesh, L.A. Zadeh, "Fuzzy Logic and the Internet," Springer, Berlin, p. 173, 2004.

[8] R. Feldman, J. Sanger, "The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data," Cambridge University Press, p. 318, 2007.

[9] G. Lustig, "Automatische Indexierung zwischen Forschung und Anwendung," Georg Olms Verlag, Hildesheim, p. 92, 1986

[10] A. Hotho, A. Nuernberger, G. Paass, "A Brief Survey of Text Mining," LDV Forum, vol. 20(1), pp. 19-26, 2005

[11] M.F. Porter, "An algorithm for suffix stripping," Program, vol. 14(3), pp. 130-137, 1980

[12] D. Carl, J. Clausen, M. Hassler, A. Zund, "Mashups programmieren," O'Reilly Germany, pp. 51-53, 2008

[13] P. Mayr, F. Tosques, "Webometrische Analysen mit Hilfe der Google Web APIs," Information Wissenschaft und Praxis, vol. 56(1), pp. 41-48, 2005.

[14] A. Hotho, "Clustern mit Hintergrundwissen," Diss., Uni Karlsruhe, p. 29, 2004.

[15] D. Thorleuchter, „Finding Technological Ideas and Inventions with Text Mining and Technique Philosophy," C. Preisach, H. Burkhardt, L. Schmidt-Thieme, R. Decker Eds. "Data Analysis, Machine Learning, and Applications," Springer, Berlin, pp. 413—420, 2008.