2011 International Conference on Uncertainty Reasoning and Knowledge Engineering

Companies Website Optimising concerning Consumer's searching for new Products

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

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

***Abstract*— In this paper, we focus on consumer's needs for new products. Normally consumers search for these new products in internet by using a web search engine. They focus on websites of specific selected enterprises where they suppose the existence of such new products. Very often, consumers only get information from enterprise websites about existing products. However, these products do not fulfill consumer's needs for new products. Then, they probably request the enterprise hotline to ask for these new products. The manual response process is time und cost consuming.**

If we identify consumer's needs for new products by time then we easily can preview consumers search queries and their intention. With this information, we can support marketing professionals to optimize the enterprise website concerning consumer's searching for new product ideas. Therefore in this paper, we present a semi-automatically approach that firstly extracts consumer's needs for new products from the internet and that secondly compares these new product ideas with existing or future products from the enterprise. After company's decision to realize the new product ideas as product in future or not, the approach thirdly gives recommendation to marketing professionals for optimizing an enterprise website in a way that consumers find information about their new needs. This increases customer satisfaction and reduces expenditure of time and cost.

Web Mining; Text Classification; Text Mining; Knowledge Discovery

I. INTRODUCTION

Today, many consumers use the internet to search for existing products. There, they probably find a corresponding product web page of an enterprise. This is because normally an enterprise provides information about existing products on the enterprise website. Additionally, consumers also have needs for new products and search for them in the internet specifically on websites of enterprises where they suppose the existence of a corresponding product [12]. Very often, such a corresponding product that fulfils consumer's needs for a new product does not exist. Then, a corresponding product web page does not exist, too. In this case, consumers do not get any information about new products that fulfill their needs from the enterprise website.

In this paper, we give recommendations to marketing professionals for optimizing company's websites concerning consumer's searching for new products. For marketing professionals two different cases are relevant. Firstly, a new product idea will be realized in future and secondly, a new product idea will not be realized in future. The third case, a new product idea is already realized as product is trivial because in this case web pages on company's website already exist describing the product and thus, the new product idea.

II. BACKGROUND

We can find information in the internet - e.g. in web logs [4] or in electronic commerce systems - that describes consumer's needs for new products [3]. There, new product ideas that are not equal to existing products are described in form of textual information. Key words that are in a description can be used to represent the corresponding new product idea.

Consumers search for these new product ideas in internet by using a web search engine. They provide a search query that consists of key words from a new product idea and they limit their search on websites of enterprises where they suppose the existence of such new products [16]. However, they do not get any information about the new product idea from the enterprise website. This is because descriptions of these new product ideas normally are not published on enterprise websites if an existing product does not realize them.

The set of key words from a new product idea can be divided in two subsets [14]. On one hand, we see key words that occur together in a description of a new product idea and of an existing product from the enterprise. In the remainder of this paper, these key words will be named known terms. Additionally, we see key words that only occur together in a description of a new product idea but not in a description of an existing product. Furthermore, these terms are named unknown terms.

A new product idea consists of both, a combination of known terms and of unknown terms. This is because if the product idea is new then not all terms that describe the new product idea must occur together in a description of an existing product (unknown terms) otherwise it is not new. Additionally, if the product idea is useful to the enterprise, some terms that describe the new product idea must occur together in a description of an existing product (known

terms) [12]. Thus, the search queries that consumers use to find products behind a new product idea also consists of a combination of known terms and of unknown terms.

III. METHODS

A general approach for optimizing consumer's searching for new products can be described as follow (see Fig. 1).

With an existing approach [13], we extract new product ideas from the internet concerning existing products of an enterprise. Each new product idea is represented by a text phrase from query result of an internet search engine. Terms (words) from each new product idea are classified automatically as known or unknown terms. In a manually process, the company decides whether a new product idea will be realized as future product or not (see Sect. IV). Then, we create recommendations to marketing professionals about optimizing consumer's searching for new product ideas (see Sect. V) and show the current optimizing status of existing web pages from the company. (see Sect. VI).

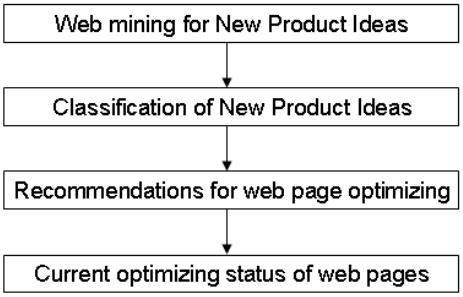


Figure 1. Processing of the website optimising approach

IV. NEW PRODUCT IDEA CLASSIFICATION

This web mining approach extracts new product ideas concerning a provided description of an existing product. The enterprise decides whether an extracted new product idea will be realized as new product or not. This decision is necessary to distinguish between product ideas that are already realized, product ideas that will be realized in future, and product ideas that will never be realized. Therefore, we support this decision by realizing a marketing recommendation and decision tool as web based application (see Fig. 2). Here, all extracted new ideas are presented and each idea can be classified manually. The web mining approach has a precision value of about 20 % (see evaluation in [13]). This means, four of five results do not represent a new and useful idea. Thus, we also offer the possibility to delete these results.

Further, we support the decision that an existing product already realizes a new product idea with text classification. For this, the user provides a description about this existing product. Then we use Jaccard's coefficient [7] to measure similarity between the new idea and the provided product description. For comparing, terms from short description of the new product idea and terms from the product description are tokenized [11] by using the term unit as word. Then, these terms are stop word filtered [10] by use of a standard strop word list and stemmed [2] by use of Porter stemmer [9]. Then all terms are aggregated to a representative set for the new idea and to a representative set for the product description. After this, both sets are transformed into binary term vectors in $\{0, 1\}^n$ concerning vector space model [15]. Here vector components are based on terms in the union of both sets. A vector component of a new product idea is one if the term occurs in the representative set of this new product idea and it is zero if the term does not occur [8]. Additionally, a vector component of a product description is one if the term occurs in the representative set of this product description and it is zero if the term does not occur [5].

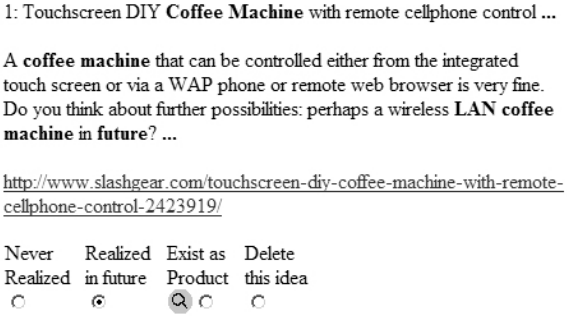


Figure 2. This is a screen shot of the recommendation tool for marketing professionals. Here new product ideas are presented with title, short description and an internet link. Users can select whether a new product idea already exists as product in the enterprise, a product idea will be realized in future, or a product idea will not be realized.

Let a be a the set of terms (words) representing a new product idea. Let b be a set of terms (words) representing the product description. Let $n = |a \cup b|$ be the cardinality of the union of both sets. Let $w \in \{0,1\}^n$ be a term vector in vector space model concerning a. Let $p \in \{0,1\}^n$ be a term vector in vector space model concerning b. Then Jaccard's coefficient is a measure for similarity between a new product idea and a product description:

$$s(w,p) = \frac{\sum_{k=1}^{n} w_k p_k}{\sum_{k=1}^{n} w_k + \sum_{k=1}^{n} p_k - \sum_{k=1}^{n} w_k \cdot p_k} \quad (1)$$

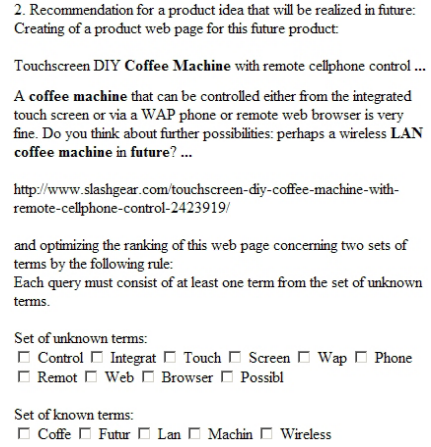
For comparing, we use the well-known Jaccard's coefficient measure because it considers the different sizes of both vectors [6]. The Jaccard's result values are always between 0 % and 100 % for each new product idea and product description combination. For this, we give the general recommendation that if the result value x is small ($x < \gamma$) then the provided product does probably not realize the selected product idea. If the result value is large ($x > \delta$) then the provided product probably realizes the selected product idea. Therefore, we define γ and δ as percentages and determine their values in the evaluation. In case of $\gamma < x < \delta$ we cannot give a certain recommendation.

V. RECOMMENDATIONS FOR WEB PAGE OPTIMIZING

To identify terms in search queries from consumer's searching for new products, we determine the length of

search queries that means the number of terms in a search query first. In this approach, the length of a search query is determined to four terms. This is because if a consumer has a new product idea in his mind and he is searching for this idea in the internet then he normally uses several terms in a search query that describe his new product idea. For this, he uses known terms that are already part of an existing product as well as unknown terms that represent the new part of a product idea. If the person uses more than four terms in a search query then he possibly does not get any results because his search query is too specific. If the person uses less than four terms in a search query then normally he cannot describe both, the part of an existing product and the part of a new product idea. Therefore, we think that a search query that consists of four terms is a good compromise.

If the enterprise decides that a new product idea will be realized as new product in future then we recommend creating and publishing a web page for this new product. It consists of information about time to market, expected price etc. To optimize the ranking of this new web page, we select key words for the new product idea. The idea is described by a short description text phrase that consists of known and unknown terms.


2. Recommendation for a product idea that will be realized in future:
Creating of a product web page for this future product:

Touchscreen DIY **Coffee Machine** with remote cellphone control ...

A **coffee machine** that can be controlled either from the integrated touch screen or via a WAP phone or remote web browser is very fine. Do you think about further possibilities: perhaps a wireless **LAN coffee machine** in **future**? ...

http://www.slashgear.com/touchscreen-diy-coffee-machine-with-remote-cellphone-control-2423919/

and optimizing the ranking of this web page concerning two sets of terms by the following rule:
Each query must consist of at least one term from the set of unknown terms.

Set of unknown terms:
☐ Control ☐ Integrat ☐ Touch ☐ Screen ☐ Wap ☐ Phone
☐ Remot ☐ Web ☐ Browser ☐ Possibl

Set of known terms:
☐ Coffe ☐ Futur ☐ Lan ☐ Machin ☐ Wireless

Figure 3. Recommendation tool for marketing professionals: Here a new product idea is presented. We also see all unknown key words and all known key words from the set of terms that represents this new product idea. We assume that it is planned to realize this new product idea by a future product of the enterprise. Then, the web page of the future product can be optimized concerning search queries that consist of terms from the set of known and unknown terms.

A consumer who searches specifically for this new product idea uses a search query that consists of these unknown terms or that consists of a combination of known and unknown terms. Known and unknown terms are automatically identified by use of tokenization (where the term unit is word), part of speech tagging, stop word filtering and stemming. The results are presented to the marketing professional who can select and discard terms for further processing. Search queries are created based on the manual selected terms.

Therefore, we create search queries from the new product idea that consist of four different selected terms in total and that consist of at least one unknown selected term from the representative set of the new product idea. Then, we recommend using these search queries for optimizing the ranking of this new web page (see Fig. 3). Examples for the created search queries in Fig. 3 are

– Control Coffe Futur Lan
– Control Integrat Touch Screen
– Control Phone Coffe Machin
– Remot Web Lan Wireless

VI. CURRENT OPTIMIZING STATUS OF WEB PAGES

To support marketing professionals by optimizing the new web page, we check the current status of the newly created web page. Thus, we compute the current ranking position of this web page concerning the recommended search queries (see Sect. V). With this information, one can optimize the web content.

The user provides the address of the enterprise website (e.g. http://www.text-mining.info) where the address of an internet link of the newly generated web page (e.g. http://www.text-mining.info/product.html) is included. Then, all search queries from the recommendation in Sect. V are executed. For this, we use web services from a search engine. A web service is a software system that is designed to support interoperable machine-to-machine interaction over a network. Web services often are realized as web based advanced programming interfaces [9]. This means, we can access to these interfaces over the internet. The web search engine Google offers such a web service [1]. Here, the requested service is to execute a search query. Then, the query result data is transferred back to an application that requested the service. By use of Google web services, the maximum number of query results is limited to one result page. Each result page consists of eight query results. This means, if we execute a search query via Google then we also have to provide a result page number and Google transfers the result data of the selected result page back.

To get more than eight query results, we have to execute each recommended search query three times. This means, we separately select query results on the first, second, and third result page. In total, we get 24 query results. Each query result consists of a title, a short description and an internet link. We compare the internet link of each query result to the provided web page. Then, for each search query, we present the first query result that equals the provided web page or that is part of the enterprise website (see Fig. 4).

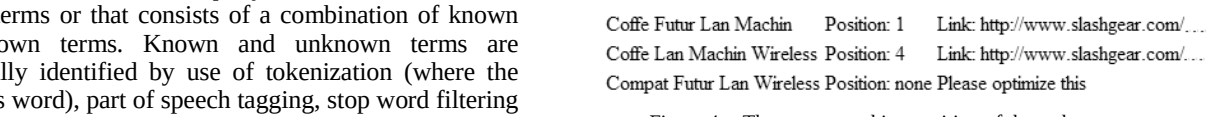

Coffe Futur Lan Machin Position: 1 Link: http://www.slashgear.com/...
Coffe Lan Machin Wireless Position: 4 Link: http://www.slashgear.com/...
Compat Futur Lan Wireless Position: none Please optimize this

Figure 4. The current ranking position of the webpage http://www.slashgear.com/touchscreen-diy-coffee-machine-with-remote-cellphone-control-2423919/ is presented concerning selected queries. We see that a web page on this website is in good position by using the search queries: (Coffe, Futur, Lan, Machin) and (Coffe, Lan, Machin, Wireless). However, using the search query (Compat Futur Lan Wireless) does not lead to any web pages on this website among the first 24 query results. Therefore, we recommend optimizing an enterprise web page concerning this search query.

VII. EVALUATION

Here, we present an approach for semi-automatically generating of recommendations. For this, we use information about new product ideas from consumers as input. We support the decision whether this new product idea is already realized by an existing product of the enterprise. This is done by comparing term vectors of new product ideas to term vectors of existing products. For this, we have to determine and evaluate the parameters γ and δ. After this, we create recommendations to marketing professionals for creating and optimizing product web pages. These recommendations are heuristically but not theoretically founded. Therefore, it is crucial to provide an evaluation to show their success.

The first aspect in our evaluation is to determine the value of the parameters γ and δ. We use these parameters for comparing term vectors of new product ideas to term vectors of existing products by using the well-known Jaccard's coefficient measure. If the Jaccard's coefficient result value x is smaller than γ then the existing product does not contain the selected product idea and therefore, the user should not select that the new product idea already exists as product in the enterprise (see Fig. 2). Additionally, if the Jaccard's coefficient result value x is greater than or equal to δ then the existing product contains the selected product idea. In this case, we recommend the user to select the corresponding option in Fig. 2. A human expert extracts about 30 new product ideas from the product idea web log mining approach. He compares these ideas to several product descriptions. As a result, he finds out that a product does not realize a new idea if the corresponding Jaccard's coefficient result value x is smaller than 30 %. Additionally, he finds out that if the Jaccard's coefficient result value is greater than or equal to 70 % then the new idea is already realized by the product. If a new idea is compared to a product and the corresponding Jaccard's coefficient result value is $30\% \leq x < 70\%$ then in some cases the new product idea is realized and in other cases it is not realized. Therefore, we set γ to 30 %, we set δ to 70 %, and we cannot give a certain recommendation for $30\% \leq x < 70\%$.

The second aspect in our evaluation is to evaluate the recommendations. We compare this approach to a baseline model because we are not aware of other approaches specifically for this kind of recommendation at the present time. As measure for the baseline, we use the chance baseline, which assigns a classification randomly. Here we have two classes (A means we get successful recommendations by the approach concerning a new product idea, B means we get unsuccessful recommendations) in our data, and we classify each instance (new idea) with a specific percentage as either A or B. For the chance baseline, we set this percentage to 50 %.

To compute this percentage for our approach, we use 100 new product ideas from the product idea web log mining approach. Then we create recommendations for each new product idea. A human expert analyses the recommendations of our approach. As a result, he finds out that 70 % of the recommendations are successful.

We do not get a higher percentage because sometimes our approach recommends optimizing web pages concerning irrelevant terms. This is because our approach distinguishes between stop words and non-stop words. Therefore, we recommend optimizing a web page concerning terms that are non-stop words. However, these terms are probably no domain specific terms and they are irrelevant for the product description. We get better results for the percentage if we focus on domain specific terms by using a domain specific stop word list. However, creation and maintenance of such a domain specific stop word list is time and cost consuming. Thus, we use a standard stop word list for our approach. This is because our approach is better than the chance baseline.

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

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

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

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

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

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

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

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

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

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

[11] D. Thorleuchter., D. Van den Poel, "Semantic Technology Classification - A Defence and Security Case Study," in: Proc. URKE 2011, IEEE Press, Los Alamitos, CA, 2011.

[12] D. Thorleuchter., D. Van den Poel, and A. Prinzie, "Mining Innovative Ideas to Support new Product Research and Development," in: Classification as a Tool for Research, H. Locarek-Junge and C. Weihs Eds. Berlin-Heidelberg-New York: Springer, 2010, pp. 587-594.

[13] D. Thorleuchter, D. Van den Poel, and A. Prinzie, "Extracting Consumers Needs for New Products - A Web Mining Approach," in: Proc. WKDD 2010, IEEE Computer Society, Los Alamitos, CA, 2010, p. 441.

[14] 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.

[15] D. Thorleuchter., D. Van den Poel, and A. Prinzie, "A compared R&D-based and patent-based cross impact analysis for identifying relationships between technologies," Technological Forecasting and Social Change, vol 77 (7), pp. 1037-1050, 2010.

[16] D. Thorleuchter., D. Van den Poel, and A. Prinzie, "Mining Ideas from Textual Information," Expert Systems with Applications, vol. 37 (10), pp. 7182-7188, 2010.