7-1-2013

An Exploratory Study On The Appropriateness Of Latent Dirichlet Allocation For Automatic Discovery Of Product Associations From User-Generated Content

Johannes Putzke
University of Cologne, Cologne, Germany, putzke@wim.uni-koeln.de

Kai Fischbach
University of Bamberg, Bamberg, Germany, kai.fischbach@uni-bamberg.de

Detlef Schoder
University of Cologne, Cologne, Germany, schoder@wim.uni-koeln.de

AN EXPLORATORY STUDY ON THE APPROPRIATENESS OF LATENT DIRICHLET ALLOCATION FOR AUTOMATIC DISCOVERY OF PRODUCT ASSOCIATIONS FROM USER-GENERATED CONTENT

Johannes Putzke, University of Cologne, Department for Information Systems and Information Management, Pohligstr. 1, 50969 Cologne, Germany, putzke@wim.uni-koeln.de

Kai Fischbach, University of Bamberg, Department of Information Systems and Social Networks, An der Weberei 5, 96047 Bamberg, Germany, kai.fischbach@uni-bamberg.de

Sonja Gensler, University of Muenster, Marketing Center Muenster, Am Stadtgraben 13-15, 48143 Muenster, Germany, s.gensler@uni-muenster.de

Javier Sanchez, University of Cologne, Department for Information Systems and Information Management, Pohligstr. 1, 50969 Cologne, Germany, sanchez@wim.uni-koeln.de

Detlef Schoder, University of Cologne, Department for Information Systems and Information Management, Pohligstr. 1, 50969 Cologne, Germany, schoder@wim.uni-koeln.de

Franziska Völckner, University of Cologne, Department for Marketing and Brand Management, Albertus-Magnus-Platz 1, 50923 Cologne, Germany, voelckner@wiso.uni-koeln.de

Abstract

Latent Dirichlet Allocation (LDA) is a method that can be used to generate word association networks from unstructured text documents. However, no study has yet examined the applicability of LDA for deriving product associations from user-generated content. In this work, we apply LDA on 9,529 unstructured and uncategorized McDonald's product reviews that were crawled from a German online review platform. We evaluate the applicability of LDA for deriving product associations from user-generated content. For this reason, we conducted a survey among 95 Information Systems undergraduate students about their associations with 17 McDonald's-related nouns. Results indicate that LDA is a valid method for deriving product associations from user-generated content.

Keywords: Product Associations, Latent Dirichlet Allocation, LDA, User-generated content

1 Introduction

Researchers in the domain of Business Intelligence (BI) intend to provide decision makers with consolidated, timely, and accurate information. Despite the vast amount of BI applications, there has been surprisingly little research on BI applications for market research (Decker & Trusov 2010, Netzer & Feldman & Goldenberg & Fresko 2012).

Latent Dirichlet Allocation (LDA) (Blei & Ng & Jordan 2003) is a method that can be used to generate word association networks (and thus "brand concept maps" (John & Loken & Kim & Monga 2006, Schnittka & Sattler & Zenker 2012)) from unstructured text documents. It is also able to account for the different semantic contexts in which the words are used. Although LDA has found wide acceptance in academic text mining research, there is surprisingly little research on the applicability of LDA for deriving word association networks. Particularly, to the best of our knowledge there are no LDA applications for consolidating user-generated content from consumers' online review platforms such as http://www.epinions.com and http://www.ciao.de into product association networks or word association networks, respectively.

Deriving product association networks from consumer online reviews is of high relevance for a managerial audience for two main reasons. First, a principal concern of marketing managers is how to position their products. In this context, product association networks are an effective tool for product management. However, it is very costly to create product association networks with traditional methods. Analyzing user-generated content with LDA provides a low-cost alternative. Furthermore, the validity bias to which deriving product association networks from Internet behavioral data is subject is different than those for methods that elicit product associations by surveying respondents in artificial environments. This should provide another perspective on the consumers' product associations, one in which they express their associations and personal experiences with a product voluntarily without being asked and knowing the aim of the study (as opposed to being asked specifically to reveal their associations). Second, marketing managers are often not certain of the best keywords to use in advertising and for search engine optimization. LDA unmasks hidden word associations from the "long tail" (Anderson, 2006) and hence may help marketing managers lower their costs for ad word campaigns.

However, no study has yet examined the applicability of LDA for deriving product associations from user-generated content. In this work, we apply LDA on 9,529 unstructured and uncategorized McDonald's product reviews that were crawled from a German online review platform. The main objective of this paper is to evaluate the applicability of LDA for deriving product associations from user-generated content. For this reason, we conducted a survey among 95 Information Systems (IS) undergraduate students about their associations with 17 McDonald's related nouns.

The present study is structured as follows: Section 2 provides an introduction into LDA as well as a short literature review about LDA. Section 3 illustrates the applicability of LDA for deriving product associations. Finally, Section 4 offers a brief summary, a discussion of the theoretical and managerial implications of the research, limitations of the research as well as an outlook towards future research.

2 Probabilistic Topic Model of Latent Dirichlet Allocation

Since LDA is not well known in the IS community, this section reviews the basics of LDA. In the first subsection, 2.1, we highlight the basic principles of probabilistic topic models. In the second subsection, 2.2, we introduce the mathematical notation of LDA models. In the last subsection, 2.3, we provide a short literature review about LDA.

2.1 Basic Principle

Probabilistic Topic Models such as LDA assume that documents from a set of *D* documents (a "document collection") are generated using a mixture of topics (Steyvers & Griffiths 2006). Each topic, in turn, is a probability distribution representation over a fixed collection of terms. Here, we consider a term as the abstraction of words or word-tokens of a document. In a document, the sequence of words "to eat or not to eat" will result in a total of 6 word-tokens (*to, eat, or, not, to, eat*) and a total of 4 terms (*to, eat, or, not*). The mixtures of topics can be understood as the latent or hidden thematic structure of a document collection.

Figure 1 illustrates how *document d1* can be generated using three topics with a particular probability. There is a probability of .65 that Topic 3 will be responsible for generating the words present in *document d1*. Topic 1 and Topic 2 have a probability of .25 and .10 of belonging to *document d1*. Like *document d1*, each document in the document collection will have a unique probability distribution of topics from which it is generated. The three different topics, in turn, are composed of 18 terms with different probabilities. In this way, we can deal with polysemy problems, identifying different semantic contexts for the same words. Therefore, each topic identified by LDA can be understood as an independent semantic interpretation of all the documents.

In our McDonald's example in Figure 1, *document d1* is most related to *ice* and *apple Danish.* It would be very likely that this document was generated from topics with words such as *ice, flurry, Smarties*, with a higher probability than from topics with other words. A document related to a McDonald's milkshake would likely be generated from a topic containing words such as *milkshake, vanilla, cup,* or *milk,* with a high probability.

Table 2 highlights three topics related to McDonald's products – *ice (topic 1), milkshake (topic 2),* and *apple Danish (topic 3)* – as well as the probabilities of each of the 18 terms in these 3 topics[1].

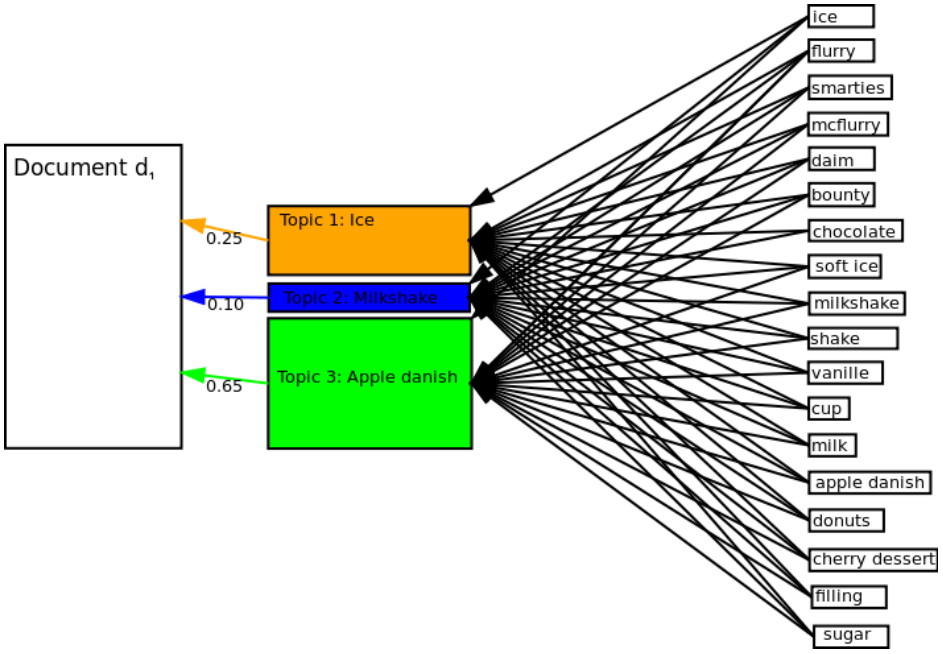

Figure 1. Document d_1 is based on 3 topics, each of them having an 18 dimensional probability distribution over terms

[1] The reviews on the German online review platform were written in the German language. Hence, some lines in the table consist of several words (e.g. Apple Danish) since we are not aware of an English translation of the original German phrase (e.g. Apfeltasche) in a single word.

Topic1: Ice	Prob.	Topic2: Milkshake	Prob.	Topic3: Apple Danish	Prob.
Ice	.3333	Milkshake	.3800	Apple Danish	.3333
Flurry	.2800	Shake	.2800	Donuts	.2800
Smarties	.1300	Vanilla	.1300	Cherry dessert	.1300
McFlurry	.0900	Cup	.0900	Filling	.0900
Daim	.0600	Milk	.0600	Sugar	.0600
Bounty	.0500	Softy	.0030	Softy	.0500
Chocolate	.0400	Chocolate	.0400	Chocolate	.0400
Softy	.0030	Ice	.0020	Milkshake	.0030
Milkshake	.0020	Flurry	.0015	Shake	.0020
Shake	.0015	Smarties	.0010	Vanilla	.0015
Vanilla	.0010	McFlurry	.0010	Cup	.0010
Cup	.0010	Daim	.0010	Milk	.0010
Milk	.0010	Bounty	.0010	Ice	.0010
Apple Danish	.0010	Apple danish	.0010	Flurry	.0010
Donuts	.0010	Donuts	.0010	Smarties	.0010
Cherry dessert	.0010	Cherry dessert	.0010	McFlurry	.0010
Filling	.0010	Filling	.0010	Daim	.0010
Sugar	.0010	Sugar	.0010	Bounty	.0010
Sum of probabilities	1		1		1

Table 2. Topic Distribution

2.2 Mathematical Notation

As Figure 1 and Table 2 illustrate, probabilistic topic models assume that all of the *D* documents from a document collection are generated from *T* different topics. Each document *d* has a probability to belong to a particular topic *j* [*P(z=j)*] and each word w_i has a probability of occurrence in a topic *j*, $P(w_i|z_i=j)$. W_i refers to the *i*th word-token in document *d*. We can represent all the *P(z=j)* of a document *d* with a *D T*-dimensional multinomial random variable θ_d. Each element in θ_d will have the probability of assigning the index element as topic number to the document. For example, the corresponding θ_d to Figure 1 would be θ_{d_1} = (.25; .10; .65). Also, all the N_d word-tokens in document *d* can be generated using a *T*-mixture of *W*-dimensional multinomial random variables ϕ_1..... ϕ_T . ϕ_z that will represent the probability distribution for topic number *z* over all the terms collection. *T* and *W* refer to the total number of topics and terms for the document collection. In our example, *W*=18 terms can be modelled in *T*=3 different topics (semantic contexts). The number of total terms need not match the N_d word tokens for each document, due to the fact that a term can be instanced more that once by several word-tokens in a document. All the word-tokens for the document collection will be represented by $N = \sum N_d$.

Each ϕ_z topic distribution over the terms collection such as *ice*, *milkshake* and *apple Danish* will be represented by ϕ_1, ϕ_2 and ϕ_3, respectively. Each ϕ_z refers to the probability distribution over all the *W* terms given topic *z*, also symbolized by $\phi_z = P(w|z)$ (Steyvers & Griffiths 2006).

Generating *document* d_1 from Figure 1 has a high probability to belong to topic 1 and topic 3, suppose that we choose from our given document-topic distribution θ_{d_1} the topic indices 1 and 3. Then, for each word *w* in document d_1 we choose one of the desired topic indices $z = 1$, $z = 3$ at a time, and sample the desired word *w* from the corresponding ϕ_z. For another document d_2 about *milkshake*, we choose the topic index $z = 2$ and, again, for each word *w* in d_2 we sample the needed words from that topic ϕ_2. This process is repeated iteratively until all *D* documents are created.

The probability of each of the word-tokens w_i for document d can be computed first by considering the probability that a topic j can be chosen for a word w_i in a document d and second by considering the probability that a word w_i was sampled from that word-topic distribution with index j (Steyvers & Griffiths 2006). The probability that a word w_i belongs to a document can, hence, be expressed as

$$P(w_i) = \sum_{j=1}^{T} P(w_i|z_i = j)\, P(z_i = j) \quad (1)$$

One could employ maximum likelihood (ML) estimation to estimate the model specified in (1). However, doing so might lead to undesirable outcomes (compare, for example, Griffiths & Steyvers 2004). Hence, Blei et al. (2003) propose in LDA the introduction of Dirichlet priors on each to the θ_d and ϕ_z. These priors are a D-dimensional $Dir_D(\alpha)$ and T-dimensional Dirichlet distribution $Dir_T(\beta)$. In the case of each hyperparameter, α from $Dir(\alpha_1 \ldots \alpha_T)$ can be understood as a prior observation count for the number of times topic j is sampled in a document before any actual words have been observed from that document (Steyvers & Griffiths 2006). This allows us to have a complete generative model for document generation. The model specifies a probabilistic procedure by which new documents can be produced given a set of topics $\phi_1 \ldots \phi_T$. This a priori knowledge can be understood as knowing in advance what the probability distribution of θ_d and ϕ_z will look like given the scalar parameters α and β, respectively.

The complete generative process described above can also be inverted to find the topics responsible for the generation of the words in the documents. In other words, given the words of the documents, with statistical techniques we can infer the topics responsible for generating the text documents. We are interested in the probability distributions $\theta_1 \ldots \theta_D$ and $\phi_1 \ldots \phi_T$ and in evaluating the posterior distribution:

$$P(z|w) = \frac{P(w,z)}{\sum_z P(w,z)} \quad (2)$$

This inference problem is resolved using the Gibbs sampling algorithm (see Steyvers & Griffiths 2006 for a detailed description of the Gibbs sampling algorithms).

2.3 Literature Review

In this subsection, we provide a short literature review about LDA. LDA is part of the larger field of probabilistic topic models (for a literature review see Blei 2012). It was introduced to solve a problem with probabilistic semantic analysis. Since its introduction it has been adapted and extended in many ways (compare Blei 2012). For example, some authors tried to relax the assumptions of LDA, such as the assumption of a fixed number of topics (Teh & Jordan & Beal & Blei 2006), the bag of words assumption (Wallach 2006), or the assumption of uncorrelated occurrence of topics (Blei & Lafferty 2007, Li & McCallum 2006). Other variations of LDA that relax some assumptions from the original model are, for example, bursty topic models (Doyle & Elkan 2009). Another stream of research that extends the classical LDA model are models that try to include metadata into the analyses, such as the author-topic model (Rosen-Zvi & Griffiths & Steyvers & Smyth 2004), the relational topic model (Blei & Griffiths & Jordan 2010), Markov topic models (Wang & Thiesson & Meek & Blei 2009) and supervised topic models (e.g. Zhu & Ahmed & Xing 2012).

3 Applicability of LDA for Deriving Product Associations

We applied LDA (e.g. Blei et al. 2003, Steyvers & Griffiths 2006) on 9,529 unstructured and uncategorized McDonald's product reviews that were crawled from a German online review platform. We picked LDA for our work because it can deal with polysemy problems of words, identifying different semantic contexts for the same word. For other advantages of applying LDA see Griffith, Steyvers and Tenenbaum (2007).

We conducted our analysis relying on a three-step approach suggested by Fayyad, Piatetsky-Shapiro, and Smyth (1996): (1) data pre-processing, (2) data mining, and (3) post-processing of data mining results.

As a first pre-processing step, we removed all html tags from the documents (i.e. reviews). We then reduced the vocabulary size of our dataset to nouns from the documents, on the assumption that nouns capture product names with a high probability. For the annotation of part-of-speech information as well as for lemmatization we used TreeTagger (Schmid, 1994). The resulting vocabulary contains 35,105 terms.

As a second step, data mining, we applied LDA. As LDA-specific parameters for the Dirichlet priors on θ_d and $\phi_{z,}$, $Dir_D(\alpha)$*and* $Dir_T(\beta)$, we chose $\alpha = 50/number of topics = 1$ and $\beta = .001$. The parameter α was chosen so that, in our model, each document has the same a priori possibility of belonging to any topic with each iteration. The parameter of $Dir(\beta)$ was chosen to be $\beta = .001$, since this was found to work well with many different text collections (Steyvers & Griffiths 2006). For our analysis, we also chose the number of topics to be 50, because a pre-study with 25, 50, 100, 150 and 200 topics showed that the LDA with 50 topics worked best for our dataset (using precision and recall as quality criteria).

We used the Gibbs sampling algorithm to sample a topic distribution for each document 1,000 times, and obtained an approximative steady state. For the LDA implementation we used the JAVA packet MALLET (McCallum 2002).

In the third step, post-processing, we computed the similarity among each pair of words present in the document collection based on the 50 discovered latent topics. Afterwards, we ranked them according to the similarity weight as follows: given a pair of words: (w1, w2), the more joint probability of words w1 and w2 to be in the same topics, the more similar they should be. That is:

$$P(w_2|w_1) = \sum_{j=1}^{T} P(w_2|z = j)P(z = j|w_1) \quad (3)$$

In order to evaluate our results we randomly selected 17 nouns from a pool of words people associate with McDonald's (see the first column in Table 3). The pool of words was created in a brainstorming session by the authors. Using LDA, we then figured out the three words with the highest similarity to each of these 17 nouns. Table 3 illustrates these results.

Noun	Word 1	Word 2	Word 3
1. McChicken	Soße (*sauce*)	Burger	Brötchen (*roll*)
2. McFlurry	Eis (*ice*)	Smarties (*M&M's*)	Daim
3. McNuggets	Chicken	Soße (*sauce*)	Curry
4. McRib	Sauce	Burger	Fleisch (*meat*)
5. Cola (*coke*)	Wasser (*water*)	Getränk (*beverage*)	Becher (*cup*)
6. Milchshake (*milkshake*)	Vanille (*vanilla*)	Geschmack (*taste*)	Schoko (*chocolate*)
7. Apfeltasche (*apple Danish*)	Donuts	Kirschtasche (*cherry Danish)*	Füllung (*filling*)
8. Kaffee (*coffee*)	Frühstück (*breakfast*)	Kakao (*cocoa*)	Milch (*milk*)
9. Maxi	Menü	Getränk (*beverage*)	Pommes (*fries*)
10. BigMac	Geld (*money*)	Mark	Gibt's
11. Hamburger	Cheesburger	Ketchup	Gurke (*cucumber*)
12. Fett (*fat*)	Pommes (*fries*)	Preis (*price*)	Kalorien (*calories*)
13. Kalorien (*calories*)	Preis (*price*)	Fett (*fat*)	Geschmack (*taste*)
14. Frühstück (*breakfast*)	Kaffee (*coffee*)	Kakao (*cocoa*)	Milch (*milk*)
15. Preis (*price*)	DM	Geschmack (*taste*)	Fazit (*result*)
16. Cheesburger	Hamburger	Ketchup	Gurke (*cucumber*)
17. Salat (*salad*)	Burger	Soße (*sauce*)	Dressing

Table 3. Words with the highest similarity to 17 nouns related to McDonald's (using LDA)

Furthermore, we also asked four of our colleagues who were not aware of the research project to suggest three words they associate with each of these nouns related to McDonald's. (The average age of the colleagues was 29.75 yrs with a standard deviation of .96 yrs). Table 4 highlights our colleagues' answers as well as our LDA results for the noun "McRib".

Colleague 1	Fleisch (*meat*)	Grill (*barbecue)*	Streifen (*stripes*)
Colleague 2	Soße (sauce)	Barbecue	Serviette (napkin)
LDA	Sauce (*sauce*)	Burger	Fleisch (*meat*)
Colleague 3	Deutschland (*Germany*)	Schwein (*pig*)	Burger
Colleague 4	Soße (*sauce*)	Zwiebel (*onion*)	Mann (*man*)

Table 4. Words associated with the noun "McRib"

Afterwards, for each of the 17 nouns we handed out a deck of five cards to 95 IS undergraduate students at a leading European research university. (The average age of the students was 21.26 yrs with a standard deviation of 2.56 yrs. 24 percent of the students were female, 76 percent of the students were male.) Each of four cards contained the three words provided by one of our four colleagues; the remaining card contained the three words our IS classified as being closest to one of the 17 terms highlighted in Table 3. The students were then asked to guess which card was created by the IS. Particularly, we asked the following question (translated from German language). "We asked 4 persons which 3 nouns they associate with the following 17 nouns related to McDonald's. We also created answers to this task with a software. For each of the nouns, 4 cards origin from real persons, one from the computer system. Please cross the card that according to your opinion was created by the computer system". We expected that the choice probability for each card should be π_{ij}=.2 in the case that LDA was a valid methodology for extracting product associations from unstructured text documents on the German online review platform.

If a student could not decide on a particular card, his or her answer to the question/noun was treated as a missing value. We could not identify any obvious pattern in the missing values. The number of missing values for each noun can be calculated from Table 5 by subtracting N from 95 (e.g. there were 95-93=2 missing values for the noun "McChicken").

Noun	$X^2(df, N), p$
1. McChicken	$X^2(4, N=93) = 38.344 , p<.01$
2. McFlurry	$X^2(4, N=90) = 13.556 , p<.01$
3. McNuggets	$X^2(4, N=91) = 44.110 , p<.01$
4. McRib	$X^2(4, N=88) = 12.455 , p<.05$
5. Cola (*coke*)	$X^2(4, N=92) = 19.630 , p<.01$
6. Milchshake (*milkshake*)	$X^2(4, N=93) = 4.366 , p<.36$
7. Apfeltasche (*apple Danish*)	$X^2(4, N=92) = 57.348 , p<.01$
8. Kaffee (*coffee*)	$X^2(4, N=93) = 26.839 , p<.01$
9. Maxi	$X^2(4, N=90) = 58.222 , p<.01$
10. BigMac	$X^2(4, N=93) = 38.022 , p<.01$
11. Hamburger	$X^2(4, N=92) = 26.370 , p<.01$
12. Fett (*fat*)	$X^2(4, N=91) = 25.209 , p<.01$
13. Kalorien (*calories*)	$X^2(4, N=92) = 67.348 , p<.01$
14. Frühstück (*breakfast*)	$X^2(4, N=92) = 24.087 , p<.01$
15. Preis (*price*)	$X^2(4, N=92) = 113.326 , p<.01$
16. Cheesburger	$X^2(4, N=92) = 41.152 , p<.01$
17. Salat (*salad*)	$X^2(4, N=92) = 74.957 , p<.01$

Table 5. Chi-Square Tests on Differences in Choice Probabilities

Contrary to our expectations, chi-square tests indicated significant differences between the choice probabilities π_{ij} for the five different cards for 16 out of 17 nouns (see Table 5). Only in the case of the noun "Milchshake" (milkshake) the hypothesis of π_{ij}=.2 could not be rejected.

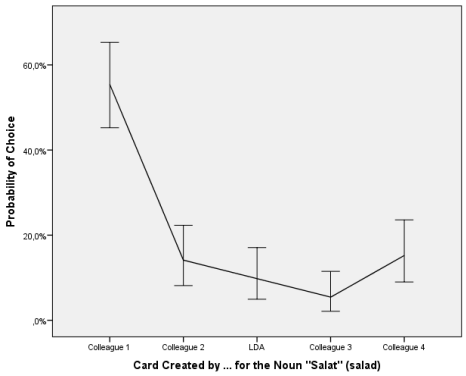

Figure 6. Choice Probabilities of Cards Containing Nouns Associated with "Salat" (salad)

Hence, we further analyzed the different choice probabilities in detail. We expected that the cards containing words from the LDA should have been chosen more often by the students than the other cards. Contrary to our expectations, the choice probabilities of the cards were similar to the ones for the noun "salad" illustrated in Figure 6. The error bars (α = .05) in Figure 6 indicate that the card that had been produced by "Colleague 1" had a very high probability of choice, whereas the remaining four cards had a low probability of choice including the card with the LDA results. Similar results were also obtained for most of the other 16 nouns (see Figure 9 in the Appendix). The only one exception in which a card with LDA results was identified as such was for the noun "McRib" (see Figure 7).

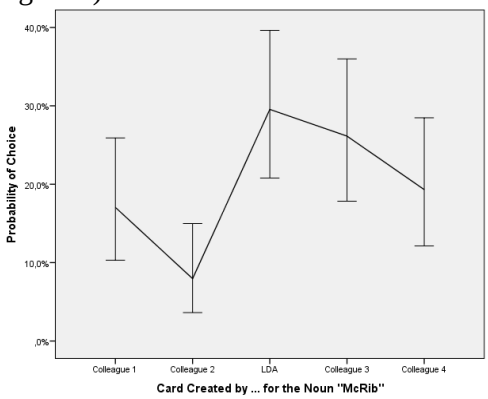

Figure 7. Choice Probabilities of Cards Containing Nouns Associated with "McRib"

In the final step of our analyses we intended to figure out whether there are some systematic variations between the choice probabilities of the cards produced by our four colleagues and the IS. Figure 8 illustrates the error bar of these choice probabilities pooled over all 17 nouns. The figure illustrates that cards produced by "Colleague 1" and "Colleague 4" were chosen with a higher probability by the students than cards produced by the other two colleagues and the card with the LDA results.

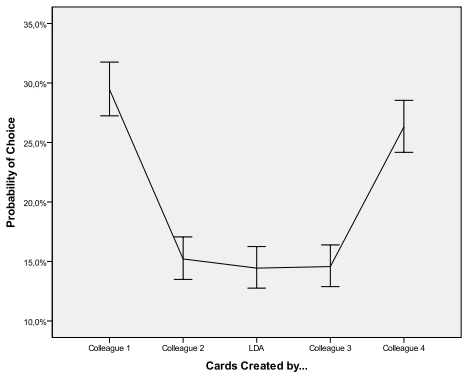

Figure 8. Choice Probabilities of Cards for all Nouns

For these findings there are two possible explanations. First, Colleague 1 and Colleague 4 have different conceptual frameworks about McDonald's products than the majority of the students and our other two colleagues. Second, cards produced by our information system base on a "shared conceptual framework" in the internet and hence are chosen with less probability than the remaining cards. However, in both cases the results indicate that LDA is a valid methodology for deriving product associations from unstructured product reviews on consumer opinion platforms such as http://www.ciao.de and http://www.epinions.com.

4 Discussion

In this paper, we applied LDA on 9,529 unstructured and uncategorized McDonald's product reviews that were crawled from a German online review platform. Results of a study with 95 IS undergraduate students indicate that LDA is a valid methodology for deriving product associations from user generated content. The results of this study have numerous theoretical and managerial implications.

Deriving product association networks with traditional methods (e.g. John et al. 2006) is an elaborate, time-consuming and expensive task. LDA offers an easy-to-conduct, low-cost alternative to these traditional methods. Although in principle LDA could replace traditional methods of eliciting product association networks, we would recommend using product association networks derived by LDA as a complement to traditional methods. This is particularly important since product association networks derived by LDA are not perfect either. Rather, they are subject to other validity biases than the product association networks derived from traditional methods. For example, product association networks derived by LDA from user-generated content are based on real behavioral data, which enhances their validity. However, these might be subject to astroturfing, which poses threats to validity.

Of course, as with any empirical study, ours is subject to some limitations that could be seen as affecting the rigor and relevance. First, as Dirichlet priors we chose $\alpha = 1$ and $\beta = .001$. However, different priors might lead to different results. Hence, future research should conduct a similar kind of study with varying values for α and β. Second, in order to obtain four reference cards we asked four of our colleagues about their associations with nouns related to McDonald's. Future research should ask a broader population about their associations. Finally, we choose a sample of 95 IS undergraduate students. Future research should conduct a similar study with a non-student sample.

Our hope is that our research will assist others in conducting these types of studies and form the basis for substantial future research into the validity of text mining for deriving product association networks.

References

Blei, D. M. (2012). Probabilistic Topic Models. Communications of the Acm, 55 (4), 77-84.

Blei, D. M., Griffiths, T. L. and Jordan, M. I. (2010). The Nested Chinese Restaurant Process and Bayesian Nonparametric Inference of Topic Hierarchies. Journal of the ACM, 57 (2).

Blei, D. M. and Lafferty, J. D. (2007). A Correlated Topic Model of Science. Annals of Applied Statistics, 1 (1), 17-35.

Blei, D. M., Ng, A. Y. and Jordan, M. I. (2003). Latent Dirichlet Allocation. Journal of Machine Learning Research, 3 (4-5), 993-1022.

Decker, R., and Trusov, M. (2010). Estimating Aggregate Consumer Preferences from Online Product Reviews. International Journal of Research in Marketing, 27 (4), 293-307.

Doyle, G. and Elkan, C. (2009). Accounting for Burstiness in Topic Models. Proceedings of the 26th Annual International Conference on Machine Learning.

Fayyad, U., Piatetsky-Shapiro, G. and Smyth, P. (1996). Knowledge Discovery and Data Mining: Towards a Unifying Framework. In U. Fayyad, G. Piatetsky-Shapiro, P. Smyth and R. Uthurusamy (Eds.), Advances in Knowledge Discovery and Data Mining (pp. 1-36). Cambridge, MA: AAI/MIT press.

Griffiths, T. L. and Steyvers, M. (2004). Finding Scientific Topics. Proceedings of the National Academy of Sciences of the United States of America, 101, 5228-5235.

Griffith, T. L., Steyvers, M. and Tenenbaum, J. B. (2007) Topics in Semantic Representation. Psychological Review, 114 (2) , 211-244.

John, D. R., Loken, B., Kim, K. and Monga, A. B. (2006). Brand Concept Maps: A Methodology for Identifying Brand Association Networks. Journal of Marketing Research, 43 (4), 549-563.

Li, W. and McCallum, A. (2006). Pachinko Allocation: Dag-Structured Mixture Models of Topic Correlations. Proceedings of the 23rd international conference on Machine learning.

McCallum, A. K. (2002). Mallet: A Machine Learning for Language Toolkit.

Netzer, O., Feldman, R., Goldenberg, J. and Fresko, M. (2012). Mine Your Own Business: Market-Structure Surveillance through Text Mining. Marketing Science, 31 (3), 521-543.

Rosen-Zvi, M., Griffiths, T., Steyvers, M. and Smyth, P. (2004). The Author-Topic Model for Authors and Documents. Proceedings of the 20th conference on Uncertainty in artificial intelligence.

Schmid, H. (1994). Probabilistic Part-of-Speech-Tagging Using Decision Trees. Proceedings of International Conference on New Methods in Language Processing.

Schnittka, O., Sattler, H. and Zenker, S. (2102). Advanced Brand Concept Maps: A New Approach for Evaluating the Favorability of Brand Association Networks. International Journal of Research in Marketing, 29 (3), 265-274.

Steyvers, M. and Griffiths, T. L. (2006). Probabilistic Topic Models. In T. Landauer, D. McNamara, S. Dennis and W. Kintsch (Eds.), Latent Semantic Analysis: A Road to Meaning (pp. 427-448). Mahwah, New Jersey: Laurence Erlbaum Associates.

Teh, Y. W., Jordan, M. I., Beal, M. J. and Blei, D. M. (2006). Hierarchical Dirichlet Processes. Journal of the American Statistical Association, 101 (476), 1566-1581.

Wallach, H. M. (2006). Topic Modeling: Beyond Bag-of-Words. Proceedings of the 23rd international conference on Machine learning.

Wang, C., Thiesson, B., Meek, C. and Blei, D. (2009). Markov Topic Models. Paper presented at the The Twelfth International Conference on Artificial Intelligence and Statistics (AISTATS).

Zhu, J., Ahmed, A. and Xing, E. P. (2012). Medlda: Maximum Margin Supervised Topic Models. Journal of Machine Learning Research, 13, 2237-2278.

Appendix

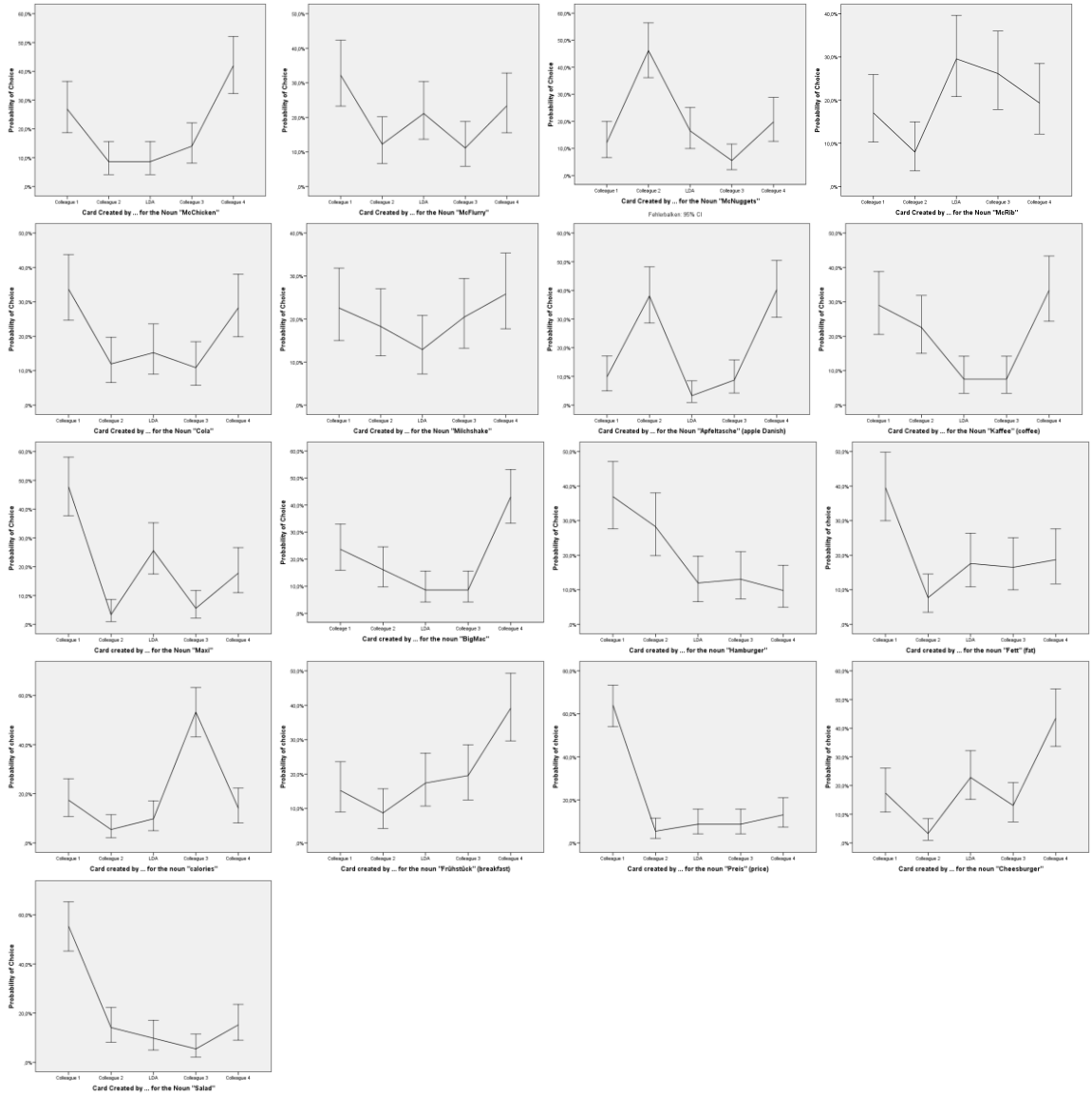

Figure 9. Choice Probabilities of all five cards (Colleague 1, Colleague 2, LDA, Colleague 3, Colleague 4) for all 17 Nouns (McChicken, McFlurry, McNuggets, McRib, Cola (coke), Milchshake (milkshake), Apfeltasche (apple Danish), Kaffee (coffee), Maxi, BigMac, Hamburger, Fett (fat), Kalorien (calories), Frühstück (breakfast), Preis (price), Cheesburger, Salat (salad))