Available online at www.sciencedirect.com

SciVerse ScienceDirect

Procedia - Social and Behavioral Sciences 26 (2011) 63 – 75

Procedia
Social and Behavioral Sciences

Collaborative Innovation Networks (COINs) 2010

Automatic Generation of Product Association Networks Using Latent Dirichlet Allocation

Javier Sanchez-Monzon, Johannes Putzke, Kai Fischbach*

Department of Information Systems and Information Management, University of Cologne, 50937 Cologne, Germany

Elsevier use only: Received date here; revised date here; accepted date here

Abstract

We present a method for extracting semantic networks of words that consumers associate with products and brands, and illustrate the method using reviews of McDonald's products from the opinion platform www.ciao.de as examples. We model the generation of each product review with the probabilistic topic model Latent Dirichlet Allocation (LDA), which enables us to discover the hidden thematic structure of all the reviews in our text collection. We conduct an association analysis of all the words used, revealing the semantic networks of words. Our approach may be highly relevant for marketing managers, for example, as they analyze brand concept maps or seek to optimize ad campaigns with the best words.

Keywords: brand concept maps; collaborative innovation networks; data mining; social networks; webmining; COINs; LDA

1. Introduction

Our main objective is to propose a method to generate automatically a network of words that consumers associate with particular products and brands, by analyzing unstructured text documents from online review platforms such as http://www.ciao.de and http://www.epinions.com. To generate association networks, we use the probabilistic topic model of Latent Dirichlet Allocation (LDA) (Blei, Ng, & Jordan, 2003), which is able to account for the different semantic contexts in which the words are used. We illustrate our approach using some of the most frequently consumed McDonald's products as examples.

The proposed method may be of high relevance for a managerial audience for three main reasons. First, a principal concern of marketing managers is how to position their brands. In this context, brand concept maps (John, Loken, Kim, & Monga, 2006) are an effective tool for brand management. However, it is very costly to create brand concept maps with traditional methods. The proposed method provides a low-cost alternative. Furthermore, our

* Corresponding author. Tel.: +49-221-470-5322; fax: +49-221-470-5393.
E-mail address: sanchez@wim.uni-koeln.de.

doi:10.1016/j.sbspro.2011.10.563

methodology is based on Internet behavioral data and hence the validity bias to which it is subject is different than those for methods that elicit brand associations by questioning respondents in artificial environments. This should provide another perspective on the consumers' brand associations, one in which they express their associations and personal experiences with a brand voluntarily without being asked and knowing the aim of the study (as opposed to being asked specifically to reveal their associations).

Second, marketing managers often are not certain of the best keywords to use in advertising and for search engine optimization. Our method unmasks hidden word associations from the ”long tail” (Anderson, 2006) and hence may help marketing managers lower their costs for ad word campaigns.

Third, scholars and managers who analyze social networks are often interested in identifying Collaborative Innovation Networks (COINs) -- teams of people that share the same vision and collaborate to achieve a common goal by sharing information, ideas, and work, often over the Web (Gloor, 2006). Our method identifies people who are closely linked to each other by analyzing the co-occurrence of their names on Web documents.

The present study is structured as follows: Section 2 highlights probabilistic topic models with a focus on Latent Dirichlet Allocation (LDA). Section 3 illustrates our approach practically, using McDonald’s product reviews as an example. Finally, Section 4 offers a brief summary, a discussion of the theoretical and managerial implications of the research, as well as an outlook to future research.

2. Probabilistic Topic Model of Latent Dirichlet Allocation

This section highlights probabilistic topic models, with a focus on 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 explain how LDA models can be estimated using Markov Chain Monte Carlo Methods.

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* d_1 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* d_1. Topic 1 and Topic 2 have a probability of .25 and .10 of belonging to *document* d_1. Like *document* d_1, 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 1.1, *document* d_1 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. The word with the highest probability in each topic represents the topic’s subject.

Table 1 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 eighteen terms in these three topics.[1]

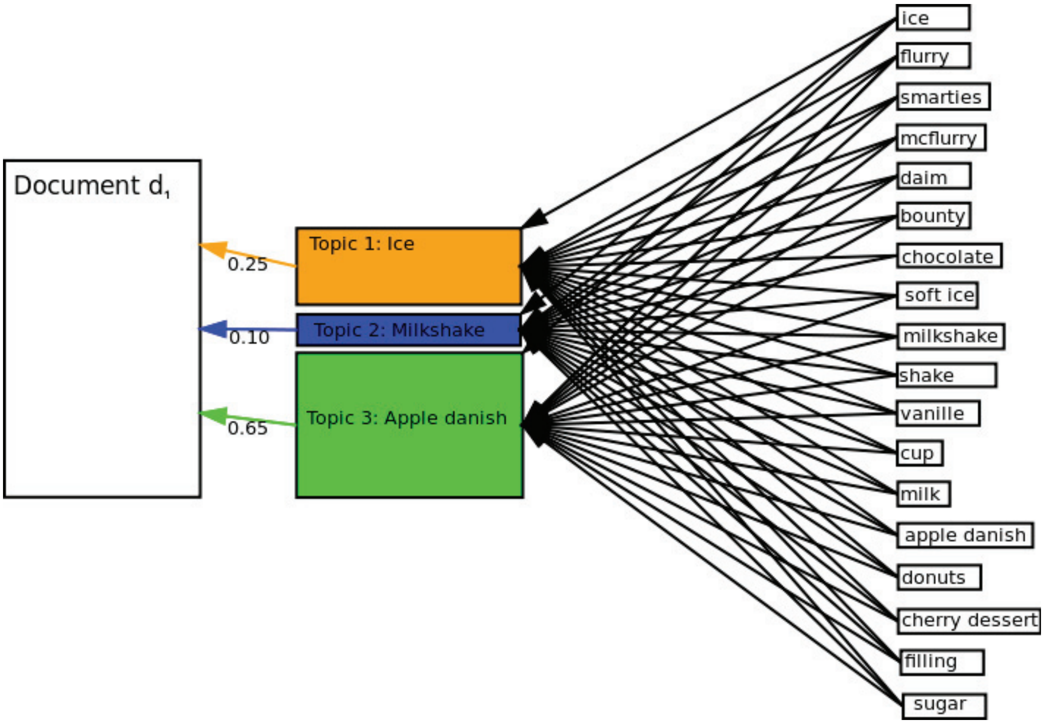

Figure 1. *Document d_1* is based on 3 topics, having each of them a 18 dimensional probability distribution over terms

[1] The reviews on ciao.de were written in German language. Hence, some lines in the table consist of several words (e.g. soft ice cream) since we are not aware of an English translation of the original German phrase (e.g. Softeis) in a single word.

Table 1. Topic Distributions

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	soft ice cream	.0030	soft ice cream	.0500
chocolate	.0400	chocolate	.0400	chocolate	.0400
soft ice cream	.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**

2.2. Mathematical Notation

As Figure 1 and Table 1 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, $P(w_i \mid z_i=j)$. W_i refers to the ith word-token in document d. We can represent all the $P(z=j)$ of a document d with a 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 $\boldsymbol{\theta_{d_1}}$ = (.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 $\square_1$..... $\square_T$. $\square_z$ 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 don't need to 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 $\square_z$ topic distribution over the terms collection such as *ice*, *milkshake* and *apple Danish* will be represented by $\square_1$, $\square_2$ and $\square_3$, respectively. Each $\square_z$ refers to the probability distribution over all the W terms given topic z, also symbolized by $\square_z = P(w|z)$ (Steyvers & Griffiths, 2006).

Generating *document* d_1 from Figure 1 that has high probability to belong to topic 1 and topic 3, suppose that we choose from our given document-topic distribution $\boldsymbol{\theta_{d_1}}$ 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 $\square_z$. 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 $\square_2$. 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

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

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 $\square_z$. 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 $Dir(\alpha_1 \ldots \alpha_D)$ 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 $\square_1 \ldots \square_T$. This a priori knowledge can be understood as knowing in advance what the probability distribution of θ_d and $\square_z$ will look like given the scalar parameters α and β, respectively.

The fact that the probability of occurrence of a word in a document depends on θ_d, $\square_z$, α, and β can be depicted as in Figure 2 or Table 2. The model in Figure 2 depicts the conditional dependency between the random variables (see Steyvers and Griffiths, 2006). The nodes represent the random variables. The plates represent how many replications of a random variable are done, and the edges between the nodes represent the dependencies between them. In the figure, the variable z_d reflects the number of words w in the document d. The occurrence probability of word w in a document d depends, in other words, on the document topic proportion θ_d or, more specifically, on the topic assignation z_d, chosen from topic proportion θ_d. Furthermore, the word w depends also on the word-topic distribution of that chosen topic $\square_z$.

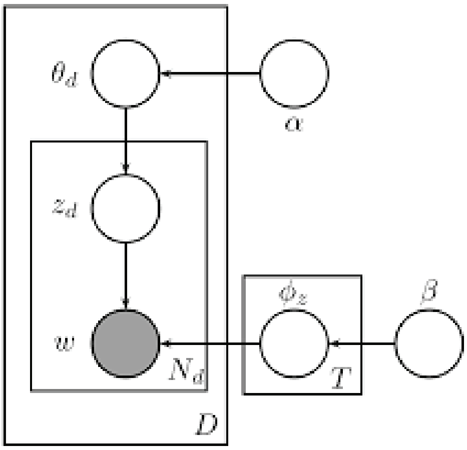

Figure 2. LDA generative process as a graphical model

Table 2. LDA document generation as pseudocode

```
Sample T distributions over terms from the Dirichlet distribution: $□_{1..}$ $□_T \sim Dir_W(\vec{\beta})$ , with parameter $\beta$
For each document d do
Sample a vector of document-topic proportions $\theta_d \sim Dir(\vec{\alpha})$, from a Dirichlet distribution with parameter $\alpha$
  For each word(w) of a document(d) do
   Sample topic index z from the multinomial variable $\theta_d$
   Sample term for word w (word in document d of position i) from the word-topic distribution $□_{z.}$
  EndFor
EndFor
```

2.3. Topic Discovery

The complete generative process described in the last section can 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, we can infer with statistical techniques the topics responsible for generating the text documents. We are interested in the probability distributions $\theta_1 \ldots \theta_D$ *and* $□_1 \ldots □_T$ and in evaluating the posterior distribution

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

This inference problem is resolved using the Gibbs sampling algorithm (Geman & Geman, 1984; Steyvers & Griffiths, 2006), by examining the posterior distribution of the topic assignations and not estimating directly the multinomial random variables: $\theta_1 \ldots \theta_D$ and $□_1 \ldots □_T$. Gibbs Sampling belongs to the set of iterative techniques to sample values from complex distributions, also called Markov Chain Monte Carlo (MCMC) estimation. The algorithm starts by randomizing topic assignations to each of the different words in the collection. Each word-topic assignation is conditioned iteratively on the topic assignations of all other words from the previous iteration. Each of the topic assignation iterations are modeled as states of a Markov Chain. The Gibbs Sampling algorithm begins after an initial period to approximate the desired topic distributions. These steps are then iterated until a steady state is reached.

The *j* topic assignation to each word token w_i on document *d* is approximated iteratively using the Gibbs Sampling algorithm, as described in (Steyvers & Griffiths, 2006), by equation (3). The posterior probability over the topic assignations is aproximated given $z_{-i}, w_i, d_i, \cdot,$ *where* z_{-i} refers to the topic assignments of all other word contained in the document and · represents all other known or observed information such as all the other words, documents, and Dirichlet parameters.

$$P(z_i = j | z_{-i}, w_i, d_i, \cdot) = \frac{P(w,z)}{\sum_z P(w,z)} \propto \frac{C^{WT}_{w_i j} + \beta}{\sum_{w=1}^{W} C^{WT}_{wj} + W\beta} \frac{C^{DT}_{d_i j} + \alpha}{\sum_{t=1}^{T} C^{DT}_{d_i t} + T\alpha} \quad (3)$$

C^{WT} and C^{DT} represent matrices of counts with dimensions *WxT* and *DxT* respectively, where *W* refers to the total number terms of the document collection. C^{WT}_{wi} contains the number of times *j* is assigned to word *w*, not including the current topic-assignation of z_i. Analog C^{DT}_{di} contains the number of times topic *j* is assigned to some word token in document *d*, again not including the actual topic-assignation of z_i (Steyvers & Griffiths, 2006). This means that the full conditional distribution $P(z_i=j|z_{-i},w_i,d_i,.)$ can be calculated by using these count matrices and the Dirichlet parameters only.

The Gibbs Sampling process begin as described above by assigning a random topic number from [*1…T*] to each word token of all the documents. For each word token, the count matrices C^{WT} and C^{DT} are decremented by one for the entries that correspond to the current topic assignment. Then, according to equation (3), a new topic number for the word token is sampled and the count matrices C^{WT} and C^{DT} are incremented with the new topic assignment. After several Gibbs Sampling iterations, we obtain an approximation of the desired values of word-topic (□) and document distributions (θ) from the count matrices, as in equations (4) and (5). The value of $□_{i,j}$ refers to the estimate of word *i* on topic *j*, and $\theta_{d,j}$ refers to the estimate of document *d* belonging to topic *j,* as described in (Steyvers & Griffiths, 2006),

$$\hat{\phi}_{i,j} = \frac{C^{WT}_{ij} + \beta}{\sum_{k=1}^{W} C^{WT}_{kj} + W\beta} \quad (4)$$

$$\hat{\theta}_{d,j} = \frac{C^{DT}_{d,j} + \alpha}{\sum_{k=1}^{T} C^{DT}_{dk} + T\alpha} \quad (5)$$

3. Practical Illustration

We illustrate our approach using 9,529 unstructured and uncategorized McDonald’s product reviews that were crawled from http://www.ciao.de. Each review is represented as an independent document. We conducted our analysis relying on a three-step approach suggested by Fayyad, Piatetsky-Shapiro, and Smyth (1996): (1) **data preprocessing**, which transforms the raw source data of the web documents into an appropriate form for the subsequent analysis; (2) **data mining,** which transforms the prepared data into latent topics responsible for the generation of all the documents of our datasets and subsequent product association analysis; and (3) **postprocessing of data mining results**, which assesses the validity and usefulness of the latent topics and association analysis.

3.1. Data Preprocessing

As a first preprocessing step, we removed all html tags from the documents (see Figure 3). We then reduced the vocabulary size of our dataset as follows. First, we used a stop word list to filter out words such as pronouns (personal, possesive, reflexive, indefinite, relative and interrogative), because these words are not representative for a document (and would hamper the analyses). Second, we used a part-of-speech tagger[2] to extract nouns from the documents, on the assumption that nouns capture product names with a high probability. The resulting vocabulary contains 35,105 terms.

[2] “TreeTagger”, http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger, accessed 12/14/2010.

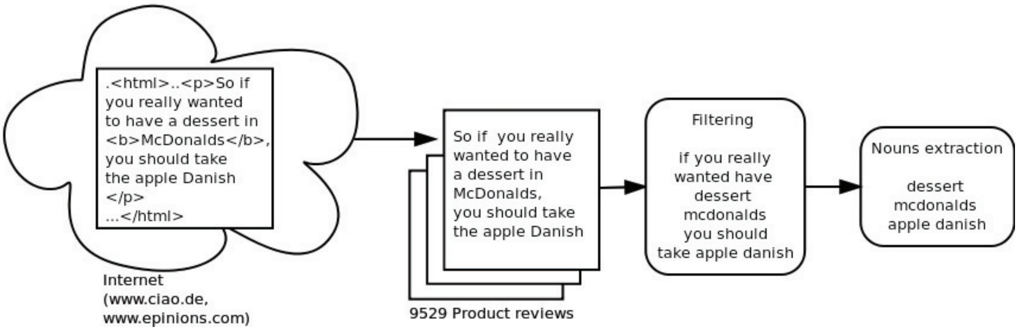

Figure 3. Text Preprocessing

Table 3: Topic Distribution Extraction on 3/50 Topics

Topic 7	{**eis**, **flurry**, **smarties**, **mcflurry**, daim, bounty, softeis, becher (cup), nuts, sorten(types), löffel (spoon), geschmacksrichtungen (tastes), vanilleeis, su ndae, mischung (mixture), sommer (summer), favorit, waffel (waffer), **schokolade**, eisdiele}
Topic 10	{**milchshake**, **shake**, vanille, geschmack (taste), schoko, **becher** (cup), milch (milk), sommer (summer), milchshakes, strohhalm (straw), erdbeer (strawberry), shakes, erdbeere, liter, deckel (cap) pappbecher (paper cup), schön (nice), zucker (sugar), pina (pineapple), colada, **schokolade**}
Topic 30	{**apfeltasche** (apple danish), **donuts**, **donut**, kirschtasche (cherry dessert), füllung (filling), zunge (tongue), inhalt (content), zucker (sugar), nachtisch (dessert), schön (nice), tasche (bag), teig (dough), apfeltaschen, **schokolade**, vorsicht (attention)}

3.2. Data Mining

3.2.1. Topic Distribution Extraction

The methods illustrated in section 2.3 were then used to extract a topic distribution of 50 topics using the JAVA packet MALLET (McCallum, 2002) (see Table 3 for three exemplary topics). Using the Gibbs sampling algorithm, we sampled a topic distribution for each document 1,000 times and obtained an approximative steady state. The $Dir(\alpha_1...\alpha_D)$ and $\mathrm{Dir}(\beta_1...\beta_T)$ Dirichlet distributions turn out to have single values for α and β. Good choices from previous research (Steyvers & Griffiths, 2006) for these parameters are $\alpha = 50/numberoftopics$ and $\beta = 0.01$. The same α value means that, in our model, each document has equal opportunity to belong to any topic by each sample. [3] In the same way, the single β value for $Dir(\beta_1...\beta_T)$ can be interpreted as having the same chance in advance for all the topics to be assigned to any word.

3.2.2. Product Association Analysis

Based on the 50 discovered latent topics and the 35,105 terms described in the last subsection, we computed the similarity among each pair of words present in the document collection and ranked them according to the similarity weight as follows: given a pair of words: (w_1, w_2), the more joint probability of words w_1 and w_2 to be in the same topics, the more similar they should be. That is:

$$P(w_2|w_1) = \sum_{i=1}^{T} P(w_2|z=j)\,P(z=j|w_1) \quad (6)$$

In Figure 4, we illustrate the 9 words with the highest joint probability to be in the same topics with the word *Schokolade*. The results in Figure 4 reflect those in Table 1. The three topics in Table 1 refer to three different semantic contexts in which the word *Schokolade* has been used. The first topic corresponds to McDonald's *ice* products, the second to *milkshake*-related words, and the third topic refers to *apple danish*.

[3] The reviews on ciao.de were written in German language. Hence, some lines in the table consist of several words (e.g. soft ice cream) since we are not aware of an English translation of the original German phrase (e.g. Softeis) in a single word.

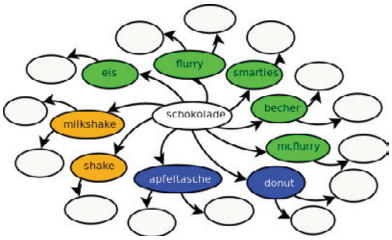

Figure 4. *Schokolade*'s probability semantic word network. The word *Schokolade* has a high probability of appearing in three topics. In this case, the topic number s7, 10, and 30 are represented by the colors green, orange, and blue, respectively.

4. Summary and Discussion

In this paper, we described a method for extracting networks of words that consumers associate with products and brands, and illustrated the method using reviews of McDonald's products from the opinion platform www.ciao.de as an example. In doing so, our method automatically accounts for the different semantic contexts in which the words have been used. The results of this study are highly relevant both for theory and practically for managers.

First, consumers increasingly use blogs, fora, and social network platforms to express opinions about products and brands. Marketing managers must harvest, analyze, understand, and manage this information. An automatic analysis of what a consumer associates with a particular product can be useful for the design of marketing campaigns for several reasons. For example, marketing managers can automatically create brand concept maps (John et al., 2006) from these data. Furthermore, they can extract ”long tail” keywords for the optimization of search engine ad word campaigns.

Second, the LDA model provides an unsupervised clustering method in which no input has an *a priori* categorization. It relies solely on the number of different semantic contexts we seek to discover, the words in the documents, and the Dirichlet parameters. By varying the last parameters, we can ensure the representation of documents or words using fewer or more topics.

Third, the proposed method can also be used to analyze text documents that contain the names of social entities such as organizations and persons. In such types of analyses, we can infer social networks of people and organizations that co-occur in unstructured text documents. This may help identify Collaborative Innovation Networks (COINs), as proposed by (Gloor, 2006). In our future research, we intend to conduct a study to identify COINs using our proposed method. Furthermore, we intend to include in our analyses adjectives, verbs, and adverbs from the text collection to capture sentiments (compare, for example, Liu, 2007).

Finally we intend to evaluate the results of our study as follows. We already randomly selected 20 nouns from a pool of words people associate with McDonald's products (see Table 4) and then asked four of our colleagues who were not aware of the research project to suggest three words they associate with each of these products. In our future work, we will hand a deck of five cards to a large number of students. Each of four cards will contain the three words provided by one of our four colleagues; the remaining card will contain the three words our computer system classified as being closest to one of the terms highlighted in Table 4. The students will be asked to guess which card was created by the IT system. We will analyze the results of this study using standard techniques of analysis of variance and from Information Retrieval (e.g. Baza-Yates & Ribeiro-Neto, 1999).

Table 4. Word evaluation list

Cheesburger, Hamburger, Bigmac, RoyalTs, Salat, Fett, Kalorien, Gesundheit (health), Milchshake, Frühstück (breakfast), Preis (price), McChicken, McFlurry, McNuggets, McRib, Cola, Apfeltasche (apple danish), Chefsalat, Kaffee, Maxi}

References

Anderson, C. (2006). *Long Tail, The, Revised and Updated Edition: Why the Future of Business is Selling Less of More*. New York: Hyperion.

Baeza-Yates, R., & Ribeiro-Neto, B. (1999). *Modern Information Retrieval*. Harlow, England: Addison-Wesley.

Blei, D.M, Ng, A.Y., & Jordan, M.I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, 3, 993-1022

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). Knowledge discovery and data mining: Towards a unifying framework. In U.M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, & R. Uthurusamy (Eds.) *Advances in knolwedge discovery and data mining* (pp. 1-36), pages 1–36. Cambridge, MA: AAI/MIT press.

Geman, S., & Geman, D. (1984). Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images. *IEEE Trans. Pattern Anal. Mach. Intell*, 6, 721-741.

Gloor, P.A. *Swarm Creativity: Competitive Advantage Through Collaborative Innovation Networks*.News York: Oxford University Press.

Griffiths, T. L., & Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National Academy of Sciences of the United States of America (PNAS)*, 101(1), 5228-5235.

John, D.R., Loken, B., Kim, K., & Monga, A. B. (2006). Brand concept maps: A methodology for identifying brand association networks. *Journal of Marketing Research*, 43(November), 549–563.

Liu, B. (2007). *Web Data Mining*. 2007. Berlin: Springer.

McCallum, A. K. (2002). Mallet: A machine learning for language toolkit. http://mallet.cs.umass.edu/.

Steyvers, M., & Griffiths, T. (2006) Probabilistic topic models. In T. Landauer, D. Mcnamara, S. Dennis, & W. Kintsch (Eds.), *Latent Semantic Analysis: A Road to Meaning* (pp.427-448). Mahwah, New Jersey: Laurence Erlbaum Associates.