Applicability of Latent Dirichlet Allocation for Company Modeling

Katsiaryna Mirylenka, Christoph Miksovic, and Paolo Scotton

IBM Research – Zurich,
Säumerstrasse 4, 8803 Rüschlikon, Switzerland
{kmi,cmi,psc}@zurich.ibm.com

Abstract. Nowadays enterprises rely on ever growing amounts of data describing the products used by their customers. The set of these products forms a so-called install base of a company. The analysis and modeling of such install base data essential to the sale forces reveal latent connections between customers and deployed products. This allows to identify discriminative features of the install base structure, which can be used to efficiently compare companies, apply targeted marketing strategies towards similar companies and recommend future products.
In this paper we study the applicability of a topic modeling technique emerging in natural language processing, namely Latent Dirichlet Allocation, to the task of company-product modeling. We formulate an approach of the best model selection using perplexity and silhouette scores on a corpus of more than 400 customer companies belonging to one industry. The results of this study demonstrate that Latent Dirichlet Allocation fits install base well in terms of both goodness of fit of the model and quality of company clusters.

1 Introduction

Marketing data about potential clients provides a unique opportunity to get market insights such as, for example, new business developments or white space[1] determination. This data typically contains various information about companies, such as the products or services that company sells or buys or insights about its internal structure. In this work, we focus on a specific type of marketing data, namely, the information technology (IT) install base[2]. This data contains information about the type of IT equipment of a company and how this equipment is distributed across its subsidiaries. It also contains estimates of the company equipment age and the confidence of its presence.

From the viewpoint of a hardware services provider, the install base information is especially useful for detecting white spaces, as it contains knowledge about the potential of companies with whom it does not yet have a business

[1] A white space represents a potential new business opportunity in terms of customer and/or product/service.
[2] Install base refers to the IT inventory of a company.

relationship. Moreover, when combined with data about established customers, competitive install base information can be used to identify companies similar to existing customers. These similar companies have an even higher potential of becoming new customers.

In this paper, we concentrate on the problem of identifying similar companies. A naive comparison of the individual product types owned by companies reveals a strong bias towards products that are common to a large number of companies. Therefore, after reviewing the related work, we focus on the Latent Dirichlet Allocation (LDA) approach that provides meaningful and interpretable company features, which are then tested for the tasks of company clustering and modeling.

2 Preliminaries

In this work, we use the install base information provided by HG Data Company, Inc. [2], which describes IT products deployed at each site of a company. Products are categorized in a hierarchical fashion. In this work, we use the "product category" level provided by HG Data Company, Inc., which consists of 91 distinct categories. Examples of the product categories are "Printers" or "Midrange Computers". Owing to the nature of our application, we restrict our study to 23 categories related to hardware and low-level hardware management software.

These categories constitute company attributes that have been used to create our corpus which is a binary company-attribute matrix. As we do not have the information about the number of products present at each company site, we consider only binary values in the company-product matrix, where 1 means that the product belongs to the install base and 0 means that it does not.

More formally, let us consider a set of N companies $C = \{c_0, \dots, c_{N-1}\}$ represented in the HG Data Company database. Each company c_i has a given set of products A_i in its install base belonging to k categories, which can also be called attributes. This set of attributes is included in the set of all possible attributes $A = \{a_0, \dots, a_{M-1}\}$ containing M elements. That is: $\forall c_i \in C\ ;\ \ c_i \longmapsto A_i = \{a_{i_0}, \dots, a_{i_{k-1}}\} \subset A$. The information about the attributes or products of A can be re-written using vectors $\mathscr{A}_i$ instead of sets A_i: $\forall c_i \in C\ ;\ \ c_i \longmapsto \mathscr{A}_i\ ,\ \ \dim(\mathscr{A}_i) = M\ ,\ \ \mathscr{A}_i = \left[\mathbb{1}_{a_0 \in A_i}, \dots, \mathbb{1}_{a_{M-1} \in A_i}\right]$.

To compare company attributes, we could immediately introduce a distance in this initial attribute space. In Section 5 we demonstrate that the initial attribute space does not discriminate the companies well enough. To overcome this obstacle, our goal is to introduce a new space of features that better represents the IT install base of a company: $\forall c_i \in C\ ;\ \ c_i \longmapsto \mathscr{B}_i \in \mathbb{R}^L\ ,\ \ L < M$.

Given this formalization, the goal of this paper is to discover the most representative features $\mathscr{B}$ of a company based on the initial company attributes A_i. The features should be representative in terms of goodness of fit of a generative model of company-product data and in terms of quality of company clusters.

3 Related work

In recent years, a lot of research has been devoted to the advancement and improvement of topic modeling methods; one of these techniques is LDA [1]. It is a generative probabilistic model for collections of discrete data, such as corpora of documents, that can be used also for collaborative filtering. Each item (document) is modeled as a finite mixture of an underlying set of topics (hidden groups). The vector of topic probabilities for an item provides an explicit representation of a document. The main advantage of LDA over other similar techniques, such as Latent Semantic Indexing, is the fact that LDA-learned features are easily interpretable.

The company-product modeling we are interested in consists of a dimension for companies and a dimension for product categories (or products). We assume that these dimensions can be mapped to Natural Language Processing (NLP) concepts for the application of LDA. In our case, companies refer to documents and product categories refer to words. The vocabulary of the words consists of the product categories in our scope. All companies from the HG Data Company database form the corpus of "company documents". We further assume that products can constitute hidden topics, which eventually construct specific and discriminative features of a company.

4 LDA Adaptation and Parameter Estimation

We train LDA both on initial binary company-product representations[3] and TF-IDF[4] representations. The type of data representation is considered as one of the parameters for LDA training. Although LDA intrinsically models data in a way that gives more weight to the most representative features, we verify whether the model improves if TF-IDF representations are given as input. Another parameter is the number of LDA latent topics.

We choose the parameters of LDA by minimizing the perplexity level of the model, which is a goodness-of-fit measure. The average per product perplexity is calculated on a test set. Perplexity shows how well the probability distribution defined by LDA, $P(\cdot)$, predicts test data, and is calculated as follows: $Perplexity = 2^{-\frac{1}{T}\sum_{i=1}^{T}\log_2 P(a_i)}$, where T is the number of products in the test set. The lower the perplexity, the better the model.

To examine how well extracted features perform in comparison with initial binary – or TF-IDF – features, we assess the quality of representations for a clustering task.

As the measure of clustering quality we use the silhouette score[5]. It is calculated as the ratio of intraclass and interclass distances. The higher the score, the better the clusters are separated from each other.

[3] Initial binary representation can also be called 'Bag Of Words' (BOW) in NLP terms.

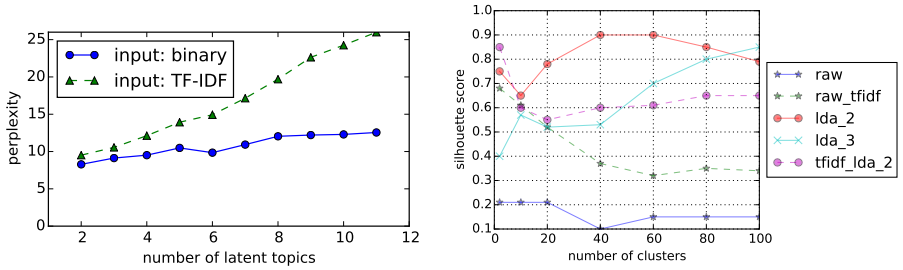
[4] Term Frequency-Inverse Document Frequency [6], in our case can be also reformulated as product frequency-inverse company frequency.

[5] For our experiments we use the silhouette score implementation from *sklearn*[5].

5 Experimental evaluation

The experiments are done for more than 400 aggregated companies. First, we estimate the perplexity of the initial $\mathscr{A}$ company representations. This is equivalent to the perplexity of the unigram BOW model. This perplexity is equal to 21.84, which is the baseline value for further experiments.

We estimate perplexity values of the LDA model for different parameter values, and use the LDA implementation from *gensim* package [7]. The perplexity curves of LDA with both inputs are shown in Figure 1a. As can be seen from



(a) Perplexity plots. (b) Silhouette curves.

Fig. 1: Experimental results.

the plot, the perplexity of LDA from raw binary input is better than that from TF-IDF preprocessed input. Hence, LDA indeed assigns higher weights to the most representative products, and thus no additional TF-IDF transformation of the input is needed. Lower numbers of hidden topics, namely 2 and 3, lead to the lowest perplexity scores, i.e., 8.28 and 9.12 for raw binary inputs.

Our next step is to see how well clustering can separate companies given the most discriminative company representations obtained with 2 and 3 hidden topics. For this purpose, we build silhouette curves for these LDA architectures. We also compare them with silhouette curves of clusters built given (a) raw BOW representations, (b) raw TF-IDF company representations and (c) LDA-based representations with TF-IDF input for 2 hidden topics (Figure 1b).

Higher scores mean that the distances between companies within one cluster are much smaller than distances between companies in different clusters. Figure 1b shows that the initial binary representation of companies is not very discriminative (blue line with stars) as the silhouette score is the lowest for almost all numbers of clusters. The initial representation with TF-IDF transformation leads to better clusters, as the silhouette curve is higher, reaching values greater to 0.3 for different numbers of clusters. LDA representations with TF-IDF input (tfidf_lda_2) perform better than raw TF-IDF. Company representations that produce the best silhouette curves are representations derived from LDA with

raw binary inputs for the number of latent topics equal to 2 and 3. These results are in line with perplexity results, meaning that LDA with these number of topics represent install bases of companies the best.

Below we describe the main contributing products to the LDA2 topics. The weights correspond to the probability of a product to belong to a certain topic.

- **Topic 1:** 0.095*Virtualization: Platform Management + 0.083*Consumer Electronics, Personal Computers & Software + 0.079*Virtualization: Application & Desktop + 0.073*Data Archiving, Back-Up & Recovery + 0.069*Network Management (Hardware) + 0.064*Server Technologies (Software) + 0.058*Communications Technology + 0.057*Virtualization: Server & Data Center + 0.057*Hypervisor + 0.053*Data Management & Storage (Hardware)
- **Topic 2:** 0.445*Database Management Software + 0.445*Operating Systems & Computing Languages + 0.067*Server Technologies (Software)

The products are well separated by the topics. The first topic contains 10 product categories with almost equal weights, whereas in the second topic two categories strongly dominate ("Database Management Software" and "Operating Systems & Computing Languages"). The topics clearly represent two different latent structures in the install base of a company. These results motivate us to continue using LDA for company-product modeling.

6 Conclusions and Future Work

In this work, we assessed the performance of LDA for modeling the install bases of companies. By assuming intrinsic hierarchies between products, companies and possibly latent internal structures, we have applied LDA and it has indeed discovered such hierarchies. We found that LDA with 2 and 3 latent topics fitted our data best.

Motivated by the results, we will model more product-company data using LDA and other techniques that can extract hidden structures in the data, such as Deep Neural Networks. We will also gather install base data with corresponding timestamps and validate the LDA-based features having historical slices of the data. We will also consider time dimension of product appearances in a company and assess the applicability of time series generative models, like Markov chains. This kind of models can be estimated incrementally using conditional heavy hitters [4], [3].

References

1. D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022, 2003.
2. HG Data Company. `http://www.hgdata.com`.
3. K. Mirylenka, G. Cormode, T. Palpanas, and D. Srivastava. Conditional heavy hitters: detecting interesting correlations in data streams. *The VLDB Journal*, 24(3):395–414, 2015.
4. K. Mirylenka, T. Palpanas, G. Cormode, and D. Srivastava. Finding interesting correlations with conditional heavy hitters. In *ICDE 2013*, pages 1069–1080, 2013.
5. F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.
6. A. Rajaraman and J. D. Ullman. Data mining. In *Mining of Massive Datasets*, pages 1–17. Cambridge University Press, 2011.
7. R. Řehůřek and P. Sojka. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta, May 2010. ELRA.