16th Industrial Conference on Data Mining, July 13-17, 2016, New York, USA

Research — Zurich

Applicability of Latent Dirichlet Allocation for Company Modeling

Katsiaryna Mirylenka, Christoph Miksovic and Paolo Scotton
{kmi, cmi, psc}@zurich.ibm.com
IBM Research — Zurich, Säumerstrasse 4, CH-8803 Rüschlikon — Switzerland

1 Motivation

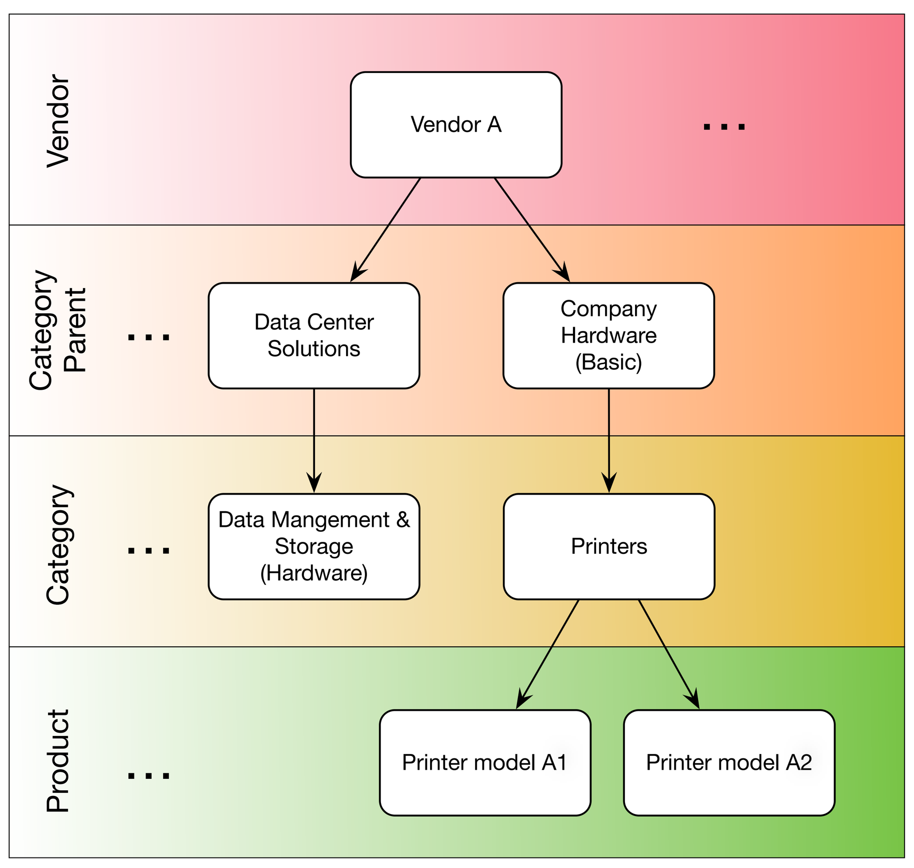
Competitive install base data provides:
- insights about the IT equipment of a company
- equipment distribution across its subsidiaries

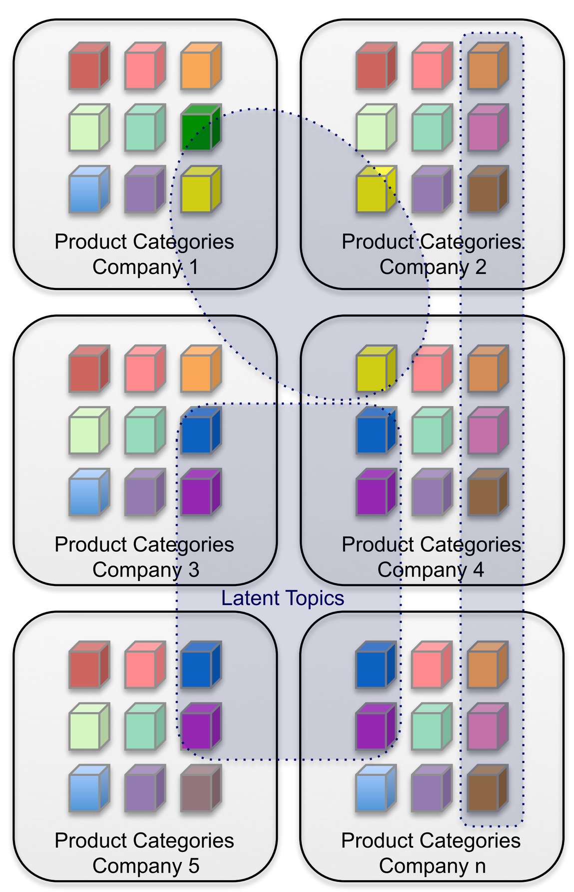
Important for hardware services providers to detect white spaces: contains knowledge about the sales potential for companies with whom it has no business.
Companies *"similar"* to existing customers are of higher interest.

2 Dataset Characterization

Install base information is provided by HG Data Company, Inc.:
- describes IT products deployed at each company site
- products are categorized in a hierarchical fashion

We restrict our study to 23 categories related to hardware and low-level hardware management software.





3 Problem Statement

$C = \{c_0, \ldots, c_{N-1}\}$ is the set of N companies. Each company c_i has a given set of products $A_i \in A$, $|A| = M$, in its install base belonging to k categories:

$$\forall c_i \in C\ ;\ \ c_i \longmapsto A_i = \{a_{i_0}, \ldots, a_{i_{k-1}}\} \subset A.$$

The information about the products from A can be re-written using vectors $\mathscr{A}_i$ instead of sets A_i:

$$\forall c_i \in C\ ;\ \ c_i \longmapsto \mathscr{A}_i\ ,\ \ \dim(\mathscr{A}_i) = M\ ,\ \ \mathscr{A}_i = \left[\mathbb{1}_{a_0 \in A_i}, \ldots, \mathbb{1}_{a_{M-1} \in A_i}\right].$$

Goal: learn the set of most representative features $\mathscr{B}_i$ of a company based on initial company product set A_i such that:

$$\forall c_i \in C\ ;\ \ c_i \longmapsto \mathscr{B}_i \in \mathbb{R}^L\ ,\ \ L < M.$$

Features should be representative in terms of:
- goodness of fit of a generative model of company-product data
- quality of company clusters

4 Solution Approach

Latent Dirichlet Allocation (LDA) is the NLP state-of-the-art technique to find hidden topics in document-word models. We associate companies with documents and products with words and extract hidden topic in company-product space.

5 Experimental Settings

$N = 1319$ companies in the pharma industry.
Goodness-of-fit is measured with the average per product perplexity, $\mathcal{P}$, which shows how well the probability distribution defined by LDA, $P(\cdot)$, predicts test data:

$$\log_2(\mathcal{P}) = -\frac{1}{T}\sum_{i=1}^{T} \log_2 P(a_i),$$

T is the number of products in the test set.
The lower the perplexity, the better the model.
Clustering quality is measured by silhouette score: the ratio of intraclass and interclass distances.

6 Experimental Results

The perplexity of the initial company representations $\mathscr{A}_i$ (equivalent to the unigram BOW model) is equal to 21.84, which is the baseline value for further experiments (Fig. 1).
The clustering performance is assessed for LDA representations with 2 and 3 topics. We also compare them with (a) raw BOW representations, (b) raw TF-IDF company representations and (c) LDA-based representations with TF-IDF input for 2 hidden topics (Fig. 2).

Figure 1: Perplexity Analysis

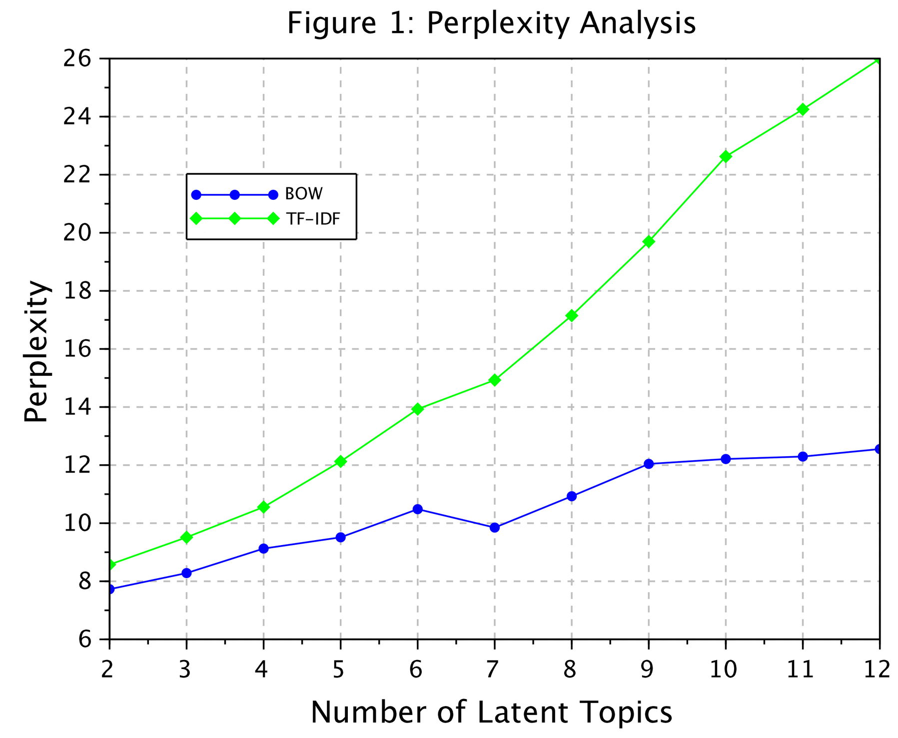
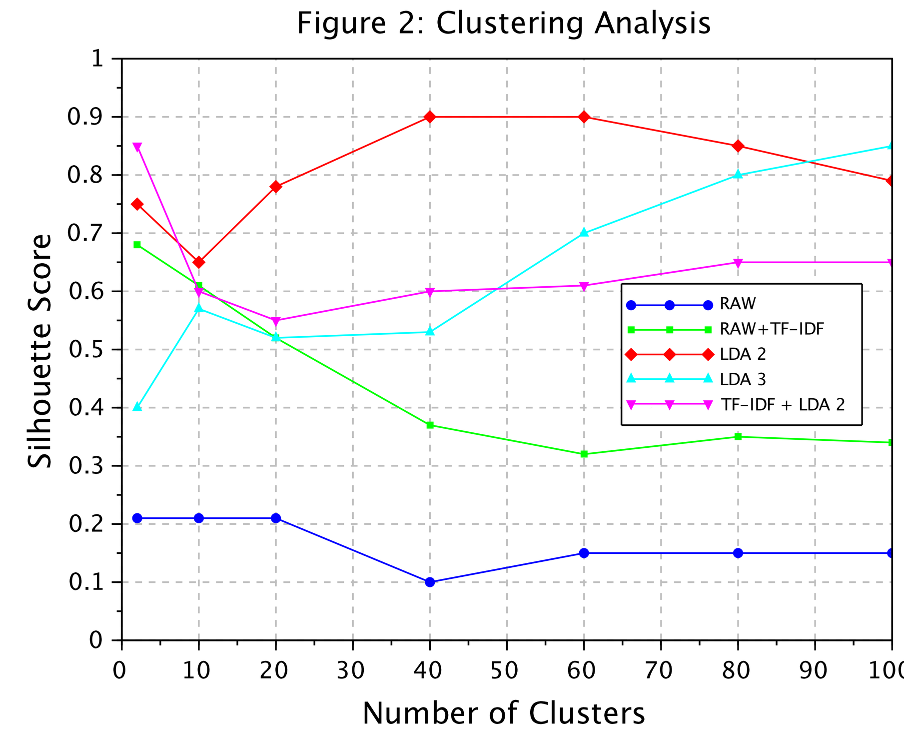


Figure 2: Clustering Analysis



LDA-learned hidden topics:
- **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)

7 Conclusions and Future Work

We found that:
- LDA performs very well for modeling install bases of companies as it reveals intrinsic hierarchies between products and companies
- LDA with 2 and 3 latent topics fitted our data best

Future work:
- comparison with other techniques that can extract hidden structures in the data e.g. Deep Neural Networks
- validation of LDA-based features having historical slices of the data
- investigate the applicability of time series generative models, like Markov chains.