LDA-BASED INDUSTRY CLASSIFICATION

Research-in-Progress

Fang Fang
Department of Information Systems
National University of Singapore
15 Computing Drive, Singapore 117418
fangfang@comp.nus.edu.sg

Kaushik Dutta
Department of Information Systems
National University of Singapore
15 Computing Drive, Singapore 117418
duttak@nus.edu.sg

Anindya Datta
Department of Information Systems
National University of Singapore
15 Computing Drive, Singapore 117418
datta@comp.nus.edu.sg

Abstract

Industry classification is a crucial step for financial analysis. However, existing industry classification schemes have several limitations. In order to overcome these limitations, in this paper, we propose an industry classification methodology on the basis of business commonalities using the topic features learned by the Latent Dirichlet Allocation (LDA) from firms' business descriptions. Two types of classification – firm-centric classification and industry-centric classification were explored. Preliminary evaluation results showed the effectiveness of our method.

Keywords: Industry Classification, LDA, Text mining, Peers Identification

Introduction

Industry analysis, which studies a specific branch of manufacturing, service, or trade, is widely used in financial analysis (Davis and Duhaime 1992; Kahle and Walkling 1996). Such analysis is useful by various groups of people: (a) asset managers need industry analysis to investigate the target company's competitive environment and growth opportunities, after which they could perform stock selection and valuation (Bhojraj and Lee 2002); (b) credit analysts need industry analysis to assess the target company's financial status through the comparison of industry average, after which they could rate the company; (c) investors need industry analysis to study the target industry's competitiveness, profitability and growth, after which they could make investment decision; (d) researchers need industry analysis to identify the industry that the target company belongs to, after which they could design appropriate control groups for their studies (Lee et al. 2012).

Before we could perform industry analysis, one crucial step to take is to define industry boundaries effectively and accurately. In other words, we need to assign firms into appropriate industries on the basis of commonalities before any further analysis could be conducted. Otherwise, further industry analysis could become impossible, or at least misleading. Appropriateness and accuracy of industry classification is the premise of an effective and valuable industry analysis.

There exist a number of *Industry Classification* schemes such as the Standard Industrial Classification (SIC)[1] and the North American Industry Classification System (NAICS)[2]. However, these schemes have two major limitations. Firstly, they are all static and assume that the industry structure is stable (Hoberg and Phillips 2013). However, Firms often introduce new products, improve old products and discontinue outdated products, and thus enter and exit various industries. In addition, due to technology innovation, some industries change or even fade out, and new industries appear. Since firms and the market are evolving with the passage of time, an effective industry classification approach should be able to capture the dynamic aspect of the industries. Researchers have started to address this problem through annually updated documents such as financial statements (Chong and Zhu 2012).

Secondly, these schemes assume binary relationship – two firms either in the same industry or from different industries – and do not measure the degree of similarity. This is particularly important when identifying rivals for a target firm. Similarities between firms within the same industry vary a lot and we would like to select the most similar firms as rivals. We believe that an effective industry classification approach should not only be able to identify industries, but also can measure differences within industry, that is, to capture the within industry heterogeneity. In order to overcome this limitation, researchers have started a line of work referred as *Peer Firms Identification*, which aims to identify the most similar firms of the target firm. Data such as input-output (IO) tables (Fan and Lang 2000), 10-K forms (Hoberg and Phillips 2013) and EDGAR[3] search traffics (Lee et al. 2012) were used for this purpose. However, as will be discussed in details in the next section, these work suffer from weaknesses such as failing to consider firms' business scales and inaccurate classification.

In this paper, we propose an industry classification methodology on the basis of business commonalities using the topic features learned by the Latent Dirichlet Allocation (LDA) (Blei et al. 2003) from firms' business descriptions. Unlike most of the existing work, which address either industry classification or peer firms identification, we address them concurrently since we believe they are essentially the same. Industry classification is to group firms with the industry center as the centroid and we refer it as *industry-centric industry classification (ICIC)*; peer firms identification is to group firms with the target firm as the centroid and it is referred as *firm-centric industry classification (FCIC)* in the current work. ICIC is applicable when there is no target firm and we just want to have an overview of the market and industries while FCIC is useful when we have a target firm to study or compare. We represent each firm's business genre by the topic features learned from firms' business descriptions, over which industries are classified and peers are identified. ICIC is achieved through a clustering algorithm and FCIC is accomplished according to the business divergence between firms.

[1] http://www.census.gov/epcd/www/sic.html [Accessed May 1, 2013]
[2] http://www.census.gov/eos/www/naics/ [Accessed May 1, 2013]
[3] http://www.sec.gov/edgar.shtml [Accessed May 1, 2013]

The rest of the paper is organized as follows: we first review related work in literature. Then we provide the intuition and overview followed by an elaboration of our proposed method. Whereafter, we present the results of our preliminary evaluation. Finally, we discuss the future work.

Related Work

Industry Classification

There are a number of industry classification schemes used by practitioners and researchers. Bhojraj et al. (2003) offered a comparison of several major industry classification schemes in a variety of applications in accounting, economics and finance, including the Standard Industrial Classification (SIC), the North American Industry Classification System (NAICS) and the Global Industry Classification Standard (GICS)[4]. We briefly introduce them below.

SIC was established in the United States in 1937 by the Central Statistical Board and classifies industries by a set of four-digit codes. Since the SIC is relatively obsolete, governmental agencies from the U.S., Canada and Mexico jointly developed the NAICS to replace SIC. Though the NAICS has largely replaced the SIC, certain government agencies, such as the U.S. Securities and Exchange Commission (SEC), are still using the SIC codes. Both SIC and NAICS are developed by governmental agencies, which may have little bearing on how investors actually perceive firm similarities (Bhojraj et al. 2003). The GICS, on the contrary, is a collaboration of Standard & Poor's and Morgan Stanley Capital International. It is based on the judgment of a team of financial analysts who read through regulatory filings to determine which firms are financially comparable and has been shown to outperform all other schemes in explaining stock return co-movements (Bhojraj et al. 2003). In the current research, we also rely on the regulatory filings. However, instead of reading them manually, we adopt text analytics techniques to automate the process.

Though quite a number of classification schemes are proposed, they all have the same limitation - they are static and assume that the industry structure is stable. Thus, they cannot capture the dynamic aspect of the industry. Researchers have started to address this problem by using annually updated regulatory filings. Chong and Zhu (2012) attempted industry classification in light of XBRL based financial information collected from the EDGAR. They modeled firms and the GAAP Taxonomy elements used by firms as a bipartite graph and applied a spectral co-clustering approach that simultaneously classified firms and financial statement elements over the network.

Peer Firm Identification

Industry classification we discussed above has one major limitation: it could only present a binary relationship - two firms are either in the same industry or from different industries. In other words, industry classification does not distinguish firms in the same industry. In order to address this limitation, researchers have started their efforts to measure the degree of relatedness between firms.

Fan and Lang (2000) employed commodity flow data from input-output (IO) tables to measure the relatedness based on whether firms share the same inputs and outputs. Their results suggested that the new IO-based measures outperformed traditional measures based on SIC codes. One shortcoming of this method is the necessity for well-specified production process, which is not available for industries such as software. Bhojraj and Lee (2002) developed "warranted multiples"- the future enterprise-value-to-sales and price-to-book ratio - for each firm with guidance from the valuation theory and identified peers as those having the closest warranted multiples. Their experimental results showed the superiority of their proposed method over methods on the basis of other techniques such as industry and size matches.

Ramnathr (2002) defined the peer firms based on analysts' choice of firm coverage. Firms that followed by at least five analysts in common are categorized as peers. The intuition is that brokerage house would assign similar firms for coverage to one analyst for the purpose of minimizing an analyst's information acquisition cost. Franco, Hope, and Larocque (2013) improved this approach by using hand-collected data of peer choice by sell-side equity analysts in their research reports. They found that analysts are more likely to choose peer firms that are similar in size, leverage, etc, and select firms with high valuations.

[4] http://www.msci.com/products/indices/sector/gics/ [Accessed May 1, 2013]

Recently, there is a growing interest in using data from EDGAR of the U.S. Securities and Exchange Commission for industry classification and peer firm identification. Lee et al. (2012) used the Internet traffic patterns from the EDGAR website and an association rules based technique to identify peers. Their intuition is that firms appearing in chronologically adjacent searches by the same individual are fundamentally similar. The experimental results suggested that traffic-based approach outperformed peer firms based on six-digit GICS groupings in explaining variations in base firms' stock returns. However, we found that some peers were clearly misidenfitied. For instance, Microsoft, a software corporation, was identified as a peer of Dow Chemical, chemical corporation.

Hoberg and Phillips (2013) used nouns and proper nouns in 10-K forms' business description section for industry classification and peer firm identification. Specifically, they utilized those words to represent firms and adopted a text clustering algorithm to group firms into industries. In addition, they calculated the cosine similarity between those words of any two firms and selected peer firms using a simple minimum similarity threshold. They showed in the experiment that their text-based approach can explain firm characteristics better than SIC and NACIS. One major drawback of this work is that it failed to consider the business scale – peers should be in comparable business scale.

The current study also contributes to this strand of work. Though we also use 10-K forms downloaded from EDGAR, as will be discussed in the next section, our approach has several key characteristics that make it quite different from Hoberg and Phillips (2013).

Solution Overview

We are interested in categorizing firms into industries based on their commonality of business. At a high level, our method consists of two steps: (1) deriving effective features from text data to represent firms' business; and (2) classifying firms into industries.

In order to represent firms' business, we utilize the "Item 1. Business" section of their 10-K form which is a required filling by the U.S. Securities and Exchange Commission (SEC) and updated annually. It describes the business of the company, i.e., what the company does, what markets it operates in, etc. There are several advantages of using the Item 1 section for business representation. First, the section is updated annually, which enables our industry classification method to capture the evolvement of the firm's business. In addition, it is legally required that firms provide accurate information, which is the premise of high quality industry classification results. Use of 10-K forms restricts the current to focus on public firms only; however, our proposed approach is generic enough to be applied in private firms, given that accurate business descriptions are provided.

We believe that each word in the "Business" section attributes to the corresponding firm's business activities. For instance, if a firm involves in the oil business, words such as "fuel", "refinery", "crude" are very likely to appear in that firm's "Business" section. The Latent Dirichlet Allocation (LDA) (Blei et al. 2003) is adopted to learn those business activities, each of which is referred as a topic and defined as a multinomial distribution over words. Those topic features are formed as vectors to represent firms' business genre.

Topic features offer several benefits over word features used in Hoberg and Phillips (2013). (1) One major issue for text analysis is its high dimensionality which is essentially the number of unique words in a collection and causes the so-called "curse of dimensionality" (Archak et al. 2011; Korn et al. 2001). Use of topic features greatly reduces the data dimensionality and avoids the dimensionality curse. Though Hoberg and Phillips (2013) does not disclose the number of unique words, given the information in the paper and Heaps' Law (Heaps 1978), we can estimate that the number of distinct word features is about 44,000. However, we only use 50 topic features in our experiment. (2) Text data are typically quite sparse – while there are a huge number of potential words, number of words in a document is actually quite small. In Hoberg and Phillips (2013), the average number of words for each firm is only around 175. Use of topic features significantly reduces the data sparsity. According to our experiment, there is no zero-valued feature in firms' business representations. (3) Each topic is a multinomial distribution over words and we can use those probabilities to weigh words with respect to a certain topic. Top weighted words could be used to describe the business activities and furthermore, the industries, which offers natural interpretations of the resulting industries. (4) Topic features enable us to filter out irrelevant content very easily. For instance, in our experiment, we found that there is one topic corresponding to introduction of

firms' management team. That kind of topics should be excluded since they are not related to firms' business. However, Hoberg and Phillips (2013) includes nouns and proper nouns in those sections into firms' business representation, which inevitably bring noises and jeopardize the accuracy of their approach.

After the business representations are constructed, we then classify firms into industries. Two types of industry classifications are proposed: *firm-centric industry classification (FCIC)* and *industry-centric industry classification (ICIC)*. FCIC is useful when there is a target firm to study or compare and ICIC is applicable when there is no focused firm and we just want to have an overview of the industry and the market. FCIC is performed according to two criteria: business genre and scale. We believe that peer firms must have comparable business size. For instance, we have two firms, Microsoft and Tiger Logic, both of which design, develop and sell software products to customers. Although they are engaged in the same business, they are not peers and comparing them is meaningless since their business scale vary too much – Microsoft have a market capitalization of 276 billion USD while Tiger Logic only have 48 million USD. ICIC is accomplished through a clustering algorithm.

Our approach has several key characteristics that make it quite different from Hoberg and Phillips (2013) which also use 10-K forms: (a) we use topic features to represent firms' business. As we have discussed, this offers a number of advantages over the word features used in Hoberg and Phillips (2013). (b) We consider the business scale in addition to the business activities, which is the only criterion considered in Hoberg and Phillips (2013). As we discussed, business scale is an indispensable criterion for peers identification. (c) We use completely different methods for industry classification and peers identification.

Solution Details

In this section, we describe the architecture of our system, and the details of each component in the architecture. We will use the piece of text from Google's Item 1 section of 10-K form "Our business is primarily focused around the following key areas: search, advertising, operating systems and platforms, enterprise and hardware products" as an example for illustrative purpose throughout the rest of the paper.

Architecture

The system architecture of our approach is depicted in Figure 1. The representation construction aims to construct features to represent firms' business genre effectively. It first performs routine text processing and then learns the topic features. After the representations are constructed, firms are classified either in a firm-centric or industry-centric way. We describe each component in Figure 1 in detail below.

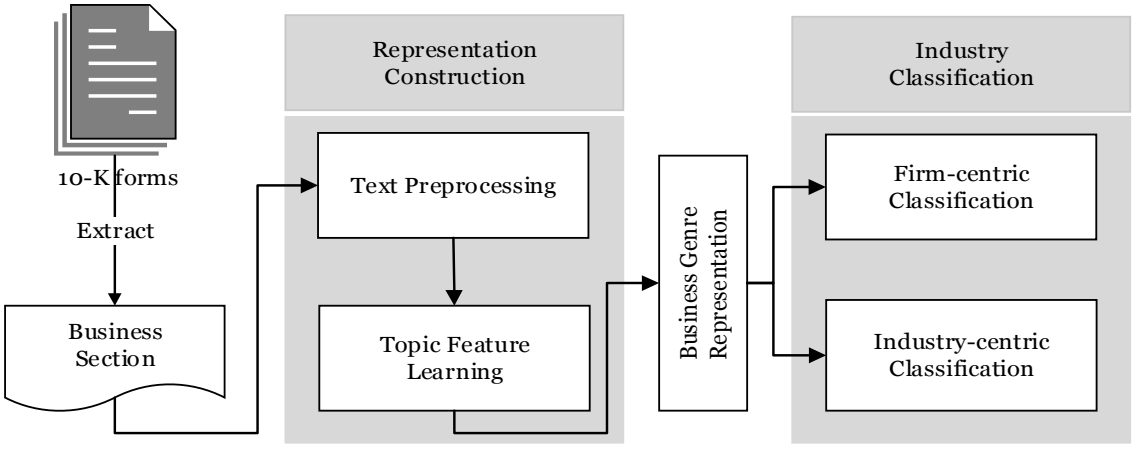

Figure 1. System Architecture

Representation Construction

Text Preprocessing

Before feeding the text data into the LDA for topic feature learning, we first carry out lemmatization on each piece of "Business" section using the Stanford Core Natural Language Processing (NLP) toolkit (Stanford NLP Group 2013). Lemmatization, which transfers inflected forms to base form, or lemma, reduces the sparseness of the data and has been shown to be effective in text related tasks (Joachims 1998). For instance, "says", "said" and "saying" will be all converted into "say". Lemmatization is closely related to stemming. The difference is that stemming operates on a single word without knowledge of the context. For example, the word "meeting" can either be a base form of a noun or an inflected form of a verb. However, lemmatization will determine this based on the contextual Part-of-Speech (POS) information, and thus, we believe it is more appropriate for our current context.

We also remove words that appear very frequent. This includes those typical stop words such as "a", "do", "be", which are not semantically informative. In addition, we also exclude common words that are used by more than 50% of all firms. We believe those common words carry little industry-specific information. After this step, we acquire a set of words that describe the business of a particular firm for topic feature construction.

Following the example in consideration, we would have "focused", "search", "advertising", "operating", "systems", "platforms", "enterprise" and "hardware" after this preprocessing step.

Topic Feature Learning

The Latent Dirichlet Allocation (LDA) (Blei et al. 2003) is selected for topic features extraction. LDA is a three-level hierarchical Bayesian model, which models a document as a finite mixture over a set of underlying topics. A graphical representation of LDA adopted from Blei et al. (2003) is presented in figure 2. W represents a specific word in a document; Z is the topic that generates W; α, β are the parameters of the Dirichlet prior on the per-document topic distribution and per-topic word distribution; θ is the topic distributions for documents and φ is the word distributions for topics; K is the number of topics, N is the number of words in a document and M is the number of documents in a collection.

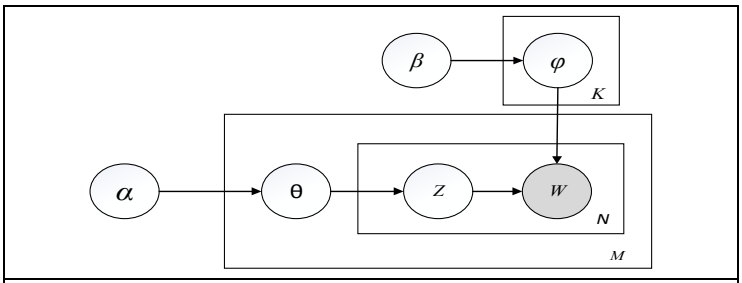

Figure 2. Plate Notation of a smoothed LDA

LDA posits that each word in a document is generated by a topic and each document is a mixture of a finite number of topics. Each topic is represented as a multinomial distribution over words. There are a number of outputs from the LDA. In the current research, we will use two of them: (1) $p(topic_k \mid document_m)$ – the probability of $topic_k$ occurring given $document_m$; (2) $p(word_w \mid topic_k)$ – the probability of $topic_k$ generating $word_w$. The first set of probabilities is used as topic features to represent firms' business genre and the second set of probabilities is used for industry description. In order to train the LDA, we need to specify the number of topics – k. In this paper, we choose topic numbers using the perplexity scores as well as manually interpretation of resulting topics. Typically, the perplexity scores decrease as topic number increases. We choose the number of topics that produces interpretable topics after the reduction of perplexity starts to decrease. The details of the LDA, including model estimation and inference, are beyond the scope of this paper and interested readers can refer to Blei et al. (2003).

We believe that the creation of each word in the "Business" section of a firm's 10-K form is attributable to the firm's business genre. For instance, an oil company would tend to use words such as "fuel", "refinery", "crude", etc; however, "broadband", "wireless" and "subscriber" are more likely to appear in the business

description of a firm in the telecommunication industry. Each business genre can be viewed as a topic, which generates words that constitute the business section according to a certain distribution. Given the words appeared, we can infer the underlying topics, or business genres, that generates the words. The probabilities $p(topic_k \mid document_m)$ are then used as topic feature to represent firms' business. For instance, if we choose the number of topics to be 5, Google's business genre representation learned from the previous piece of text would be the likes of ("0.0079", "0.0031", "0.0143", "0.2523", "0.0015").

Industry Classification

Firm-centric Industry Classification

Firm-centric Industry Classification (FCIC) aims to find comparable firms for the target firm. In this type of classification, each firm has its own set of peer firms which constitute an industry. This is useful when we have a target firm to study or compare. We identify comparable firms considering two criteria. First, peer firms should be engaged in similar business activities. In addition, peer firms should have comparable business scales. As we have discussed previously, if two firms vary too much in business scale, even though in the same business genre, they are hardly peers.

We measure the similarity between two firms in terms of the Kullback–Leibler Divergence (KLD) (Kullback and Leibler 1951) of two firms' business genre representation we constructed in the previous section. KLD is widely used to calculate the divergence between two probability distributions and the KLD of firm F_2 from firm F_1 can be calculated as follows:

$$D_{KL}(F_1 \mid\mid F_2) = \sum_{T_i} F_1(T_i) \times \log \frac{F_1(T_i)}{F_2(T_i)} \quad \text{Eq. 1}$$

where $F_1(T_i)$ and $F_2(T_i)$ are topic features we learned in the previous section. Form the equation we can see that KLD is asymmetrical and thus it not a distant metric. To overcome this, we use the following equation to compute the business genre divergence:

$$D_{genre}(F_1, F_2) = D_{KL}(F_1 \mid\mid F_2) + D_{KL}(F_2 \mid\mid F_1) \quad \text{Eq. 2}$$

where $D_{KL}(F_1 \mid\mid F_2)$ is the KLD of firm F_2 from firm $F_1$and $D_{KL}(F_2 \mid\mid F_1)$ is the KLD of firm F_1 from firm F_2. $D(F_1, F_2)$ measures the divergent between business genres of two firms, and therefore, the smaller the value, the more similar the two firms are. As we discussed previously, peer firms should have comparable business scale. We measure firms' business scales using the market capitalization, which is the total value of the issued shares of a publicly traded company. It can be calculated as follows:

$$market\ cap = share\ price \times number\ of\ outstanding\ shares \quad \text{Eq. 3}$$

Ratio of market cap of two firms is used to measure the business scale comparability. Specifically, we use the following equation:

$$D_{scale}(F_1, F_2) = \log_{10} \max\left(\frac{\text{market cap}_{F_1}}{\text{market cap}_{F_2}}, \frac{\text{market cap}_{F_2}}{\text{market cap}_{F_1}}\right) \quad \text{Eq. 4}$$

$D_{scale}(F_1, F_2)$ would be close to zero if they have similar business scale. Finally, the business divergence of two firms is measured using the following equation:

$$D_{business}(F_1, F_2) = D_{genre}(F_1, F_2) + D_{scale}(F_1, F_2) \quad \text{Eq. 5}$$

We can then rank firms with respect to the target firm according to the business divergence and select top firms with lowest divergence as peer firms that constitute the industry for the target firm.

Industry-centric Industry Classification

In some cases, we might not have a target firm and just want to have an overview of the industry and the market. In order to fulfill this type of needs, we propose the *Industry-centric Industry Classification (ICIC)*, which is analogous to SIC and NACIS. However, our method can capture the evolvement of the industries since our business genre representations are updated annually to represent the current business of firms. In addition, business divergence between any two firms could be easily calculated through equation 2. In other words, our method is able to overcome the two limitations of existing industry classification schemes aforementioned.

Any clustering algorithm, which can group firms into industries, can fulfill this task. In this paper, we select the spectral clustering approach proposed in Ng et al. (2001) due to its good performance in terms of accuracy. One input parameter for almost all clustering algorithms is the number of clusters. We chose the appropriate cluster number by the sum of squared error (SSE), which is defined as the sum of the squared distance between each point of a cluster and its cluster center. Generally, the value of SSE should decrease as the cluster number increases. We select the number where the reduction of SSE slows dramatically as the cluster number since increasing the number of clusters does not have a substantial impact on the SSE.

Preliminary Evaluation

Dataset

We selected the constituents of the S&P Total Market Index (Standard & Poor's 2013) as our firm sample. The index includes all common equities listed on the NYSE (including NYSE Arca), the NYSE Alternext, the NASDAQ Global Select Market, the NASDAQ Global Market and the NASDAQ Capital Market (Standard & Poor's 2013). We acquired the constituent list from Standard & Poor's website, which includes 3756 firms. However, we found that some firms in the list were either delisted or acquired and finally, we have 3734 firms. Then we downloaded 10-K forms filed by those 3734 firms from 2008 to 2012 from the EDGAR database. Please note that for years before 2012, the number of firms is less since some firms have not been made public tradable yet. 10-K forms are reported in nonstandardized HTML files and it is hard to extract information from them (Huang and Li 2011). We found that most of the 10K forms provide links to specific section in the Table of Contents. Taking use of those links as well as the titles of each section in 10K forms, we are able to extract the business section for about 96% of all 10-K forms. Thus, we believe our extraction approach is quite effective. We collected the capitalization information from Yahoo! Finance and used GibbsLDA++ (Phan and Nguyen 2007) to learn the topic features.

Preliminary Results

Examples of our FCIC and ICIC results are presented below. According to the methods discussed in the previous section, we chose the number of clusters to be 60 and number of topics to be 50. For the sake of space, in this paper we are only able to present one industry and peers of one firm.

Table 1. Payment Industry
Visa Inc (V)
Mastercard Inc (MA)
American Express Co (AXP)
Global Payments Inc (GPN)
Discover Financial Services (DFS)

Table 2. Top 10 Peers of Dow Chemical in 2008			
1	DuPont (DD)	6	Ashland Inc (ASH)
2	PPG Industries (PPG)	7	Westlake Chemical (WLK)
3	Eastman Chemical (EMN)	8	Praxair Inc (PX)
4	Albemarle Corp (ALB)	9	Valspar Corp (VAL)
5	FMC Corp (FMC)	10	Rock-Tenn (RKT)

Table 1 presents 5 firms from the payment industry in 2012 classified by our industry classification method. Those 5 firms all clearly belong to the payment industry. Table 2 presents the top 10 peers for Dow Chemical in 2008. We list this for comparison of results in Lee et al. (2012). Dow Chemical is the one of the three largest chemical companies, together with BASF and DuPont. Since BASF is not listed in the U.S. market, it is not in our sample. All 10 firms in Table 2 produce same products as Dow Chemicals such as chemicals, coatings, etc. Compared with Lee et al. (2012), our approach clearly offers much better results. In Lee et al. (2012), companies such as Microsoft and General Electric, which are engaged in quite different business, were identified as peers of the Dow Chemicals. Lee et al. (2012) also identified Rohm and Haas, and Union Carbide as peers; but these two firms are actually subsidiaries of Dow Chemicals.

Expected Contribution and Future Work

In this paper, we proposed a novel approach for industry classification based on the topic features learned by the LDA model. Two types of classification – firm-centric classification and industry-centric classification were explored. Preliminary evaluation results showed the effectiveness of our method.

Our research contributes to the industry classification literature by adding a novel industry classification approach. We introduced the use of topics as features for firm business genre representation, which overcomes the so-called "curse of dimensionality" and sparse data issue. In addition, we considered the business scale as an important factor for firm-centric classification, which avoids identifying two firms with distinct business sizes as peers. Thirdly, our approach take use of the annually updated business description in 10K forms and conduct industry classification every year, which allows to adjust the industries as the firms' business change. Fourthly, our approach is capable of measuring the similarity between any two firms, which captures the within industry heterogeneity. Finally, our preliminary evaluation results showed the effectiveness of our method compared with existing approaches in literature.

Besides contributing to literature, this study also benefits the practitioners. Asset managers could use our approach to investigate the target company's competitive environment and growth opportunities for stock selection and valuation. In addition, credit analysts could use our approach to assess the target company's financial status through the comparison of industry average for company rating. Thirdly, investors could use our approach to study the target industry for investment decision-making. Finally, researchers could use our approach to design appropriate control groups for their studies.

We will complete this work by conducting a comprehensive and rigorous evaluation and demonstrating the effectiveness of the proposed method conclusively. Firstly, following Bhojraj et al. (2003), Lee et al. (2012) and Hoberg and Phillips (2013), we will investigate the extent to which different industry classification methods help to explain co-movements in base firms' various financial ratios. Specifically, we will evaluate how well peer firms identified by various methods explain base firm's return and other financial ratios by comparison of the R^2 values. We will compare our method with SIC, GICS, Lee et al. (2012) and Hoberg and Phillips (2013). This will demonstrate the superiority of our approach. In addition, we will conduct robustness check for the peer size when evaluating how peer firms explain return co-movements.

Furthermore, we will investigate how peer firms of a base firm evolve with the passage of time. We are going to select a number of firms and see how their peers change from 2008 to 2012. This will prove the necessity of a dynamic industry classification method that can capture the evolvement of firms' business. In addition, we will also show how certain industries are evolving and why capturing it is important to business. Finally, in addition to U.S. market, we would like to apply our approach to other markets.

References

Archak, N., Ghose, A., and Ipeirotis, P. G. 2011. "Deriving the Pricing Power of Product Features by Mining Consumer Reviews," *Management Science* (57:8), pp. 1485–1509.

Bhojraj, S., and Lee, C. M. C. 2002. "Who Is My Peer? A Valuation-Based Approach to the Selection of Comparable Firms," *Journal of Accounting Research* (40:2), pp. 407–439.

Bhojraj, S., Lee, C., and Oler, D. 2003. "What's My Line? A Comparison of Industry Classification Schemes for Capital Market Research," *Journal of Accounting Research* (41:5), pp. 745–774.

Blei, D. M., Ng, A. Y., and Jordan, M. I. 2003. "Latent Dirichlet Allocation," *Journal of Machine Learning Research* (3:1), pp. 993–1022.

Chong, D., and Zhu, H. 2012. "Firm Clustering based on Financial Statements," in *In proceedings of 22nd Annual Workshop on Information Technologies and Systems (WITS)*, Orlando, Florida, USA, pp. 43–48.

Davis, R., and Duhaime, I. 1992. "Diversification, Vertical Integration, and Industry Analysis: New Perspectives and Measurement," *Strategic Management Journal* (13:7), pp. 511–524.

Fan, J. P. H., and Lang, L. H. P. 2000. "The Measurement of Relatedness: An Application to Corporate Diversification," *The Journal of Business* (73:4), pp. 629–660.

Franco, G. De, Hope, O.-K., and Larocque, S. 2013. "Analysts' Choice of Peer Companies," http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2046396, [Accessed March 1, 2013].

Heaps, H. S. 1978. *Information Retrieval: Computational and Theoretical Aspects*, Academic Press, pp. 206–208.

Hoberg, G., and Phillips, G. M. 2013. "Text-Based Network Industries and Endogenous Product Differentiation," http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1520062, [Accessed March 1, 2013].

Huang, K.-W., and Li, Z. 2011. "A Multi-Label Text Classification Algorithm for Labeling Risk Factors in SEC Form 10-K," *ACM Transactions on Management Information Systems* (2:3), pp. 18:1–18:19.

Joachims, T. 1998. "Text Categorization with Support Vector Machines: Learning with Many Relevant Features," in *Proceedings of the 10th European Conference on Machine Learning (ECML' 98)*, Chemnitz, Germany, pp. 137–142.

Kahle, K. M., and Walkling, R. A. 1996. "The Impact of Industry Classifications on Financial Research," *Journal of Financial and Quantitative Analysis* (31:3), pp. 309–335.

Korn, F., Pagel, U., and Falutsos, C. 2001. "On the 'Dimensionality Curse' and the 'Self-Similarity Blessing'," *IEEE Transactions on Knowledge and Data Engineering* (13:1), pp. 96–111.

Kullback, S., and Leibler, R. 1951. "On Information and Sufficiency," *Annals of Mathematical Statistics* (22:1), pp. 79–86.

Lee, C. M. C., Ma, P., and Wang, C. C. Y. 2012. "Identifying Peer Firms: Evidence from EDGAR Search Traffic," http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2171497, [Accessed March 1, 2013].

Ng, A., Jordan, M., and Weiss, Y. 2001. "On Spectral Clustering: Analysis and an algorithm," in *In proceedings of 15th Annual Conference on Neural Information Processing Systems*, , pp. 849–856.

Phan, X.-H., and Nguyen, C.-T. 2007. "GibbsLDA++: A C/C++ Implementation of Latent Dirichlet Allocation," http://gibbslda.sourceforge.net/, [Accessed August 10, 2013].

Ramnathr, S. 2002. "Investor and Analyst Reactions to Earnings Announcements of Related Firms: An Empirical Analysis," *Journal of Accounting Research* (40:5), pp. 1351–1376.

Standard & Poor's. 2013. "The S&P Total Market Index," http://us.spindices.com/indices/equity/sp-total-market-index-tmi, [Accessed May 1, 2013].

Stanford NLP Group. 2013. "CoreNLP Tookit," Http://nlp.stanford.edu/software/corenlp.shtml, [Accessed May 1, 2013].