Industrial Engineering
& Management Systems
Vol 13, No 4, December 2014, pp.432-441
ISSN 1598-7248 │ EISSN 2234-6473 │

http://dx.doi.org/10.7232/iems.2014.13.4.432
© 2014 KIIE

Business Model Mining: Analyzing a Firm's Business Model with Text Mining of Annual Report

Jihwan Lee, Yoo S. Hong*
Department of Industrial Engineering, Seoul National University, Seoul, Korea

(Received: November 5, 2014 / Revised: November 18, 2014 / Accepted: November 19, 2014)

ABSTRACT

As the business model is receiving considerable attention these days, the ability to collect business model related information has become essential requirement for a company. The annual report is one of the most important external documents which contain crucial information about the company's business model. By investigating business descriptions and their future strategies within the annual report, we can easily analyze a company's business model. However, given the sheer volume of the data, which is usually over a hundred pages, it is not practical to depend only on manual extraction. The purpose of this study is to complement the manual extraction process by using text mining techniques. In this study, the text mining technique is applied in business model concept extraction and business model evolution analysis. By concept, we mean the overview of a company's business model within a specific year, and, by evolution, we mean temporal changes in the business model concept over time. The efficiency and effectiveness of our methodology is illustrated by a case example of three companies in the US video rental industry.

Keywords: Business Model, Text Mining, Keyword Extraction, 10-K Annual Report, Business Model Concept, Business Model Evolution

* Corresponding Author, E-mail: yhong@snu.ac.kr

1. INTRODUCTION

Market globalization and the rising expectations of customers have drawn companies into fierce competition. Business model innovation is one of the most prominent strategies for meeting this challenge. By doing its business differently, a company could find novel revenue streams and reduce their cost (Johnson *et al.*, 2008). The business model is a logical description about how a company does its business with its product and service (Magretta, 2002). It consists of interrelated plans for capturing the potential values from the company's offering and converting them into economic values. As the business model is receiving considerable attention these days (Pohle and Chapman, 2006), the ability to collect business model related information has become essential requirement for a company. Information about the competitor's business model or the emerging trend could help a decision maker to better plan their business model. Fortunately, we could easily extract such information from various types of documents including blogs, database, electronic newspapers, or even business model patent. In this sense, what is important is mining valuable knowledge from the flood of information.

Among those documents, we propose to use the 10-K annual report in such knowledge mining process. The form 10-K is an annual report required by the US Securities and Exchange Commission (SEC) who regulates publicly traded companies in the United States to disclose their information on an ongoing basis (SEC, 2012).

Table 1. Business model related sentences from 10-K annual report (Netflix, 2011)

Extracted sentences from Form 10-K annual report	Related business model perspective (Osterwalder and Pigneur, 2010)
Our subscribers can instantly watch **unlimited TV shows and movies streamed over the Internet** to their TVs, computers and mobile devices (from ITEM 1.)	Value proposition
We market our service through various channels, **including online advertising, broad-based media, such as television and radio**, as well as various strategic partnerships (from ITEM 1.)	Distribution channel Partner network
We are continuously improving the customer experience, with a focus on expanding our streaming content, enhancing our user interface and extending.	Customer relationship
We obtain content from various **studios** and other **content providers** through streaming **content license agreements**, DVD direct purchases and DVD **revenue sharing agreements** (from ITEM 1.)	Key activities Key resources
The company has three operating segments: Domestic streaming, International streaming and domestic DVD (from ITEM 1.)	Customer segment
We derive our revenues from monthly **subscription fees** and recognize subscription revenues ratably over each subscriber's monthly subscription period (from ITEM 7.)	Revenue streams
Cost of subscription revenues consists of expenses related to the **acquisition** and **licensing** of **content**, as well as content **delivery** costs related to providing streaming content and shipping DVDs to subscribers (from ITEM 7.)	Cost structure

Each report contains comprehensive information about a company's business and financial performance in a fiscal year. Every year, about ten thousand public companies whose assets are above certain class are required to submit the 10-K report. The entire collection of 10-K reports is available to the public throughout EDGAR online database (http://edgar.sec.gov).

The 10-K annual report contains rich information on the company's business model. As an illustrative example, a number of sentences which are sampled from the annual report of Netflix Inc. in 2011 are listed in Table 1 (Netflix, 2011). According to what types of business-model-related information they provide, each sentence is related with the nine business model perspectives which were proposed by Osterwalder and Pigneur (2010). Covering nine perspectives of the business model, those sentences provide the key information for figuring out the company's business model.

The above example shows that 10-K annual report could be a promising source for the business model analysis. However, given the sheer volume of data, which is usually over a hundred pages for each report, it is not practical to depend only on human effort. The purpose of this study is to complement such manual extraction process by using text mining techniques. Text mining is an automated process of extracting valuable knowledge from a collection of unstructured textual data (Delen and Crossland, 2008). This automated process provides a new opportunity for finding uncovered patterns or trends from the huge textual data.

In this study, text mining technique is applied in analyzing the business model concept and its evolution throughout the time period. By concept, we mean overview of a company's business model within a specific year. In order to extract the business model concept, our methodology automatically structures 10-K annual reports and derives a collection of keywords within the structured data. Then, the evolution of the business model is analyzed by investigating time-series patterns of each keyword. With a support of visualization tool, we can diagnose the overview of emerging or obsolete concepts of a company's business model.

To the authors' knowledge, our study is the first attempt to automatically extract business model knowledge from empirical data set. Based on qualified data sources, our methodology provides high quality knowledge which could be incorporated into business model planning. Moreover, the automation of the knowledge extraction process proposed in our methodology may enhance the efficiency of the business model planning.

This paper is organized as follows. In Section 2, we introduce the related literature. In Section 3, we illustrate the general text mining procedures for obtaining keyword vector. Section 4 introduces our methodology for business model concept extraction. In Section 5, methodology for business model evolution analysis is proposed with an illustrative example. Finally, in Section 6, we outline the study's conclusion and discuss future works.

2. BACKGROUND AND RELATED LITERATURE

Whereas the term business model is widely used in practice, research on the business model is still at the beginning stage. Without unified definition or dominant theory, most of research attempt to identify the characteristics of the business model. One of such attempts is to develop taxonomy for existing business model. Timmers (1998)'s taxonomy for internet business or Linder and Cantrell (2000)'s taxonomy for generic industries are such examples. Another attempt is to propose a conceptual model for explaining a business model. These models represent a business model throughout various conceptual schemes, such as a combination of strategy and tactics (Casadesus-Masanell and Ricart, 2010), or a collection of value creation activities (Zott and Amit, 2010).

There are few empirical studies on the business model. Most research in this field tries to figure out relationship between business model types and firms' performances. Investigating revenue segment of several companies, Malone *et al.* (2006) identifies whether certain types of assets or rights sold to the customer is important to the company's performance. Similarly, Zott and Amit (2008) identifies the relationship between company's activity configurations and it's financial performance.

Text mining refers to semi-automated process of extracting valuable knowledge from the text. With the combination of various research fields, such as data mining, computational linguistics, and statistics, it aims to derive knowledge from the text (Miner *et al.*, 2012). Depending on the objective of the analysis, the knowledge could be hypothesis, decision rules, or trends. Although many of text-mining researches focus on developing new text-mining methodologies, such as Mikawa *et al.* (2012), it has also been applied to numerous documents in order to extract interesting patterns from them. For example, Tseng *et al.* (2007) applies text-mining techniques to patent document in order to detect new technology opportunity. Zhang and Jiao (2007) analyzes customer interaction documents in order to extract customization strategy. On the other hands, Cho *et al.* (2014) uses text-mining techniques to journal articles in order to identify the trend of research methodologies applied in industrial engineering domains.

There is some research which applies text mining technique to the 10-K annual report. From economics, Hoberg and Phillips (2010) proposes new industry classification established by a similarity of product description between companies within 10-K reports. From accountings area, Li (2008) investigates the relationship between readability of annual reports and firm's financial performance. From risk managements, Martin and Rice (2007) proposes a case study of automatically profiling risk themes of the company based on the company's disclosure about their risk factors in 10-K reports.

In terms of focusing on the development of practical method for business model planning, our study is distinguished from other conceptual studies on the business model. Moreover, our study is the first attempt to automatically extract business model using 10-K annual report.

3. TEXT MINING PROCEDURE FOR GENERALIZED VECTOR SPACE MODEL

In this study, we use the vector space model which represents a concept of the document as a vector of keywords. Therefore, a business model concept is compacttly represented by a single keywords vector. In this chapter, we propose common standard procedures in generating a keywords vector from the document. Based on Miner *et al.* (2012)'s procedure, we point out some of the important steps which are related to our methodology.

First step is *establishing corpus*. The collection of documents that contains relevant information is referred to the *corpus*. Textual documents from various sources, such as web pages, e-mail, newspapers or research, are collected with the support of automated techniques, such as web crawling program. As a result of this step, the collected documents are transformed into same format that is ready to be used in next steps.

After the corpus is established, *tokenization* is followed. In this step, the text is split into a single word which is referred to token. There are various tokenization methods according to the language being analyzed.

The next step is *stop-words removing*. The stop-word is a word which should be excluded from the analysis since they cannot provide any meaningful information to the analyzer. Such words might be auxiliary verb or definitive article which has less meaning or specific words which are defined according to a user's domain expertise. The stop-word removing, thus, reduces noise of the input data and hence, enables a text miner to obtain more reliable result.

The next step is *stemming*. Stemming is an operation that transforms a word of different part of speech into a single reduced root form (also known as stem). For example, after the stemming operations, the words 'stemming,' 'stemmed,' 'stemmer' are transformed into its root words 'stem.' This operations reduces the dimension of the entire keywords spaces, hence could improve computational efficiency of the text-mining process.

Finally, the keyword vector is generated. The collection of keyword vectors are represented by a term-document matrix. The element within this matrix is represented by a frequency of extracted keyword that occurs in specific documents. Since the term-document matrix contains numerical values, various data mining techniques could be applied to the matrix. Usually, a number of data transformations techniques could be performed to the matrix in order to obtain more reliable result. The most commonly used transformation is *inverse document frequencies transformation.* In this transformation, the number increases by the frequency of the keyword appear in a document, but also offset by the frequency of the keyword appear in entire collection of the document (the corpus). Therefore, we can control the too general keyword which might be less meaningful for the knowledge extraction.

4. OVERALL METHODOLOGY

This chapter describes our methodology in a detailed manner. Our methodology consists of two steps: 1) business model concept extraction and 2) business model evolution analysis. The concept means overview of a company's business model within a specific year, while the evolution means temporal change of the business model concept over time.

4.1 Business Model Concept Extraction

This section describes the detail procedures for extracting a business model concept in a detailed manner. Each step is based on standard text mining procedures described in the previous chapter. In most steps, the practical extraction and report language (PERL) is used as a language for the implementation. All of the PERL modules used in this study can be accessible via 'www.cpan.org.'

4.1.1 Corpus extraction

Every 10-K annual report consists of 15 sub-chapters which begin in item 1 (business) and end in item 15 (exhibits, financial schedule signature). After a thorough investigation of several annual reports, we concluded that item 1 (business description) and item 7 (managerial discussion and analysis) are the most desirable part for the business model analysis. Both parts provided narrative descriptions about their businesses done or intended to be done. However since item 7 (managerial discussion and analysis) is more focused on in-depth analysis of its business in the perspective of firm's financial performance within specific period, it is more likely to have noisy information about the business model. For this reason, information quality often varied across companies in item 7. In this sense, only item 1 part is determined as the corpus to be extracted.

In order to establish the corpus, we first collect entire 10-K filings from EDGAR online database. Throughout the web crawling modules (LWP::UserAgent, WWW::Mechanize), we automatically extracted every document from the database. In order to extract only required part from the document, some of the text processing techniques are applied. Since every 10-K annual report is in html format, we first eliminate unnecessary HTML tags from the document in order to obtain pure textual data. After that, a regular expression matcher which could extract only item 1 chapter out of entire documents is devised.

4.1.2 Sentence filtering

In this step, sentences that are only relevant to the firm's business model are extracted out of item 1. Although item 1 is the most relevant part about firm's business model, it also contains noisy information. For example, sentences reporting numerical indicators, locations, people's name or website are less important for representing a business model. In order to collect only the relevant information, we developed a word index for detecting only business model-related sentences. After careful investigation on the word usage patterns in many sentences, we have found that much of business model information is contained in the introductory sentence. Such introductory sentence uses nouns '*we*,' '*our*,' '*us*,' '*registrant*,' '*strategy*' or '*company*' as their subject, and uses '*allow*' '*enable*' '*sell*' and '*help*' as their verbs. Such word pattern is described in Table 2. Moreover, since numbers or locations should not be included in sentences, in order to improve the information quality, we have also developed the sentence exclusion rules defined by regular expression matcher for filtering irreverent sentences.

4.1.3 Keywords extraction

In this step, keywords are extracted from the collection of sentences. We decided to extract only the noun word as a keyword for business model representation. Because the aim of our study is to extract information on business model, other parts of speech such as the adjective which is frequently used in analyzing subjective opinion are not required. In this sense, we used an 'Lingua::EN::Tagger' module which could extract only the noun words out of sentences. The algorithm tokenizes sentences into individual words. After that, part-of-speech tagger classifies each token according to their parts of speech. Among those tagged tokens, we can easily extract only the noun keywords. In this study, we do not apply stemming procedure to extracted noun words in order to preserve their meaning as much as possible. Currently, the proposed method is not perfect since it does not take into account multi-word phrases which could provide more specific information. We leave this issue for the future works.

4.1.4 Keyword vector generation and interpretation

In this step, a collection of extracted keywords is represented by a single vector. Each keyword is sorted by its frequency based on the assumption that a keyword appears more frequently in the sentence is more important for their business model. The vector representation enables an analyst to quickly grasp a concept about a company's business model. Although keywords with high frequency means dominant concept for its business model, the keywords with less frequency should not be dis-

Table 2. Include/stop words for sentence extraction

Include-word list for sentence filtering	Stop-word list for sentence filtering
we, company, our, us, registrant, sell, allow, enable, provide, product, service, include, ⋯	numbers, peoples name locations, ⋯

carded in the analysis because some of them might indicate future sign about their new business plans or changes. This issue will be discussed in the trend analysis.

4.2 Business Model Evolution Analysis

This section describes procedures for business model evolution analysis. By evolution, we mean temporal change of the business model concept over time. Since we represent the concept as a collection of keywords, we can describe the evolution of the concept throughout the change of its distribution over time. In order to investigate temporal change of the keyword, we develop the unified measure for indicating whether a certain keywords is emerging or disappearing over time. Then, we propose the business model evolution map where each keyword is visualized according to its degree of growth and average frequencies.

4.2.1 Derivation of temporal keywords matrix

First of all, the time period for the analysis is defined by the user. Depending on a user's interest, it might vary from two successive years to entire lifetime of the company. A set of annual reports corresponding the time period are extracted for the analysis. Applying keyword vector generation procedures to each annual report, a series of keyword vectors is obtained. Those vectors are represented by a term-period matrix where the value of (i, t) element in the matrix represents a frequency of a keyword i in year t. Therefore, a time series of a specific keyword's frequency is represented by a raw vector of the matrix. On the other hands, a company's business model on a specific year is represented by a column vector.

4.2.2 Inverse document frequency transformation

In this step, each keyword of the matrix is weighted by its relative importance for trend representation. As a keyword weighting scheme, we apply the term frequentcy - inverse document frequency (TF-IDF) weighting E which is the product of two statistics, term frequency T and inverse document frequency I.

The logic behind term frequency T is that the keyword with high frequently receives much attention for trend detection because it represents major concept of the business model. Therefore, if we denote the frequency of a keyword i in period t by $F_{i,t}$, the term frequency of that keyword is represented in Eq. (1).

$$T_{i,t} = \frac{F_{i,t}}{\max\{F_{w,t} : w \in t\}} \tag{1}$$

In order to avoid bias toward longer document, we also normalize the value by dividing the maximum frequency of the keyword given a period.

On the other hand, the logic behind inverse document frequency I is that the keyword which appears too common throughout the time period receives less attention because it is not changed over time. Therefore, the inverse document frequency of keyword i is obtained by dividing the total number of period by the number of documents containing the keyword i. Therefore the inverse document frequency I of keyword i is shown in Eq. (2).

$$I_i = \log \frac{T+1}{D_i} \tag{2}$$

where T is the total number of time period, and D_i is the number of documents that includes keyword i. Since we take log transformation, the inverse document frequency approaches zero when the keyword appears in entire time periods. Note that we use T+1 instead of T in order to avoid the value becoming zero. Finally, TF-IDF weight of keyword i in period t is calculated as the product of term frequency and inverse frequency as shown in Eq. (3).

$$E_{i,t} = T_{i,t} \times I_i \tag{3}$$

4.2.3 Measurement of degree of growth for keywords

In this step, degree of growth for each keyword is measured. We can identify the degree of growth of the keyword by investigating a pattern of its TF-IDF values over time. A keyword whose value increases with time can be regarded an emerging trend. Otherwise, a keyword with opposite pattern can be regarded as an obsolete trend.

In this study, we apply the linear regression analysis in order to measure the degree of growth of each keyword. Linear regression attempts to model the relationship between two variables by fitting a linear equation to observed data. A linear regression line has an equation of the form $Y = a+bX$ where X is the independent variable and Y is the dependent variable. The linear equation is determined by coefficients a and b. The quantity of a is the 'baseline' value of Y when the $X = 0$. The quantity of b is the slope of the line which indicates how a change of the independent variable affects the value taken by the dependent variable. In the regression analysis, these coefficients are chosen in a manner such that the resulting line is the 'best fit' of the data.

Applying linear regression to the each keyword's time series data, we can also obtain a best fitted linear equation. The slope coefficient of the linear equation could be a good indicator of degree of growth. The sign of the coefficient indicates whether a keyword is emerged (positive) or obsolete (negative) over time. Moreover, the absolute value of the coefficient indicates the 'rate' of its emergence or disappearance.

In this sense, from T ordered pairs $(1, E_{i,1})$, $(2, E_{i,2})$, $\cdots, (t, E_{i,T})$, we define the degree of growth of the keyword i, R_i as the slope coefficient of the regression line obtained by Eq. (4).

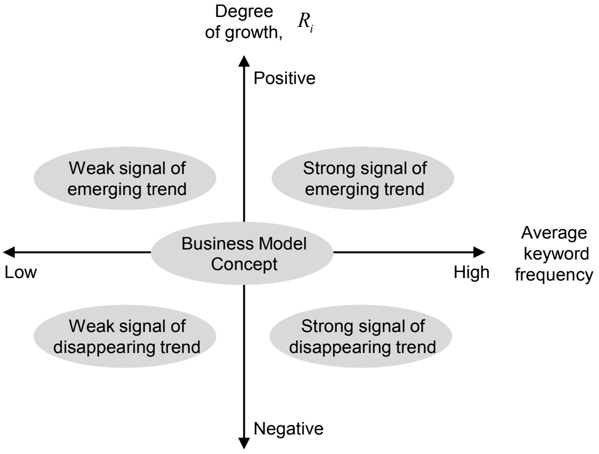
Degree of growth, R_i
Positive
Weak signal of emerging trend
Strong signal of emerging trend
Business Model Concept
Low
High
Average keyword frequency
Weak signal of disappearing trend
Strong signal of disappearing trend
Negative

Figure 1. Business model evolution map.

$$R_i = \frac{\sum_{t=1}^{T}(t-\bar{t})(E_{i,t}-E_i)}{\sum_{t=1}^{T}(t-\bar{t})}, \quad \text{where } \bar{t} = \frac{\sum_{t=1}^{T} t}{T}$$

$$\text{and } E_i = \frac{\sum_{t=1}^{T} E_{i,t}}{T} \qquad (4)$$

4.2.4 Visualization through business model evolution map

In this step, each keyword is visualized throughout the business model evolution map which is shown in Figure 1. In this map, each keyword which has been appeared in the collection of documents is plotted according to the degree of growth (vertical axis) and average frequency (horizontal axis).

By investigating relative position of the keywords on the map, we can briefly identify the overview of the business model evolution. Keywords which are located on the centroid indicate that they are relatively stable over times, and hence provide static concept of the business model. Keywords from upper-right part indicate that they are both emerging and important concept to the business model. On the other hand, keywords on lower-right part are interpreted as obsolete concept as time goes. Keywords on left part are considered as weak signal because they are appeared less frequently from the collection of documents. However they are also required to be investigated because some of them might provide future change about their business model.

5. CASE EXAMPLE

As an illustrated example, we extracted business model concept of three different companies—Blockbuster, Surf a Movie Solution, and Netflix. Those companies are classified as video tape rental industries by Standard Industrial Classification (SIC). Business model concepts of each company are shown in Table 3.

Although those three companies come from same industry, their keyword vectors show the difference of their business model. In order to clarify the extracted concept, extracted keywords are classified according to the business model perspective as illustrated in Table 4. In this study, three major perspectives, i.e., value proposition, revenue model, and distribution channels, are used for such classification. Blockbuster operates mostly offline oriented business model. The keyword such as store, operation, retailer, franchisee, and rental strongly support this argument. The keyword movie and game indicates that their major contents to be rent. In this sense, the distribution of the contents throughout their franchise network might be essential activities for the company.

On the other hand, Netflix doesn't operate offline store. Their major service is internet streaming service. Keywords, such as *internet*, *web* and *vod*, strongly support this argument. Keywords, such as *shipping, delivery, distribution* or *center*, indicate that they also provide contents delivery service which was known as major service of the Netflix. But keywords frequencies show that the company now focuses more on online streaming services. Their contents also differ from Blockbuster. While Blockbuster mainly provides DVD contents such as *movie* or *games*, Netflix also provides TV contents (*tv, episode*).

Surf a Movie Solution's business model lies between Blockbuster and Netflix. The keywords vector contains both words which indicate offline-based model (*store*) and online-based model (*online, web, internet*). Therefore the company might be typical bricks-and-clicks company which integrates both offline and online presence.

As a pilot test of the business model evolution analysis, we analyzed the Netflix's 10-K annual report from 2004 to 2010. Total 437 keywords which appear more than average one times are collected throughout the time periods. As illustrated in Figure 2, the collected keywords are plotted on the keyword portfolio map according to its degree of growth and average frequency. As a boundary for classifying emergence of each keyword, we use 1 standard deviation from average degree of growth which are shown as dotted horizontal lines in Figure 2. Therefore we regard a keyword within the boundary as stable concept which does not varies over time. On the other hands, a keyword outside the boundary is considered dynamic concept which emerges or obsoletes over time. In a similar way, dotted vertical line represents +1 standard deviation of average frequency. We use this line as a boundary for indicating strength of the signal. Therefore, keywords which are located at the right side represent strong signal of the business model concept. On the other hands, keywords at the left side means weak signal about the concept.

Keywords located in each region are tabulated in Table 5. Keywords from upper right region (strong signal of emerging concept) highlight topics which have

Table 3. Business model concept of companies in video tape rental industry (SIC: 7840)

Blockbuster Inc. (2010)		Surf A Movie Solutions Inc. (2010)		Netflix Inc. (2010)	
Keyword	Freq.	Keyword	Freq.	Keyword	Freq.
store	118	video	55	subscriber	62
product	83	store	40	content	38
operation	77	owner	22	dvd	38
game	74	online	20	title	30
video	68	customer	19	service	26
customer	66	web	16	Netflix	23
sale	63	internet	15	internet	18
consumer	59	movie	12	web	17
service	55	business	11	tv	16
distribution	55	page	11	subscription	15
cash	53	site	10	technology	14
studio	52	software	9	merchandising	14
movie	49	product	8	shipping	14
term	48	service	7	site	14
ability	46	information	7	delivery	12
market	42	law	7	time	11
state	41	newsletter	6	video	11
blockbuster	41	market	6	recommendation	10
retailer	40	content	6	center	10
rental	40	shopping	6	library	10
content	40	product	6	distribution	10
home	39	customer	6	studio	10
addition	37	category	5	episode	9
franchisee	36	order	5	device	9
future	36	list	5	customer	9
entertainment	36	company	5	business	8
impact	36	rental	5	provider	8
DVD	36	download	5	VOD	7
debt	36	growth	5	revenue	7
entertainment	36	email	4	consumer	7

Table 4. Keyword classification according to three major business model perspectives

Business model perspective	Blockbuster Inc. (2010)	Surf A Movie Solutions Inc. (2010)	Netflix Inc. (2010)
Value proposition	entertainment, content, game, video, movie, DVD	content, information, video, order, movie	delivery, video, recommendation, tv, episode, VOD, library
Revenue model	rental, cash	rental, download	subscriber, subscription
Distribution channel	store, operation, distribution, retailer, franchisee	newsletter, internet, store, web, email	device, internet, content, shipping, dvd, distribution

become major interest to the company. Keywords related to the company's major service offering, such as *web*, *tv*, and *internet*, are identified as both emerging and major concept. From those keywords we could identify that the company which was originally kwon as a DVD rental service provider now has changed their business model into streaming video service providers. Moreover, keywords, such as *content*, *dvd*, *episode*, and *movie*, provide a company's emphasis on contents acquisition.

Keywords from lower right region (strong signal of obsolete concept) highlight topics which have become less interesting to the company. Among those keywords, recommendation, rating, selection and preference is interesting. They are related to contents recommendation interface of the website. We could interpret that those topic have already become a standard of the industry hence does not considered a source of competency of their business model.

Keywords from left region (weak signal) could be interpreted in a similar way. Those keywords are also required to be investigated because some of them might provide future sign of the new business model. As an emerging weak signal, keywords such as telecommunication, device looks interesting. Those words could be interpreted as their future distribution channel through which could provide contents streaming service to the customer.

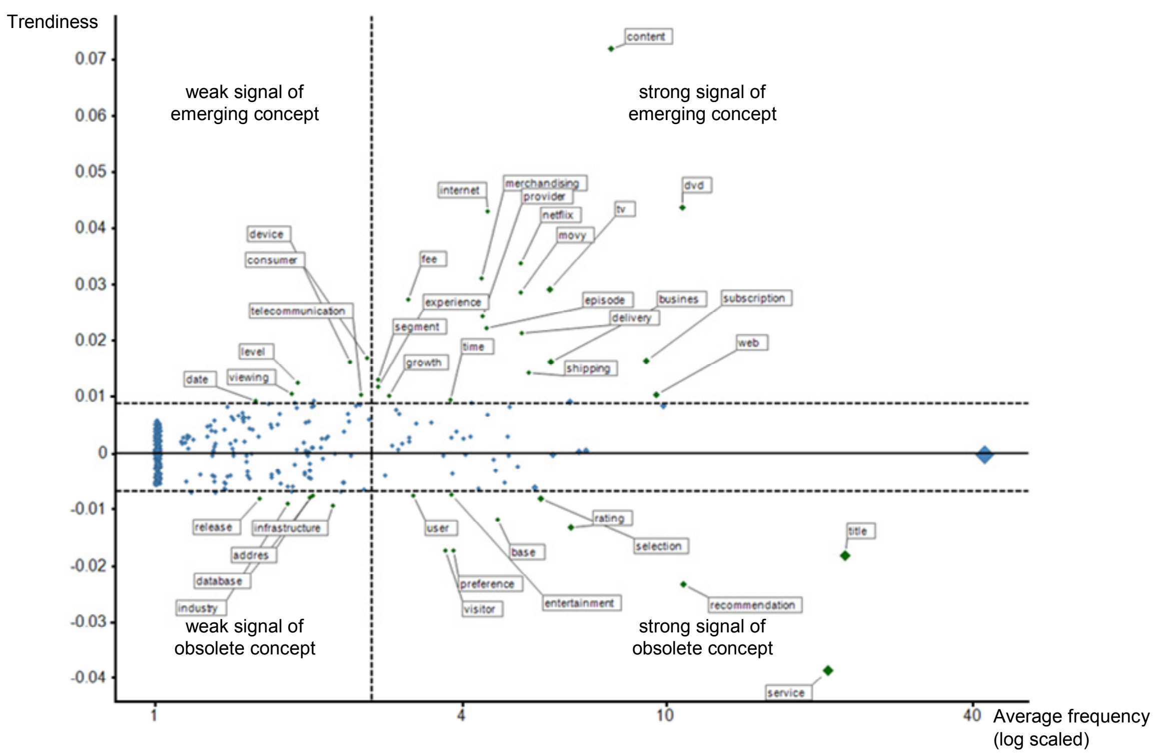

Figure 2. Business model evolution map of Netflix (2004-2010).

Table 5. Keywords indicating business model evolution

Strong signal of emerging concept			Strong signal of obsolete concept			Weak signal of emerging concept			Weak signal of obsolete concept		
Keywords	Trend.	Fr.	Keywords	Trend.	Fr.	Keywords	Trend.	Fr.	Keywords	Trend.	Fr.
dvd	0.044	10.71	title	-0.018	22.28	consumer	0.017	2.60	infrastructure	-0.009	2.20
web	0.011	9.42	service	-0.039	20.57	telecommunication	0.011	2.50	database	-0.008	2.00
subscription	0.016	9.28	recommendation	-0.023	10.71	device	0.017	2.40	address	-0.008	2.00
content	0.072	7.85	rating	-0.013	6.57	competitor	0.009	2.00	industry	-0.009	1.83
technology	0.009	6.57	selection	-0.008	5.71	level	0.013	1.85	release	-0.008	1.57
business	0.016	6.00	base	-0.012	4.57	viewing	0.011	1.85	maximum	-0.007	1.33
tv	0.029	6.00	entertainment	-0.008	3.85	date	0.009	1.60	position	-0.006	1.20
netflix	0.033	5.28	preference	-0.017	3.83						
shipping	0.014	5.28	visitor	-0.018	3.75						
delivery	0.021	5.28	user	-0.007	3.16						
movy	0.028	5.14									
internet	0.043	4.50									
episode	0.022	4.50									
provider	0.025	4.28									
merchandising	0.031	4.25									
time	0.009	3.85									
fee	0.027	3.14									
growth	0.011	2.85									
experience	0.012	2.71									
segment	0.013	2.71									

6. CONCLUSION AND FUTURE WORKS

As the term business model has received considerable attention these days, the ability to collect information about the business model has become essential requirement. In this study, we proposed a text-mining methodology for extracting business model related knowledge from annual 10-K reports. Our methodology is applied in 1) business model concept extraction and 2) business model evolution analysis. The utility of our methodology is illustrated by a case example of a real company.

However, there are further research issues to enhance our methodology. First, more reliable methodology for extracting business model related knowledge is required. For example, more sophisticated rule for judging business model related sentence is required. Various machine learning algorithms and conceptual schemes about the business model could be utilized in such sentence filtering process. Also, the keyword extraction process should be enhanced. Currently, we only extracted single noun keywords from the text. Considering multi-words phrases for the keyword might provide more specific and desirable information.

Secondly, the scope of our methodology could be extended. Currently, our methodology is limited to handling a single company's business model. By incorporating multiple companies into the analysis, we can detect industry-wide trends or opportunity for a new business model. More advanced data mining techniques, such as clustering methods or singular vector machine algorithm, might be applicable to such pattern classification.

Finally, we can also incorporate numerical data into the analysis. The 10-K annual report contains various types of numerical data, such as security prices or financial statements. Combining those numerical data with textual data may open the path for more advanced business model analyses, such as performance prediction or risk analysis.

ACKNOWLEDGMENT

This work was supported by the "Core Technology Development Program for Knowledge-Based Service" funded by the Korean Government (MOTIE) (Project No. 10048090, Title: Manufacturing Servitization Support Framework). The authors deeply appreciate the administrative support during the project period from Engineering Research Institute (ERI) of Seoul National University.

REFERENCES

Casadesus-Masanell, R. and Ricart, J. E. (2010), From strategy to business models and onto tactics, *Long Range Planning*, **43**(2), 195-215.

Cho, G. H., Lim, S. Y., and Hur, S. (2014), An analysis of the research methodologies and techniques in the industrial engineering using text mining, *Journal of Korean Institute of Industrial Engineering*, **40**(1), 52-59.

Delen, D. and Crossland, M. D. (2008), Seeding the survey and analysis of research literature with text mining, *Expert Systems with Applications*, **34**(3), 1707-1720.

Hoberg, G. and Phillips, G. M. (2010), Text-based network industries and endogenous product differentiation, *NBER Working Paper No. 15991*, National Bureau of Economic Research, Cambridge, MA.

Johnson, M. W., Christensen, C. M., and Kagermann, H. (2008), Reinventing your business model, *Harvard Business Review*, **86**(12), 57-68.

Li, F. (2008), Annual report readability, current earnings, and earnings persistence, *Journal of Accounting and Economics*, **45**(2), 221-247.

Linder, J. and S. Cantrell (2000), Changing business models: surveying the landscape, *Working Paper*, Institute for Strategic Change at Accenture, Dublin, Ireland.

Magretta, J. (2002), Why business models matter, *Harvard Business Review*, **80**(5), 86-93.

Malone, T., Weill, P., Lai, R., D'Urso, V., Herman, G., Apel, T., and Woerner, S. (2006), Do some business models perform better than others? *MIT Sloan Research Paper No. 4615-06*, MIT Sloan School of Management, Cambridge, MA.

Martin, N. J. and Rice, J. L. (2007), Profiling enterprise risks in large computer companies using the Leximancer software tool, *Risk Management*, **9**(3), 188-206.

Miner, G., Delen, D., Elder, J., Fast, A., Hill, T., and Nisbet, R. A. (2012), *Practical Text Mining and Statistical Analysis for Non-Structured Text Data Applications*, Waltham, MA: Academic Press.

Mikawa, K., Ishida, T., and Goto, M. (2011), An optimal weighting method in supervised learning of linguistic model for text classification, *Industrial Engineering and Management Systems*, **11**(1), 87-93.

Netflix (2011), Form 10-K (annual report), Available from: http://www.sec.gov/edgar/searchedgar/companysearch.html.

Osterwalder, A. and Pigneur, Y. (2010), *Business Model Generation: A Handbook for Visionaries, Game Changers, and Challengers*, Hoboken, NJ: Wiley.

Pohle, G. and Chapman, M. (2006), IBM's global CEO report 2006: business model innovation matters, *Strategy and Leadership*, **34**(5), 34-40.

US Securities and Exchange Commission (2012), Form 10-K, Available from: http://www.sec.gov/answers/form10k.htm.

Timmers, P. (1998), Business models for electronic markets, *Electronic Markets*, **8**(2), 3-8.

Tseng, Y. H., Lin, C. J., and Lin, Y. I. (2007), Text mining techniques for patent analysis, *Information Processing and Management*, **43**(5), 1216-1247.

Zhang, Y. and Jiao, J. R. (2007), An associative classification-based recommendation system for personalization in B2C e-commerce applications, *Expert Systems with Applications*, **33**(2), 357-367.

Zott, C. and Amit, R. (2008), The fit between product market strategy and business model: implications for firm performance, *Strategic Management Journal*, **29**(1), 1-26.

Zott, C. and Amit, R. (2010), Business model design: an activity system perspective, *Long Range Planning*, **43**(2), 216-226.