Advanced Engineering Informatics xxx (2015) xxx–xxx

Contents lists available at ScienceDirect

Advanced Engineering Informatics

journal homepage: www.elsevier.com/locate/aei

Graphical causal inference and copula regression model for apple keywords by text mining

Jong-Min Kim [a], Sunghae Jun [b,*]

[a] Statistics Discipline, Division of Sciences and Mathematics, University of Minnesota-Morris, Morris, MN 56267-2134, USA
[b] Department of Statistics, Cheongju University, Chungbuk, 360-764, Republic of Korea

ARTICLE INFO

Article history:
Received 6 April 2015
Received in revised form 21 July 2015
Accepted 4 October 2015
Available online xxxx

Keywords:
Apple keywords
Text mining
Patent analysis
Graphical causal inference
Copula regression

ABSTRACT

Apple is a leading company of technological evolution and innovation. This company founded and produced the Apple I computer in 1976. Since then, based on its innovative technologies, Apple has launched creative and innovative products and services such as the iPod, iTunes, the iPhone, the Apple app store, and the iPad. In many fields of academia and business, diverse studies of Apple's technological innovation strategy have been performed. In this paper, we analyze Apple's patents to better understand its technological innovation. We collected all applied patents by Apple until now, and applied statistics and text mining for patent analysis. By using graphical causal inference method, we created the causal relations among Apple keywords preprocessed by text mining, and then we carried out the semiparametric Gaussian copula regression model to see how the target response keyword and the predictor keywords are relating to each other. Furthermore, Gaussian copula partial correlation was applied to Apple keywords to find out the detailed dependence structure. By performing these methods, this paper shows the technological trends and relations between Apple's technologies. This research could make contributions in finding vacant technology areas and central technologies for Apple's R&D planning.

1. Introduction

In technological evolution and innovation, Apple has been a leading company. The innovative products and services of Apple, such as the iPod, iTunes, the iPhone, the Apple app store, and the iPad have evolved over a period of time and have been one of the most sought after products in the global marketplace. From its foundation in 1976, Apple's technological innovation has become more impressive. Technological innovation of a company is very important to improve its competitiveness, so many studies on technological innovation have been performed [1–3]. Most previous studies on technological innovation were based on qualitative and subjective approaches [3–5], or used descriptive statistics such as frequency, summary statistics, and visualization [6]. Also, the study on examination of Apple's technological innovation using linear model, clustering and social network analysis (SNA) was published [7]. This was focused on the analysis of International Patent Classification (IPC) codes from the Apple's patent documents, because an IPC code represents its assigned technology. However, this is limited to only the technological innovation of Apple because such simple analytical methods are restrictive.

Therefore, in this study we want to examine how the target response key term and the predictor key terms are relating to each other, as well as show the technological trends and relations between Apples technologies. To achieve these goals, we will use graphical causal inference and copula methods in this paper. Copula has been increasingly popular as a flexible tool for modeling the dependence of multivariate data in many fields of application such as biostatistics, medical research, econometrics, finance, actuarial science, and hydrology because a copula does not require a normal distribution and independent, identical distribution assumptions [8–10]. Furthermore, the invariance property of copula has been attractive in the area of finance. Kim et al. [11] applied a copula method for modeling directional dependence of genes as an alternative method of Bayesian network which is a probabilistic graphical method. Kim and Kim [9] proposed the improved copula method for modeling directional dependence of genes. The copula arises from Sklar's theorem [12], which proves that for any q continuous random variables, there exists a unique function, which is a copula that couples q univariate marginal distributions into the q-dimensional distribution function.

* Corresponding author. Tel.: +82 10 7745 5677; fax: +82 43 229 8432.
E-mail address: shjun@cju.ac.kr (S. Jun).

http://dx.doi.org/10.1016/j.aei.2015.10.001
1474-0346/© 2015 Elsevier Ltd. All rights reserved.

The main advantages of using the copula method are as follows: (1) one can model the dependence structure of the joint distribution and its marginal distributions separately and, when the variables are transformed by increasing transformations such as rank, the copula describing these transformed variables is the same as the copula for the original variables, and (3) one can use a rank-based approach to the inference on the joint dependence by means of copulas, independent of the marginal. Standard references for a detailed overview of copula theory and applications include the books by Joe [13] and Nelson [14]. Wang and Hua [15] proposed a semiparametric Gaussian copula text regression model for predicting financial risks from earnings calls. The method used kernel density function of text features as marginal distribution and then used Gaussian copula for modeling joint conditional distribution for dependencies with multivariate text features. Wang and Hua [15] compared their copula method with statistical machine learning methods. So, our proposed copula regression model is the first paper to apply copula regression to a structured text data by parsing from the text corpus in text mining processing. By using graphical causal inference, we clarify response variables and predictor variables among Apple keywords and then we apply semiparametric Bayesian Gaussian copula regression model to Apple keywords, which consist of large datasets. The computation of our proposed copula regression model for large data is fast by using the semiparametric inference for copula models via a type of rank likelihood function for the association parameters proposed by Hoff [16].

The copula model was researched in statistics originally [8–10,16]. Also it has been used in diverse fields such as bioinformatics and finance, and shown good performances in their applications [11,15]. But, no copula research has ever been in patent data analysis. So we applied the copula model to patent analysis for understanding Apple's technologies. In this paper, we did not proposed new method or algorithm for data analysis, and we got the analytical result of Apple's technologies. Therefore we could not compare our results to other analytical methods by traditional evaluation measures such as predictive accuracy, misclassification rate, or computing time. The main contribution of this paper is to discover the technological relationship between Apple's technologies using graphical causal inference and copular regression. Our research result can be used for the R&D groups to study the technological innovation of the Apple. Most products and services have been based on the technology. Apple is also depended on his holding technologies for developing new products and services. So it is significant to find the technological relationship between the technologies for understanding Apple's innovation. Our case study of Apple can be expanded to the companies which want to learn the Apple's technological innovation.

The remainder of this paper is organized as follows. Section 2 introduces the text mining to find important keywords from the data. We apply the graphical causal inference method to data in order to find cause and effective keywords from data in Section 3. The semiparametric Bayesian Gaussian copula method will be applied to data for regression modeling in Section 4. The Gaussian copula partial correlation will be applied to understand the conditional dependence structure of causality inference in Section 5. Conclusions are presented in Section 6.

2. Text mining for Apple patent

The research question of our paper is how to find the technological relationship between the Apple's holding technologies, because this is important to perform the Apple's R&D planning. Also this is meaningful for the research groups who want to know Apple's innovation. So we use the Apple's keywords extracted from Apple's patent documents by text mining techniques. Text mining is an emerging research area that has become popular in the last decade, and it includes interdisciplinary techniques from linguistics, computer science, mathematics, and statistics, such as natural language processing, keywords extraction, text segmentation, term association, text clustering, and big data [17–24]. Also text mining is important techniques to analyze patent documents for technology analysis [25–31]. Since the most natural form of storing information is as text data as non-structured or less structured [32–34], we should transform them into structured data for data analysis. By an automated preprocess using text mining, text data extracted from the document is stored in a structured data describing documents [18,19]. Recently, studies about extracting novel patterns have not only been applied to text based journals, or newspapers, but applications to complicated and non-traditional data such as facial image, and social network service, have also been proposed [35–37]. In this paper, we collected patent documents applied by Apple from patent databases [38,39]. These patent data contain diverse results of developed technology such as patent title and abstract, IPC code, application number and date, figures and drawings, claims, and so on [40,41]. Jun and Park [7] extracted IPC codes from the Apple's patent data, and analyzed them for examining technological innovation of Apple.

In this paper, we used keywords from Apple's patent documents and analyze them by graphical causal inference and copula regression model. Keywords are defined as words providing representative meaning of a document [17]. We used R as a language for text mining and keyword analysis [42]. In addition, the package 'tm' based on R was applied to our text mining of Apple's patent documents [43]. We get Apple's keywords by the following text mining process (see Fig. 1).

In this paper, we selected titles and abstracts from collected Apple's patent documents. These text data were transformed into text corpus, collection of titles, and abstracts for natural language processing. Using the parsing technique of text mining, we built structured text data for keywords extraction. The structured data consist of a matrix composed of patent (row) and term (column). The element of this matrix is occurred frequency of a specific term in each patent. After removing common words such as "and", "the", and "is" from the matrix, we extracted Apple's keywords by frequency of each term. The Apple keywords with at least 1000 frequencies are presented in Table 1.

In Table 1, the number of keywords with at least 1000 frequencies is 31. The first keyword "device" has 6919 patents, the second keyword "data" has 6676 patents, and the last keyword has 1059

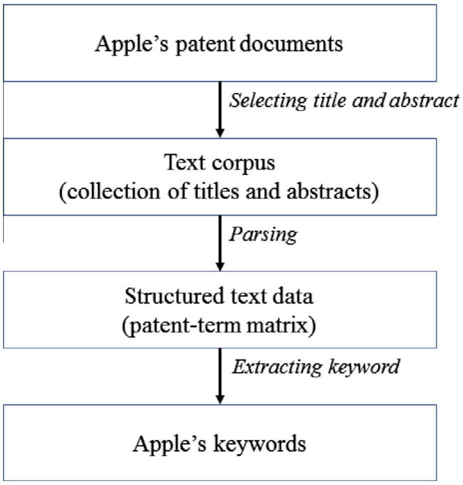

Fig. 1. Text mining procedure of Apple's keywords.

Please cite this article in press as: J.-M. Kim, S. Jun, Graphical causal inference and copula regression model for apple keywords by text mining, Adv. Eng. Informat. (2015), http://dx.doi.org/10.1016/j.aei.2015.10.001

Table 1
Apple's keywords with at least 1000 frequencies.

Order	Keywords	Frequency
1	Device	6919
2	Data	6676
3	System	5936
4	User	4767
5	Media	4157
6	Display	3468
7	Computer	3249
8	Electronic	3179
9	Interface	2391
10	Information	2243
11	Image	2230
12	Memory	2058
13	Power	1991
14	Control	1722
15	Signal	1680
16	Video	1643
17	Plurality	1609
18	Audio	1589
19	Portable	1499
20	Application	1498
21	Invention	1469
22	Circuit	1346
23	Present	1343
24	Set	1330
25	Portion	1298
26	Digital	1231
27	Content	1172
28	Object	1167
29	Color	1120
30	Time	1078
31	Network	1059

frequencies. In the whole of this paper, we used these keywords for Apple's patent analysis. To represent Table 1 visually, we made a word-cloud of Table 1 in Fig. 2.

We could easily understand the distribution of Apple's keywords through the word-cloud plot in Fig. 2. We were able to verify that the keywords of device "data", "system", "user", "media", "display", "computer", "electronic", etc. were observed most often. So we could find Apple's technologies were dependent on these keywords.

Next we present another analysis using frequency-based evaluation with Apple's keywords. We computed correlation coefficients between the keywords to find associations amongst Apple's keywords. By the minimal correlation for identifying valid association, we got visualization result of the correlations within patent–keyword matrix, our structured data, in Fig. 3.

data
computer
electronic
invention
application
content
system
digital
memory
user
audio
present
signal
set
circuit
time
color
network
image
information
object
portable
video
plurality
portion
media
device
power
control
interface
display

Fig. 2. Word-cloud plot of Apple's keywords.

We determined the minimal coefficient was greater than or equal to 0.1. In Fig. 3, we found that Apple's keywords were generally divided into two technological groups generally. The keywords of the first group are associated to each other around "power" and "portable'. In the second group, the keywords are connected by centers of "system" and "information". Other keywords not belonging to two groups mean relatively independent technologies. Accordingly, as the value of minimal coefficient changes, the graphic structure of visualization in Fig. 3 changes. So, we are able to find associations between Apple's keywords through text mining and roughly through the visualization thereof, also the main goal of text mining was to extract the keywords from Apple's patent documents. Next, we will analyze Apple's keywords in earnest by using graphical causal inference and copula regression model.

3. Causal inference with directed acyclic graph

Understanding cause–effect relationships between Apple keywords is of primary interest in this study. We therefore consider the problem of inferring causal information from Apple keywords observational data. Under some assumptions, the PC algorithm [44], the FCI algorithm [44,45], and the RFCI algorithm [46] can infer information about the causal structure from observational data. These algorithms can show which variables could or could not be a cause of some variable of interest. In this paper, we used the PC algorithm function in R package "pcalg" to find the causal structure for Apple keywords data. So we assume that there are no hidden variables and no feedback loops in the underlying causal system. The causal structure of such a system can be conveniently represented by a Directed Acyclic Graph (DAG), where each node represents a variable and each directed edge represents a direct cause [47,48]. Each linkage in the DAG means a linear regression model. That is, the dependent variable (keyword) is in the beginning of the direct arrow and, the dependent variable (keyword) is in the end of the arrow. In the Apple's keyword analysis by the DAG, the technology based on independent variable affects the technological development of dependent variable. In this paper, the relation is evaluated by t-value of statistical testing.

To do graphical causal inference and copula regression model for Apple keyword data, we converted the discrete data from 31 Keywords to continuous data. We can justify our approach by introducing Denuit and Lambert [49]: given integer value X_i, consider a continuous random variable $X_i^* = X_i + U_i$ where U_i is uniform on (0, 0.001) and independent of X_i. As shown in Denuit and Lambert [49] and Madsen and Fang [50], the original variable can be recovered from its continuous extension, and the distribution function of the original variable is exactly the same as that of its continuous extension. Furthermore, this approach randomly breaks the ties in the data. Note that Kojadinovic and Yan [10] verified that the randomization (designed to randomly break the ties) does not change the results for the copula inference. We compared summary statistics with the discrete, original version of Apple keywords and its continuous version with at least 2000 patents in Tables 2 and 3.

We can find that the results from Tables 2 and 3 have almost the same values. By these theoretical background and results, we can avoid any controversy by using the continuous version of the Apple keywords data rather than the original discrete data. Table 3 shows that Apple keywords continuous version data with at least 2000 patents have positive skewness so that the data does not follow

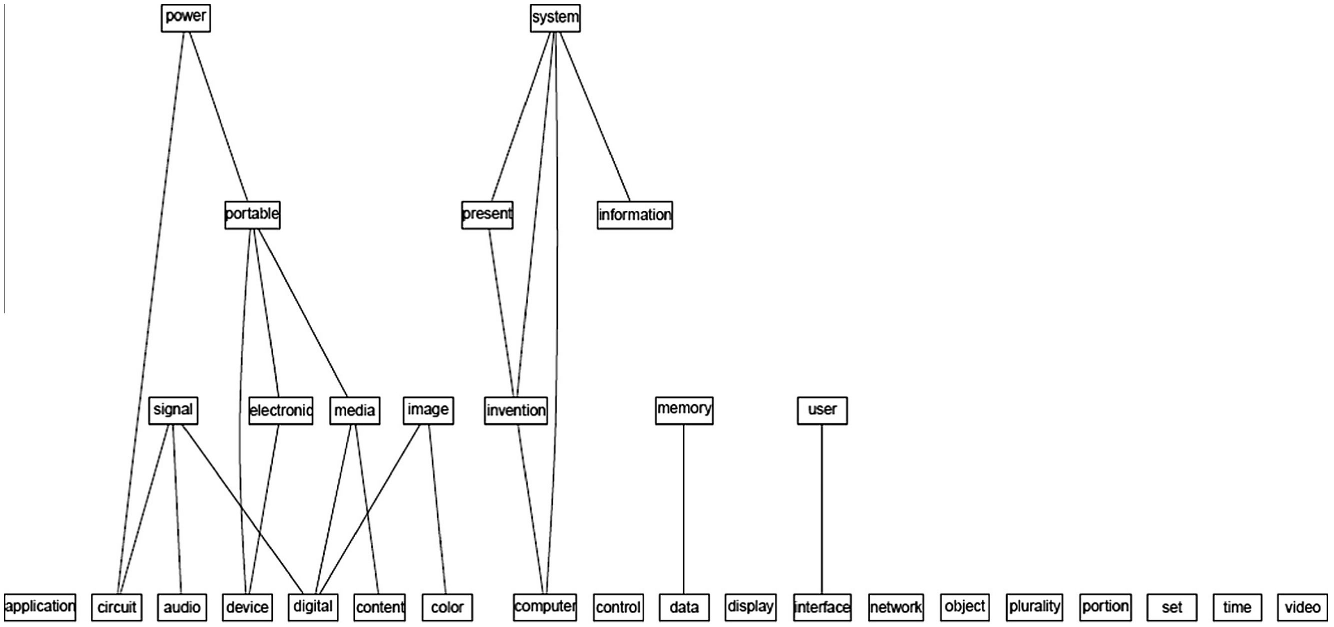

Fig. 3. Visualization of associations between Apple's keywords.

Table 2
Summary statistics of Apple keywords discrete original version with at least 2000 patents.

Keyword	Min	1st quartile	Median	Mean	3rd quartile	Max
Computer	0	0	0	0.4002	0	13
Data	0	0	0	0.8223	0	32
Device	0	0	0	0.8522	1	18
Display	0	0	0	0.4271	0	18
Electronic	0	0	0	0.3916	0	16
Image	0	0	0	0.2747	0	26
Information	0	0	0	0.2763	0	18
Interface	0	0	0	0.2945	0	12
Media	0	0	0	0.512	0	22
Memory	0	0	0	0.2535	0	24
System	0	0	0	0.7311	1	16
User	0	0	0	0.5871	0	14

a normal distribution. This is a good motivation to use copula method to Apple keywords in the following sessions. So Table 4 shows the empirical Kendall's nonparametric correlations of Apple keywords continuous version with at least 2000 patents.

From Table 4, we can see three relatively high correlation cases that the correlation of keywords "user" and "interface" is 0.231, the correlation of keywords "electronic" and "device" is 0.21, and the correlation of keywords "computer" and "system" is 0.112. Next with the PC function in R package "pcalg", we made a DAG with 31 Keywords with 2000 patents in Fig. 4.

In Fig. 4, we can see the keywords of "display", "information", "electronic", "data", "user", etc. are closely related to other keywords. So we know the technologies based on these keywords can be candidates for core technologies of Apple. To know more detailed association between major keywords, we show a clear graphical causal inference with 12 Keywords with 2000 patents in Fig. 5.

From the DAG in Fig. 5, we found the technological impact of Apple's technology. For example, "computer" technology affects the technologies of "data", "electronic", and "display". Also, technology based on "data" influences to the technologies based on "system", "memory", "electronic", "image", and "display". In addition, "device" based technology affects the technology of "media" directly and influences technology of "media" indirectly via "display".

With the causal structure information from Figs. 4 and 5, we define the terminal node as the response keyword and define keywords with each directed edge representing a direct cause as predictor variables in this paper. With the defined response keyword and predictor keywords, we proceed to make the copula regression model in the following session.

Table 3
Summary statistics of Apple keywords continuous version with at least 2000 patents.

Keyword	Mean	Median	Minimum	Maximum	St.D.	Skewness	Kurtosis
Computer	0.40067	0.00063	0	13.00071	1.04993	4.1257	26.78634
Data	0.82277	0.00066	0	32.00033	2.03863	3.80393	24.34827
Device	0.8527	0.00075	0	18.00039	1.66994	2.85572	13.70574
Display	0.42764	0.00061	0	18.00027	1.26041	4.66809	31.91295
Electronic	0.39206	0.0006	0	16.00054	1.18712	4.45706	28.96742
Image	0.27517	0.00055	0	26.00041	1.29164	7.34101	75.76041
Information	0.27676	0.00056	0	18.00091	0.9793	5.69403	53.41646
Interface	0.295	0.00059	0	12.00075	0.89776	4.66688	33.74265
Media	0.51251	0.00058	0	22.00047	1.93433	5.15915	33.90798
Memory	0.25398	0.00054	0	24.00009	1.20984	7.49059	79.71095
System	0.73162	0.00072	0	16.00066	1.45136	2.98936	15.44064
User	0.58764	0.00065	0	14.00049	1.42895	3.64265	19.66393

Table 4
Empirical Kendall's correlations of Apple keywords with at least 2000 patents.

Keyword	Computer	Data	Device	Display	Electronic	Image
Computer	1.000	0.016	−0.079	0.046	−0.068	0.003
Data	0.016	1.000	−0.008	−0.049	−0.065	0.011
Device	−0.079	−0.008	1.000	0.007	0.210	−0.012
Display	0.046	−0.049	0.007	1.000	−0.024	0.021
Electronic	−0.068	−0.065	**0.210**	−0.024	1.000	−0.023
Image	0.003	0.011	−0.012	0.021	−0.023	1.000
Information	0.019	0.041	0.016	−0.008	−0.022	−0.001
Interface	0.028	0.015	−0.007	0.05	−0.019	−0.02
Media	−0.033	0.009	0.040	−0.001	0.008	−0.004
Memory	0.016	0.047	−0.006	−0.023	−0.012	0.013
System	**0.112**	0.082	−0.077	−0.003	−0.070	0.024
User	0.052	−0.01	0.000	0.088	0.014	−0.023
	Information	Interface	Media	Memory	System	User
Computer	0.019	0.028	−0.033	0.016	0.112	0.052
Data	0.041	0.015	0.009	0.047	0.082	−0.010
Device	0.016	−0.007	0.040	−0.006	−0.077	0.000
Display	−0.008	0.050	−0.001	−0.023	−0.003	0.088
Electronic	−0.022	−0.019	0.008	−0.012	−0.070	0.014
Image	−0.001	−0.020	−0.004	0.013	0.024	−0.023
Information	1.000	0.013	0.020	0.009	0.055	0.036
Interface	0.013	1.000	0.023	0.007	0.028	0.231
Media	0.020	0.023	1.000	−0.001	−0.003	0.037
Memory	0.009	0.007	−0.001	1.000	0.029	−0.037
System	0.055	0.028	−0.003	0.029	1.000	0.017
User	0.036	**0.231**	0.037	−0.037	0.017	1.000

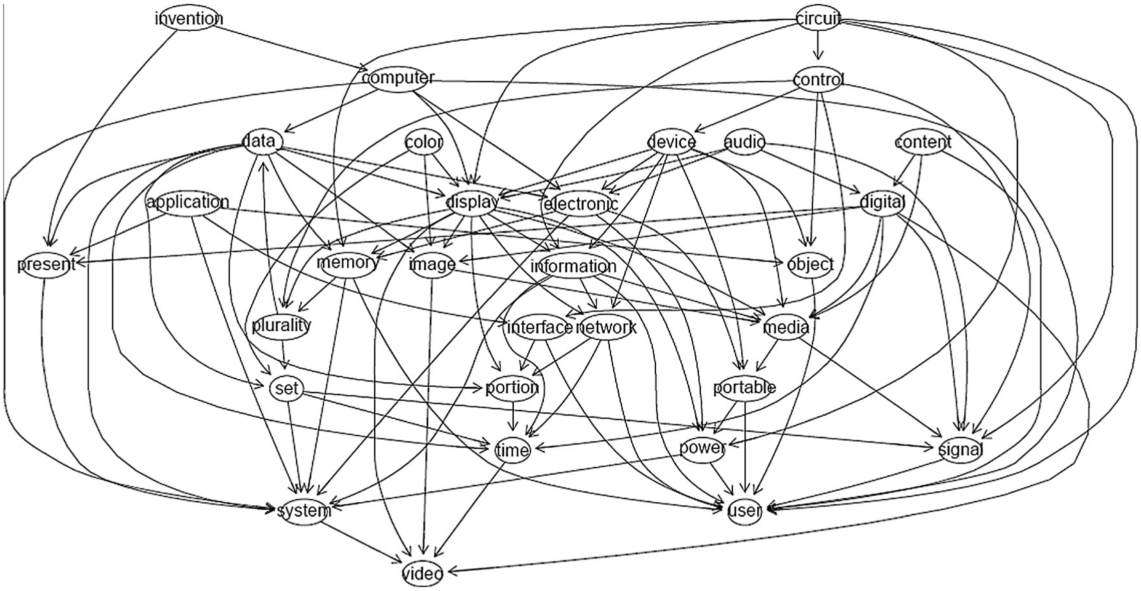

Fig. 4. Directed Acyclic Graph (DAG) with 31 Keywords with 1000 patents.

4. Copula method

A copula is a multivariate distribution function defined on the unit $[0,1]^n$, with uniformly distributed marginals. In this paper, we focus on a bivariate (two-dimensional) copula, where $n = 2$. Sklar [12] shows that any bivariate distribution function, $F_{XY}(x,y)$, can be represented as a function of its marginal distributions of X and Y, $F_X(x)$ and $F_Y(y)$, by using a two-dimensional copula $C(\cdot,\cdot)$. More specifically, the copula may be written as:

$$F_{XY}(x,y) = C(F_X(x), F_Y(y)) = C(u, v),$$

where u and v are the continuous empirical marginal distribution function $F_X(x)$ and $F_Y(y)$, respectively. Note that u and v have uniform distribution $U(0,1)$. Therefore, the copula function represents how the function, $F_{XY}(x,y)$, is coupled with its marginal distribution functions, $F_X(x)$ and $F_Y(y)$. It also describes the dependence mechanism between two random variables by eliminating the influence of the marginals or any monotone transformation of the marginals.

Let X, Y be random variables with continuous distribution functions $F_X(x)$ and $F_Y(y)$, respectively, and let X and Y be continuous random variables with copula C and marginal distribution functions $F_X(x)$ and $F_Y(y)$ so that $X \sim F_X(x), Y \sim F_Y(y)$, and $(X,Y) \sim F_{XY}(x,y)$, and let $u = F_X(x)$, $v = F_Y(y)$, and $(u, v) \sim C$. Then Spearman's ρ and Kendall's τ are given, respectively, by

$$\rho_C = 12 \int_0^1 \int_0^1 [C(u, v) - uv] du dv,$$

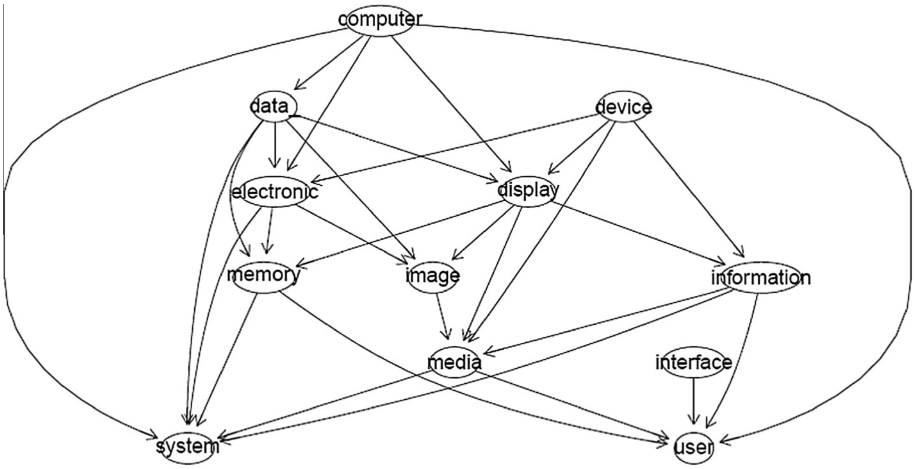

Fig. 5. Directed Acyclic Graph (DAG) with 12 Keywords with 2000 patents.

and

$$\tau_C = 4\int_0^1\int_0^1 C(u,v)dC(u,v) - 1,$$

[14].

4.1. Copula modeling with 12 Keywords with 2000 patents

Before obtaining the ranks for each keyword, however, we noticed that 12 Keywords with 2000 patents are discrete variables valued in a subset of the set of the integers, and there are ties in the data. Note that the presence of ties in the data may substantially affect the copula estimation. In order to fully facilitate the good properties of the copula, one needs to deal with the discreteness of the variable and the presence of ties in an appropriate way. Thus, we obtained a continuous extension of 12 Keywords by using the randomization technique proposed by Denuit and Lambert [49] in Section 3 and applied a continuous extension of 12 Keywords to copula regression.

To do a copula regression with large dataset, we used the semiparametric inference for copula models via a type of rank likelihood function for the association parameters by Hoff [16] because of its simplicity and fast computation for large dataset. The semiparametric inference is based on a generalization of marginal likelihood, called an extended rank likelihood, which does not depend on the univariate marginal distributions of the data. Estimation and inference for parameters of the Gaussian copula are available via a straightforward Markov Chain Monte Carlo algorithm based on Gibbs sampling. Specification of prior distributions or a parametric form for the univariate marginal distributions of the data is not necessary. Before applying a continuous extension of 12 Keywords to semiparametric copula regression, we transformed data to standard uniform distributed data through its empirical distribution function. And then by using the semiparametric Bayesian Gaussian copula estimation, we made posterior quantiles, mean and standard deviation of regression coefficients after the number of iterations is 250 times in the following Tables in this section. If we look at the t-value of regression coefficients in list which tables, then we can find the statistically significant estimates in again list which tables for individual regression model. From Fig. 5, we defined response variable and predictor variables with a continuous extension of 12 Keywords. In Table 5, “system” is a response variable because the “system” node in Fig. 5 is an end node and the “system” node has six direct arrows from “computer”, “data”, “electronic”, “information”, “media” and “memory” which are predictor variables. Tables 5 shows that “computer”, “data”, “electronic,” and “information” are statistically significant at the 5% significance level.

Table 5
Estimates of parameters of system = computer + data + electronic + information + media + memory with 12 Keywords with 2000 patents.

Keyword	2.5‰	Median	97.5‰	Mean	St.D.	t-value
Computer	0.3	0.37	0.44	0.37	0.03764	**9.83**
Data	0.16	0.24	0.32	0.24	0.03888	**6.17**
Electronic	−0.26	−0.2	−0.12	−0.2	0.03521	**−5.68**
Information	0.11	0.19	0.25	0.19	0.03988	**4.76**
Media	−0.07	0.01	0.08	0.01	0.03869	0.26
Memory	−0.01	0.07	0.14	0.07	0.03916	1.79

In this table, we show some descriptive statistics. The 2.5‰ and 97.5‰ are the numbers that have 2.5% and 97.5% of the regression parameter values from minimum. Also by median, mean, standard deviation (St.D.), we can explain the summary of the parameters. From Table 5, we find that development of “system” technology is dependent on the technology developments of “computer”, “data”, “electric”, and “information”, because absolute t-values of these keywords are larger than the threshold of the critical region.

In Table 6, “user” is the response variable because the “user” node in Fig. 5 is the end node, and the “user” node has five direct arrows from “computer”, “information”, “interface”, “media” and “memory” which are predictor variables. Table 6 shows that “computer”, “information”, “interface”, “media” and “memory” are statistically significant at the 5% significance level.

In Table 7, “media” is the response variable because media node in Fig. 5 is the middle node but “media” node has four direct arrows from “device”, “display”, “image,” and “information” which are predictor variables. Table 7 shows that “device” and “image” are statistically significant at the 5% significance level.

Table 6
Estimates of parameters of user = computer + information + interface + media + memory with 12 Keywords with 2000 patents.

Keyword	2.5‰	Median	97.5‰	Mean	St.D.	t-value
Computer	0.08	0.17	0.23	0.16	0.04020	**3.98**
Information	0.04	0.11	0.18	0.11	0.03627	**3.03**
Interface	0.74	0.82	0.88	0.82	0.03755	**21.84**
Media	0.05	0.12	0.2	0.12	0.03806	**3.15**
Memory	−0.2	−0.14	−0.06	−0.14	0.03810	**−3.67**

Please cite this article in press as: J.-M. Kim, S. Jun, Graphical causal inference and copula regression model for apple keywords by text mining, Adv. Eng. Informat. (2015), http://dx.doi.org/10.1016/j.aei.2015.10.001

Table 7
Estimates of parameters of media = device + display + image + information with 12 Keywords with 2000 patents.

Keyword	2.5‰	Median	97.5‰	Mean	St.D.	*t*-value
Device	0.08	0.14	0.22	0.15	0.04122	**3.64**
Display	−0.1	−0.02	0.05	−0.02	0.03793	−0.53
Image	0.04	0.11	0.19	0.11	0.03708	**2.97**
Information	−0.09	−0.01	0.06	−0.01	0.03762	−0.27

We then constructed a technology path diagram from the results of Tables 5–7. Fig. 6 shows this diagram from copula modeling with 12 Keywords with 2000 patents.

In Fig. 6, we found that the keywords of “device” and “image” affect “media”, and technology of “media” influences the technology of “use”. Also, “computer” and “information” affect the two keywords of “system” and “user” at the same time. This technology path diagram can be used for R&D planning of Apple. In the next section, we deal with more spacious copula modeling for Apple’s keywords.

4.2. Copula modeling with 31 Keywords with 1000 patents

Similar to Section 4.1, with a continuous extension of 31 Keywords with 1000 patents, we transformed data to standard uniform distributed data through its empirical distribution function. Then, by using the semiparametric Bayesian Gaussian copula estimation, we made posterior quantiles, mean and standard deviation of regression coefficients after the number of iterations is 250 times in the following Tables. If we look at the *t*-value of the regression coefficients in the Tables, then we can find the statistically significant estimates in the Tables for each individual regression model. From Fig. 4, we defined the response variable and the predictor variables with a continuous extension of 31 Keywords. In Table 8, “system“ is the response variable because the “system” node in Fig. 4 is the middle node but the “system” node gets nine direct arrows from “application”, “computer”, “data”, “electronic”, “information”, “memory”, “present”, “power” and “set” which are the predictor variables.

Table 8 shows that “application”, “computer”, “data”, “electronic”, “information”, “present”, “power” and “set” are statistically significant at the 5% significance level. In Table 9, “user” is the response variable because user node in Fig. 4 is the end node and the user node has nine direct arrows from “circuit”, “computer”, “content”, “information”, “interface”, “memory”, “network”, “object”, “portable”, “power” and “signal” which are the predictor variables.

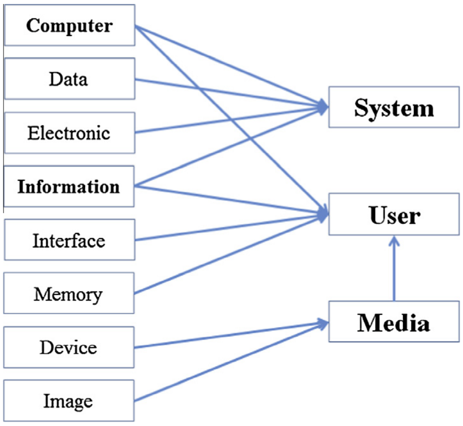

Fig. 6. Technology path diagram from copula modeling with 12 Keywords with 2000 patents.

Table 8
Estimates of parameters of system = application + computer + data + electronic + information + memory + present + power + set with 31 Keywords with 1000 patents.

Keyword	2.5‰	Median	97.5‰	Mean	St.D.	*t*-value
Application	0.06	0.15	0.22	0.15	0.04052	**3.70**
Computer	0.29	0.36	0.44	0.36	0.03875	**9.29**
Data	0.16	0.24	0.31	0.24	0.03923	**6.12**
Electronic	−0.26	−0.19	−0.11	−0.19	0.03751	**−5.07**
Information	0.08	0.17	0.24	0.17	0.04045	**4.20**
Memory	0	0.06	0.13	0.06	0.03788	1.58
Present	0.21	0.28	0.35	0.28	0.03696	**7.58**
Power	0.02	0.1	0.17	0.1	0.0391	**2.56**
Set	0.02	0.1	0.17	0.1	0.03782	**2.64**

Table 9
Estimates of parameters of user = circuit + computer + content + information + interface + memory + network + object + portable + power + signal with 31 Keywords with 1000 patents.

Keyword	2.5‰	Median	97.5‰	Mean	St.D.	*t*-value
Circuit	−0.19	−0.11	−0.03	−0.11	0.03902	**−2.82**
Computer	0.09	0.15	0.22	0.15	0.03603	**4.16**
Content	0.04	0.11	0.19	0.11	0.03967	**2.77**
Information	0.02	0.11	0.19	0.11	0.04183	**2.63**
Interface	0.75	0.82	0.91	0.82	0.04222	**19.42**
Memory	−0.21	−0.14	−0.07	−0.14	0.03814	**−3.67**
Network	−0.07	0	0.08	0	0.03849	0.00
Object	0.03	0.11	0.2	0.11	0.04143	**2.66**
Portable	−0.02	0.06	0.13	0.06	0.03817	1.57
Power	−0.16	−0.09	−0.02	−0.09	0.03634	**−2.48**
Signal	−0.13	−0.07	0.01	−0.06	0.03663	**−1.64**

Table 9 shows that “circuit”, “computer”, “content”, “information”, “interface”, “memory”, “object”, “power” and “signal” are statistically significant at the 5% significance level. In Table 10, “media” is the response variable because the “media” node in Fig. 4 is the middle node but media node has six direct arrows from “content”, “device”, “digital”, “display”, “image” and “information” which are the predictor variables.

Table 10 shows that “content”, “device”, “digital” and “information” are statistically significant at the 5% significance level. In Table 11, time is the response variable because time node in Fig. 4 is the middle node but the “time” node gets six direct arrows from “data”, “digital”, “information”, “network”, “portion” and “set” which are the predictor variables.

Table 11 shows that “data”, “information” and “network” are statistically significant at the 5% significance level. In Table 12, “video” is the response variable because the “video” node in Fig. 4 is the end node and the “video” node has five direct arrows from “digital”, “display”, “image”, “system” and “time” which are the predictor variables.

Table 12 shows that “digital”, “display” and “image” are statistically significant at the 5% significance level. In Table 13, “signal”

Table 10
Estimates of parameters of media = content + device + digital + display + image + information with 31 Keywords with 1000 patents.

Keyword	2.5‰	Median	97.5‰	Mean	St.D.	*t*-value
Content	0.08	0.14	0.21	0.14	0.03522	**3.98**
Device	0.08	0.15	0.21	0.15	0.0364	**4.12**
Digital	0.03	0.11	0.17	0.11	0.03768	**2.92**
Display	−0.1	−0.02	0.05	−0.02	0.03713	−0.54
Image	−0.09	−0.02	0.05	−0.02	0.0373	−0.54
Information	0.04	0.12	0.18	0.11	0.03655	**3.01**

Table 11
Estimates of parameters of time = data + digital + information + network + portion + set with 31 Keywords with 1000 patents.

Keyword	2.5‰	Median	97.5‰	Mean	St.D.	*t*-value
Data	0.07	0.15	0.22	0.15	0.03793	**3.95**
Digital	−0.05	0.03	0.1	0.03	0.03954	0.76
Information	0.02	0.1	0.17	0.1	0.03768	**2.65**
Network	0.01	0.09	0.16	0.09	0.03785	**2.38**
Portion	−0.09	−0.01	0.06	−0.01	0.03834	−0.26
Set	−0.06	0.02	0.1	0.02	0.04177	0.48

Table 12
Estimates of parameters of video = digital + display + image + system + time with 31 Keywords with 1000 patents.

Keyword	2.5‰	Median	97.5‰	Mean	St.D.	*t*-value
Digital	0.04	0.1	0.19	0.11	0.03926	**2.80**
Display	0.04	0.12	0.2	0.12	0.03966	**3.03**
Image	0.01	0.07	0.16	0.08	0.03696	**2.16**
System	0	0.07	0.14	0.07	0.03808	1.84
Time	-0.03	0.04	0.11	0.04	0.03682	1.09

Table 13
Estimates of parameters of signal = audio + circuit + control + digital + media + set with 31 Keywords with 1000 patents.

Keyword	2.5‰	Median	97.5‰	Mean	St.D.	*t*-value
Audio	−0.01	0.06	0.13	0.06	0.03796	1.58
Circuit	0.08	0.16	0.23	0.16	0.03961	**4.04**
Control	0.03	0.1	0.17	0.1	0.0376	**2.66**
Digital	−0.05	0.02	0.09	0.02	0.03783	0.53
Media	−0.13	−0.05	0.02	−0.05	0.03836	−1.30
Set	−0.06	0.02	0.1	0.02	0.04174	0.48

is the response variable because the "signal" node in Fig. 4 is the middle node but the "signal" node has six direct arrows from "audio", "circuit", "control", "digital", "media" and "set" which are the predictor variables.

Table 13 shows that "circuit" and "control" are statistically significant at the 5% significance level. In Table 14, "display" is the response variable because the "display" node in Fig. 4 is the middle node but the "display" node has five direct arrows from "audio", "circuit", "color", "computer", "data" and "device" which are the predictor variables.

Table 14 shows that "color", "computer" and "data" are statistically significant at the 5% significance level. Fig. 7 shows a technology path diagram from copula modeling with 31 Keywords with 1000 patents by the results of from Tables 8–14.

Using copula modeling with Apple keywords, we found core technologies for technology management of Apple. At first, the technologies of Apple were based on six keyword groups, which were "system", "user", "media", "time", "video", "display", and signal. Secondly the six keyword groups had their associated sub-keywords. For example, the keyword of "time" is influenced by keywords of "data", "information", and "network". In this paper, we assign keyword to technology. So each keyword represents technology based on it. Also the keyword of "content" is contained within "user" and "media" concurrently. This means that content based technology is mediated technology between technologies of user and media. In the next section, we perform another analysis for Apple keywords.

Table 14
Estimates of parameters of display = audio + circuit + color + computer + data + device with 31 Keywords with 1000 patents.

Keyword	2.5‰	Median	97.5‰	Mean	St.D.	*t*-value
Audio	−0.1	−0.03	0.06	−0.03	0.04192	−0.72
Circuit	−0.13	−0.05	0.02	−0.05	0.03762	−1.33
Color	0	0.08	0.14	0.08	0.03835	**2.09**
Computer	0.11	0.18	0.25	0.18	0.03872	**4.65**
Data	−0.22	−0.15	−0.08	−0.15	0.03635	**−4.13**
Device	0	0.07	0.15	0.07	0.03986	1.76

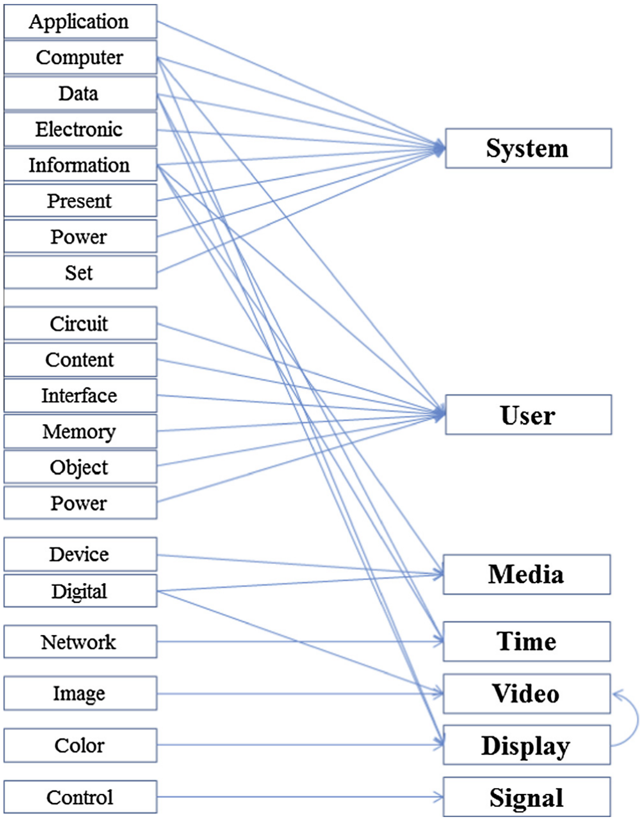

Fig. 7. Technology path diagram from copula modeling with 31 Keywords with 1000 patents.

5. Gaussian copula partial correlation with 12 Keywords with 2000 patents

Kim et al. [8] proposed the Gaussian copula partial correlation and applied the method to histone gene data. In this paper, we apply the Gaussian copula partial correlation to Dengue infection data to see the dependence structure. Given an n-dimensional distribution function F with continuous marginal (cumulative) distributions $F_1, \ldots, F_n$, there exists a unique n-copula C: $[0, 1]^n \rightarrow [0, 1]$ such that

$$F(x_1, \ldots, x_n) = C(F(x_1), \ldots, F(x_n)).$$

Suppose Y and Z are real-valued random variables with conditional distribution functions

$$F_{2|1}(y|x) = P(Y \leqslant y|X = x)$$

and

$$F_{3|1}(z|x) = P(Z \leqslant z|X = x)$$

Then the basic property of

$$U = F_{2|1}(Y|X) \text{ and } V = F_{3|1}(Z|X)$$

is as follows: suppose, for all x, $F_{2|1}(y|x)$ is continuous in y and $F_{3|1}(z|x)$ is continuous in z. Then U and V have uniform marginal distributions.

Please cite this article in press as: J.-M. Kim, S. Jun, Graphical causal inference and copula regression model for apple keywords by text mining, Adv. Eng. Informat. (2015), http://dx.doi.org/10.1016/j.aei.2015.10.001

Likewise, if $X_1, \ldots, X_n$ is a vector of n random variables with absolutely continuous multivariate distribution function F, then the n random variables

$$U_1 = F_1(X_1), U_2 = F_{2|1}(X_2|X_1), \cdots, U_n = F_{1|2,\ldots,n}(X_n|X_1, \ldots, X_{n-1})$$

are *i.i.d. U* (0, 1).

The conditional distribution of Z_1 given Z_2 is also normal with mean vector

$$\upsilon_1 = \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(z_2 - \mu_2)$$

and covariance matrix

$$Q_1 = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$$

It follows that the conditional density function $f_{1|2}(\cdot|z_2)$ of Z_1, when $Z_2 = z_2$, is specified at the point z_1 by the equation

$$f_{1|2}(z_1|z_2) = \frac{f(z_1, z_2)}{f(z_2)} = \left(\frac{1}{2\pi}\right)^{\frac{p}{2}} \sqrt{\frac{|\Sigma_{22}|}{|\Sigma|}} \exp\left\{-\frac{(z_1 - \upsilon_1)^T Q_1^{-1}(z_1 - \upsilon_1)}{2}\right\}$$

The cumulative distribution function is

$$f_{1|2}(z_1|z_2) = \int_{-\infty}^{z_p} \cdots \int_{-\infty}^{z_1} f_{1|2}(x_1|z_2)dx_1 \ldots dx_p \qquad (1)$$

where $z_1 = (z_1, \ldots, z_p)$ and $z_1, \ldots, z_p \in R$.

By using Eq. (1), we can derive the Gaussian conditional distributions, and then by using the CML method by Genest et al. [51] and the IFM method by Joe [13], we can estimate the Gaussian copula parameter, a n-th order conditional correlation, $\rho_{Y\,X|Z_1,Z_2,\ldots,Z_n}$, using the following:

$$F_{xy|z_1,\ldots,z_n}(Y, X|Z_1, \ldots, Z_n) = C^{Ga}\left(F_{x|z_1,\ldots,z_n}(X|Z_1, \ldots, Z_n), F_{y|z_1,\ldots,z_n}(Y|Z_1, \ldots, Z_n);\ \rho_{YX|Z_1,\ldots,Z_n}\right)$$

The following Tables are about Gaussian copula partial correlation between two keywords given the other variable(s). Based on Fig. 5, we made Gaussian copula partial correlations. Looking at Fig. 5, we can find a starting node and end node from DAG, so that we did partial correlation of starting node and end node given the nodes which are located in between starting node and end node. Table 15 shows Gaussian copula partial correlations conditioning on one keyword. Especially, the Gaussian copula partial correlation of system (end node) and computer (start node) given on data is 0.103 and the Gaussian copula partial correlation of system (end node) and computer (start node) given on electronic is 0.099.

From Table 15, we show top five paths of Gaussian copula partial correlation conditioning on one keyword in Fig. 8.

The technology of “computer” affects development of “system” technology via “data” technology. Other paths are explained like this. Table 16 shows Gaussian copula partial correlations conditioning on two keywords. Especially, the Gaussian copula partial correlation of system (end node) and computer (start node) given on E = (information, display) is 0.102, the Gaussian copula partial correlation of system (end node) and computer (start node) given on F = (media, display) is 0.104, the Gaussian copula partial correlation of system (end node) and computer (start node) given on I = (memory, display) is 0.103, and the Gaussian copula partial correlation of system (end node) and computer (start node) given on J = (memory, electronic) is 0.099.

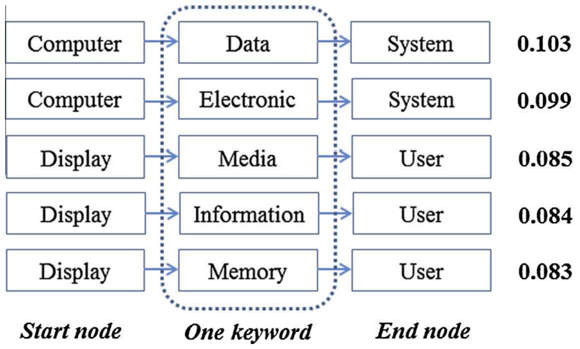

Fig. 8. Top five paths of Gaussian copula partial correlation conditioning on one keyword.

Table 15
Gaussian copula partial correlation conditioning on one keyword.

Keyword	Data	Display	Electronic	Image	Information	Media	Memory
System, computer	0.103	nep	0.099	nep	nep	nep	nep
System, data	nep	nep	0.080	nep	nep	nep	0.081
System, device	nep	nep	−0.060	nep	nep	−0.077	nep
System, display	nep	nep	nep	nep	nep	−0.007	−0.005
System, electronic	nep	nep	nep	nep	nep	nep	−0.067
System, image	nep	nep	nep	nep	nep	0.024	nep
System, information	nep	nep	nep	nep	nep	0.046	nep
User, computer	nep	nep	nep	nep	nep	nep	nep
User, data	nep	nep	nep	nep	nep	nep	−0.004
User, device	nep	nep	nep	nep	−0.004	−0.004	nep
User, display	nep	nep	nep	nep	**0.084**	**0.085**	**0.083**
User, electronic	nep	nep	nep	nep	nep	nep	0.014
User, image	nep	nep	nep	nep	nep	−0.022	nep
User, information	nep	nep	nep	nep	nep	0.030	nep
Media, computer	nep	−0.028	nep	nep	nep	nep	nep
Media, data	nep	0.007	nep	0.007	nep	nep	nep
Media, device	nep	0.035	nep	nep	0.035	nep	nep
Media, display	nep	nep	nep	0.006	0.006	nep	nep
Media, electronic	nep	nep	nep	0.005	nep	nep	nep
Memory, computer	0.011	0.012	0.012	nep	nep	nep	nep
Memory, data	nep	0.039	0.039	nep	nep	nep	nep
Memory, device	nep	−0.008	−0.004	nep	nep	nep	nep
Image, computer	0.005	0.004	0.003	nep	nep	nep	nep
Image, data	nep	0.009	0.006	nep	nep	nep	nep
Image, device	nep	−0.011	−0.003	nep	nep	nep	nep

nep = non-existent path.

Table 16
Gaussian copula partial correlation conditioning on two keywords.

Keyword	A	B	C	D	E
System, computer	Nep	nep	nep	nep	**0.102**
System, data	nep	nep	nep	nep	0.081
System, device	nep	nep	nep	nep	−0.079
System, display	nep	nep	nep	nep	nep
System, electronic	nep	nep	nep	nep	nep
User, computer	nep	nep	nep	nep	0.045
User, data	nep	nep	nep	nep	−0.006
User, device	nep	nep	nep	nep	−0.006
User, display	nep	nep	nep	nep	nep
User, electronic	nep	nep	nep	nep	nep
Media, computer	−0.028	−0.028	−0.028	nep	−0.029
Media, data	nep	nep	0.007	0.007	0.006
Media, device	nep	nep	0.035	nep	0.035
	F	G	H	I	J
System, computer	**0.104**	nep	nep	**0.103**	**0.099**
System, data	0.084	0.083	nep	0.082	0.077
System, device	−0.078	nep	−0.078	−0.077	−0.059
System, display	nep	−0.008	−0.007	nep	nep
System, electronic	nep	−0.067	nep	nep	nep
User, computer	0.047	nep	nep	nep	nep
User, data	−0.007	−0.007	nep	nep	nep
User, device	−0.006	nep	nep	nep	nep
User, display	nep	**0.085**	nep	nep	nep
User, electronic	nep	0.013	nep	nep	nep
Media, computer	nep	nep	nep	nep	nep
Media, data	nep	nep	nep	nep	nep
Media, device	nep	nep	nep	nep	nep

A = (display, data), *B* = (image, data), *C* = (image, display), D = (image, electronic), *E* = (information, display), *F* = (media, display), *G* = (media, image), *H* = (media, information), *I* = (memory, display), and *J* = (memory, electronic).
nep = non-existent path.

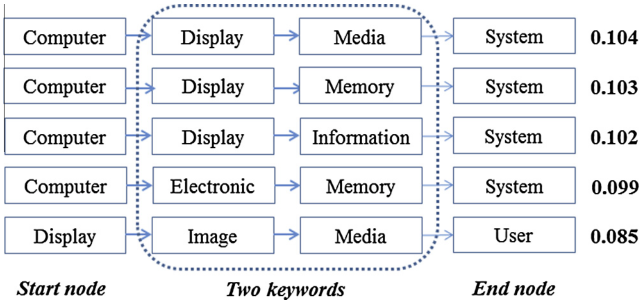

Fig. 9. Top five paths of Gaussian copula partial correlation conditioning on two keywords.

Like Table 15 and Fig. 8, we represent the top five paths of Gaussian copula partial correlation conditioning on two keywords in Fig. 9.

In Fig. 9, we knew there were diverse paths from computer to system. Every one of all paths is meaningful for Apple's technology management. Other paths except from computer to system are similar to this. Table 17 shows Gaussian copula partial correlations conditioning on three keywords. Gaussian copula partial correlation of media (end node) and computer (start node) given on image, electronic and data is −0.02746, Gaussian copula partial correlation of media (end node) and computer (start node) given on information, display and data is −0.02877, and Gaussian copula partial correlation of user (end node) and computer (start node) given on information, display and data is 0.0453.

According to the growing the number of keywords, the number of paths is decreased. Fig. 10 shows the highest paths of Gaussian copula partial correlation conditioning on three keywords.

Using the keyword paths in Figs. 8–10 from Tables 15–17, we can find meaningful associations for understanding Apple's technologies. Also these can be used for R&D planning of Apple.

Table 17
Gaussian copula partial correlation conditioning on three keywords.

Keyword	Image, electronic, data	Information, display, data
Media, computer	−0.02746	−0.02877
User, computer	nep	0.0453

nep = non-existent path.

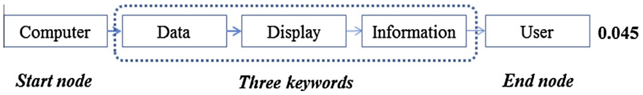

Fig. 10. Highest path of Gaussian copula partial correlation conditioning on three keywords.

6. Conclusions

In this paper, we were interested in Apple's technologies and its technological innovation. To understand and analyze Apple's technologies, we collected all patent documents applied by Apple in the world. We used Apple's keywords extracted from searched patent data, because the keywords contain technological aspects of Apple. From this study, we found that how the target response keyword and the predictor keywords were related to each other by copula modeling. Technologies of predictor keywords affect the technological developments of target response keywords. The associations between predictor and response keywords will provide novel information for Apple's R&D planning. By using Gaussian copula partial correlation, we also found the detailed dependence structure of Apple keywords. By performing these methods, this paper showed the technological trends and relations between

Please cite this article in press as: J.-M. Kim, S. Jun, Graphical causal inference and copula regression model for apple keywords by text mining, Adv. Eng. Informat. (2015), http://dx.doi.org/10.1016/j.aei.2015.10.001

Apples technologies. Also expert groups interesting in Apple's technologies can use our experimental results. This research contributes to efficient R&D planning such as intellectual property R&D strategy of a company as well as Apple.

The main goal of this paper is to find the technological relationship between Apple's holding technologies. To perform this goal, we used the copula model in this paper. So we did not consider other companies beyond Apple. Of course, we think that the comparison study between Apple and other companies such as Samsung or Google is very meaningful for understanding their technological competitions. This is another valuable research topic, and we will perform the research in our future work. If we select more detailed keywords in this paper, the implication of our research will be highly limited. Because the more specific the determined keywords are, the more limited the application scope is. To deciding the delicate and detailed keywords representing the Apple's technologies will be possible with the research and development members of Apple. The keywords of our paper are general, and also these can be the technological keywords for other companies such as Samsung or Google. But even if using the same keywords, the relationship structure between the keywords can be different according to the companies. For example, though the keyword of ''Application" affects the keyword of ''System" in Apple's technologies, in the case of Google, this result might not like Apple. Therefore in this paper, we suggested the result based on general keywords for understanding Apple's technologies. The effective use and application of our result are the role of the experts in Apple. Therefore, the patent analysis by detailed keywords is new challenge in partnership with Apple experts. Also, We focused on keyword analysis of Apple's patent documents. But there is so much information in patent documents such as citation, applied and issued date, claims, and IPC codes except keyword. In our future work, we will apply our analytical models to diverse information of patent. In addition, we will research advanced models for efficient patent analysis.

References

[1] Y. Cho, J. Hwang, D. Lee, Identification of effective opinion leaders in the diffusion of technological innovation: a social network approach, Technol. Forecast. Soc. Change 79 (2012) 97–106.

[2] D.L. Mann, Better technology forecasting using systemic innovation methods, Technol. Forecast. Soc. Change 70 (2003) 779–795.

[3] Y. Sun, Y. Lu, T. Wang, H. Ma, G. He, Pattern of patent-based environmental technology innovation in China, Technol. Forecast. Soc. Change 75 (2008) 1032–1042.

[4] S. Cesaratto, S. Mangano, G. Sirilli, The innovative behaviour of Italian firms: a survey on technological innovation and R&D, Scientometrics 21 (1) (1991) 115–141.

[5] A.J.C. Trappey, C.V. Trappey, C. Wu, C. Lin, A patent quality analysis for innovative technology and product development, Adv. Eng. Inform. 26 (2012) 26–34.

[6] D. Chen, W.C. Lin, M. Huang, Using essential patent index and essential technological strength to evaluate industrial technological innovation competitiveness, Scientometrics 71 (1) (2007) 101–116.

[7] S. Jun, S. Park, Examining technological innovation of Apple using patent analysis, Ind. Manage. Data Syst. 113 (6) (2013) 890–907.

[8] J.-M. Kim, Y.-S. Jung, T. Choi, E.A. Sungur, Partial correlation with copula modeling, Comput. Stat. Data Anal. 55 (3) (2011) 1357–1366.

[9] D. Kim, J.-M. Kim, Analysis of directional dependence using asymmetric copula-based regression models, J. Stat. Comput. Simul. 84 (9) (2014) 1990–2010.

[10] I. Kojadinovic, J. Yan, Modeling multivariate distributions with continuous margins using the copula R package, J. Stat. Softw. 34 (9) (2010) 1–20.

[11] J.-M. Kim, Y.-S. Jung, E.A. Sungur, K. Han, C. Park, I. Sohn, A copula method for modeling directional dependence of genes, BMC Bioinform. 9 (2008) 225.

[12] A. Sklar, Fonctions de repartition a *n*-dimensions et leurs marges, Inst. Stat. Univ. Paris 8 (1959) 229–231 (French).

[13] H. Joe, Multivariate Models and Dependence Concepts, Chapman & Hall, London, 1997.

[14] R. Nelsen, An Introduction to Copulas, second ed., Springer, New York, 2006.

[15] W.Y. Wang, Z. Hua, A semiparametric Gaussian copula regression model for predicting financial risks from earnings calls, in: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL 2014), 2014, pp. 1155–1165.

[16] P.D. Hoff, Extending the rank likelihood for semiparametric copula estimation, Ann. Appl. Stat. 1 (1) (2007) 265–283.

[17] M.W. Berry, J. Kogan, Text mining Applications and Theory, Wiley, 2010.

[18] I. Feinerer, A Text Mining Framework in R and its Applications, Dissertation, Department of Statistics and Mathematics, Vienna University of Economics and Business Administration, 2008.

[19] Y. Tseng, D. Juang, Y. Wang, C. Lin, Text mining for patent map analysis, in: Proceedings of IACIS Pacific Conference, 2005, pp. 1109–1116.

[20] Y. Tseng, C. Lin, Y. Lin, Text mining techniques for patent analysis, Inf. Proc. Manage. 43 (5) (2007) 1216–1247.

[21] P. Manrique, H. Qi, A. Morgenstern, N. Velasquez, T.C. Lu, N. Johnson, Context matters: improving the uses of big data for forecasting civil unrest: emerging phenomena and big data, in: Proc. of IEEE Int. Conf. on Intel. and Sec. Info. (ISI), 2013, pp. 169–172.

[22] R.Y. Zhong, G.Q. Huang, Q.Y. Dai, T. Zhang, Mining SOTs and dispatching rules from RFID-enabled real-time shopfloor production data, J. Intell. Manuf. 25 (4) (2014) 825–843.

[23] R. Kaur, M.M. Goyal, A survey on the different text data compression techniques, Int. J. Adv. Res. Comput. Eng. Technol. 2 (2) (2013) 711–714.

[24] D. Angus, S. Rintel, J. Wiles, Making sense of big text: a visual-first approach for analysing text data using Leximancer and Discursis, Int. J. Soc. Res. Method. 16 (3) (2013) 261–267.

[25] G. Jin, Y. Jeong, B. Yoon, Technology-driven roadmaps for identifying new product/market opportunities: use of text mining and quality function deployment, Adv. Eng. Inform. 29 (1) (2015) 126–138.

[26] C.V. Trappey, A.J.C. Trappey, H. Peng, L. Lin, T. Wang, A knowledge centric methodology for dental implant technology assessment using ontology based patent analysis and clinical meta-analysis, Adv. Eng. Inform. 28 (2) (2014) 153–165.

[27] A.J.C. Trappey, C.V. Trappey, C. Wu, C. Lin, A patent quality analysis for innovative technology and product development, Adv. Eng. Inform. 26 (1) (2012) 26–34.

[28] C.V. Trappey, H. Wu, F. Taghaboni-Dutta, A.J.C. Trappey, Using patent data for technology forecasting: China RFID patent analysis, Adv. Eng. Inform. 25 (1) (2011) 53–64.

[29] M. Bermudez-Edo, M. Noguera, N. Hurtado-Torres, M.V. Hurtado, J.L. Garrido, Analyzing a firm's international portfolio of technological knowledge: a declarative ontology-based OWL approach for patent documents, Adv. Eng. Inform. 27 (3) (2013) 358–365.

[30] A.J.C. Trappey, F.C. Hsu, C.V. Trappey, C.I. Lin, Development of a patent document classification and search platform using a back-propagation network, Exp. Syst. Appl. 31 (2006) 755–765.

[31] J. Choi, Y.S. Hwang, Patent keyword network analysis for improving technology development efficiency, Technol. Forecast. Soc. Change 83 (2014) 170–182.

[32] W. Glanzel, B. Thijs, Using 'core documents" for the representation of clusters and topics, Scientometrics 88 (2011) 297–309.

[33] N. Azam, J. Yao, Comparison of term frequency and document frequency based feature selection metrics in text categorization, Exp. Syst. Appl. 39 (2012) 4760–4768.

[34] Y.S. Lin, T.Y. Liao, S.J. Lee, Detecting near-duplicate documents using sentence-level features and supervised learning, Exp. Syst. Appl. 40 (2013) 1467–1476.

[35] X. Lin, Text-mining based journal splitting, in: Proceedings of Seventh International Conference on Document Analysis and Recognition, 2003, pp. 1075–1079.

[36] S. Vijayarani, M.M. Vinupriya, An efficient edge detection algorithm for facial images in image mining, Int. J. Eng. Sci. Res. Technol. 2 (10) (2013) 2880–2884.

[37] M.M. Bartere, P.R. Deshmukh, Cluster oriented image retrieval system, in: Proceedings of Emerging Trends in Computer Science and Information Technology (ETCSIT2012), 2012, pp. 25–27.

[38] The United States Patent and Trademark Office (USPTO). <http://www.uspto.gov/> (accessed on January 1, 2015).

[39] WIPS Corporation (WIPSON). <http://www.wipson.com/> (accessed on July 24, 2014).

[40] D. Hunt, L. Nguyen, M. Rodgers, Patent Searching Tools & Techniques, Wiley, New York, 2007.

[41] A.T. Roper, S.W. Cunningham, A.L. Porter, T.W. Mason, F.A. Rossini, J. Banks, Forecasting and Management of Technology, John Wiley & Sons, 2011.

[42] R Core Team, R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. <http://www.R-project.org/> (accessed on June 1, 2014).

[43] I. Feinerer, K. Hornik, Package 'tm' Ver. 0.6, Text Mining Package, CRAN of R project, 2014.

[44] P. Spirtes, C. Glymour, R. Scheines, Causation, Prediction, and Search: Adaptive Computation and Machine Learning, second ed., MIT Press, Cambridge, 2000.

[45] P. Spirtes, C. Meek, T.S. Richardson, An algorithm for causal inference in the presence of latent variables and selection bias, in: C. Glymour, G.F. Cooper (Eds.), Computation, Causation and Discovery, MIT Press, 1999, pp. 211–252.

Please cite this article in press as: J.-M. Kim, S. Jun, Graphical causal inference and copula regression model for apple keywords by text mining, Adv. Eng. Informat. (2015), http://dx.doi.org/10.1016/j.aei.2015.10.001

[46] D. Colombo, M.H. Maathuis, M. Kalisch, T.S. Richardson, Learning high-dimensional DAGs with latent and selection variables, Ann. Stat. 40 (1) (2012) 294–321.

[47] M. Kalisch, P. Buhlmann, Estimating high-dimensional directed acyclic graphs with the PC-Algorithm, J. Mach. Learn. Res. 8 (2007) 613–636.

[48] N. Harris, M. Drton, PC algorithm for nonparanormal graphical models, J. Mach. Learn. Res. 14 (2013) 3365–3383.

[49] M. Denuit, P. Lambert, Constraints on concordance measures in bivariate discrete data, J. Multivariate Anal. 93 (1) (2005) 40–57.

[50] L. Madsen, Y. Fang, Joint regression analysis for discrete longitudinal data, Biometrics 67 (3) (2011) 1171–1175.

[51] C. Genest, K. Ghoudi, L.P. Rivest, A semiparametric estimation procedure of dependence parameters in multivariate families of distributions, Biometrika 82 (3) (1995) 543–552.

Please cite this article in press as: J.-M. Kim, S. Jun, Graphical causal inference and copula regression model for apple keywords by text mining, Adv. Eng. Informat. (2015), http://dx.doi.org/10.1016/j.aei.2015.10.001