ビッグデータ時代のテキストマイニング
～マーケティング活用事例～

立教大学 経営学部 教授
佐々木宏

本日の内容

Ⅰ　トレンド：ビッグデータ

Ⅱ　ビッグデータをキーワードにしたテキストマイニングの事例

Ⅲ　テキストマイニングの効率化

本日のView Point

- デジタル・アカシックレコードの探索*
  - ライフログ：個人が単位
  - デジタル・アカシックレコード：社会が単位(人類の歴史がWeb上に瞬々刻々と刻み込まれている)
- Follows仮説
  - 投げ縄型曲線**
    - ある事象が普及するとき、別な普及曲線から前兆を知ることができる
  - スパイラル曲線
    - 2つの異なる位相をもつ波動がスパイラル型進化を起こす
      - 組織波動進化仮説：ゆらぎと組織パフォーマンスの関係***
    - 異なるビッグデータを同期化させる

*拙稿(2009)「ブログリサーチ」(同文舘)

**拙稿(2009)「ブログリサーチ」(同文舘),p.139

***拙稿(1993)「情報戦略と戦略策定組織のスパイラル進化－ゆらぎと波動変換の場の創造－」,情報システムフォーラム,No.376,pp.58-63,日本情報システム・ユーザ協会」)

Ⅰ　トレンド:ビッグデータ

データマイニングはどう変わったか。

1. データソース
2. データマイニングの着眼点
3. ツールやサービス
4. 学術的厳密性と実務的整合性
5. ビッグデータ・ニーズ

データマイニングはどう変わったか。
1. データソース

- データソース
  - 3V： ソースの量、多様性、速度
  - 4つ目のV：Veracity（正確さ）
- データのタイプ
  - トランザクション・データ、顧客データベース、テキスト・データ(VOC、ソーシャル・メディア)、Webログデータ、静止画、動画、音声、GPSデータ(位置・時刻・利用者情報)、スマートメーター(次世代電力計)からの収集データ、RFID、センサー　etc.
- データの所在
  - 個人：ライフログ
  - 業務：トランザクション・データ、Webログデータ、センサーデータ*など
  - 社会：**デジタル・アカシックレコード**（メール、SNS、ブログなど)

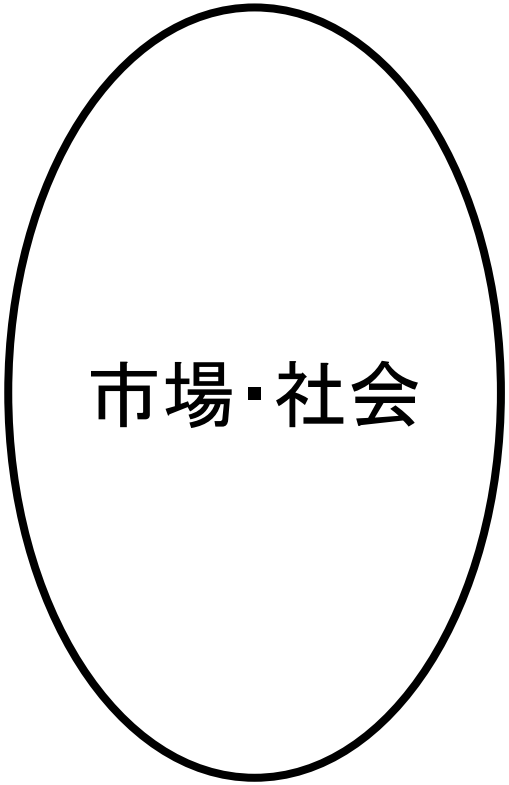
*移動車両や家電機器などに組み込まれたセンサーからの情報

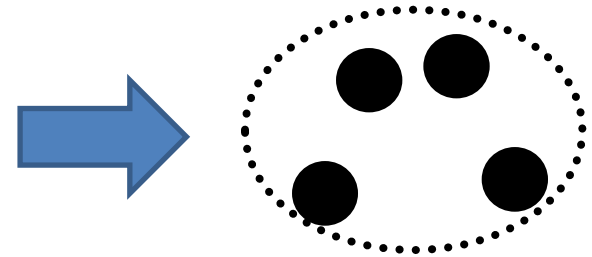
データマイニングはどう変わったか。
2. データマイニングの着眼点

- 空間軸
  - 関係性の構築:　点＝＞線＝＞面
- 時間軸
  - 離散＝＞連続
  - リアルタイム(分散処理)
  - 前兆と未来予測
- 多様なソースから関連性のあるデータの抽出
  - 同種ビッグ・データ内
  - 異種ビッグ・データ間の同期化
- 関係性(ネットワーク)の構築と分析

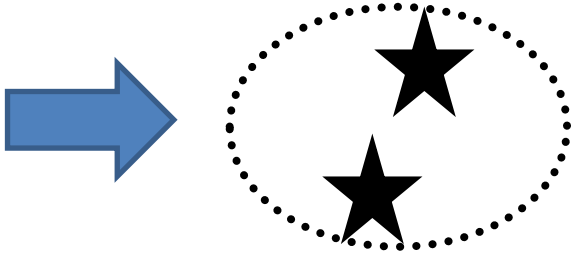
データマイニングはどう変わったか。
2. データマイニングの着眼点

平均値　vs　外れ値





統計的手法による母集団の推定
（基準：全体の傾向、期待値、有意確率・・・）



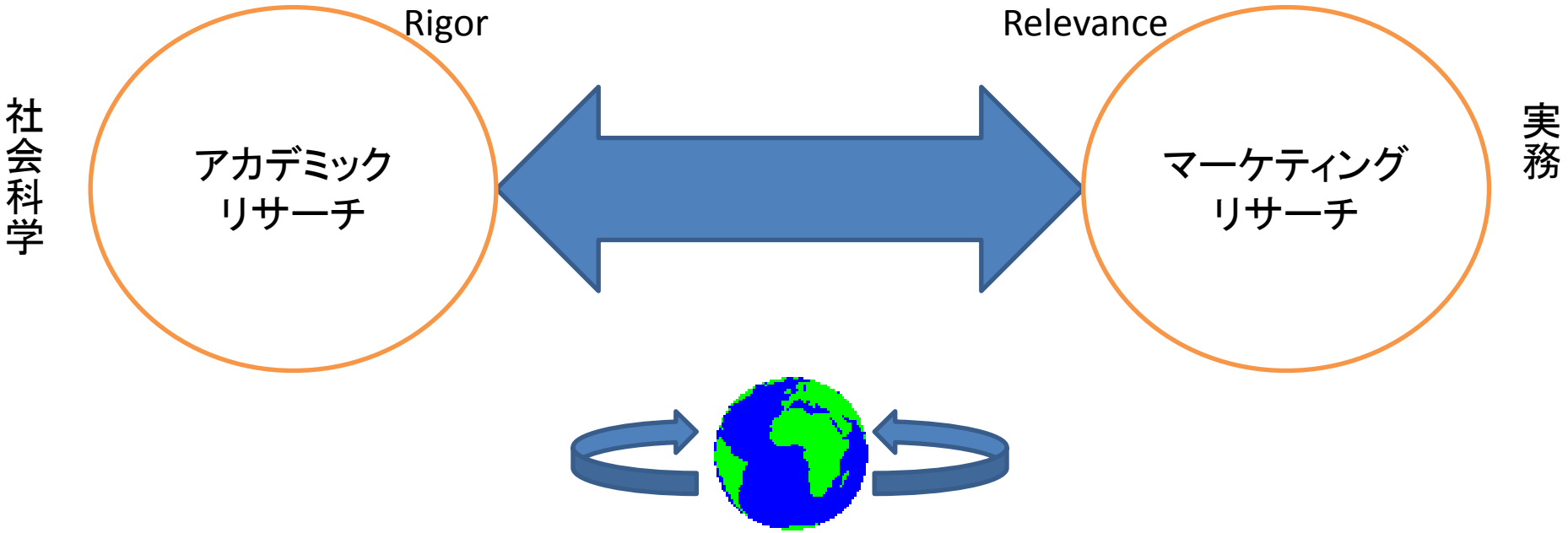
シグナルの発見
(基準：驚き、外れ値）

データマイニングはどう変わったか。
3. ツールやサービス

1. 新しいデータ処理プロセス
2. データの構造化
3. 既存手法とのリンケージ
4. ツールやソフトウェア
5. ITベンダー/インテグレータ

データマイニングはどう変わったか。
4. 学術的厳密性と実務的整合性

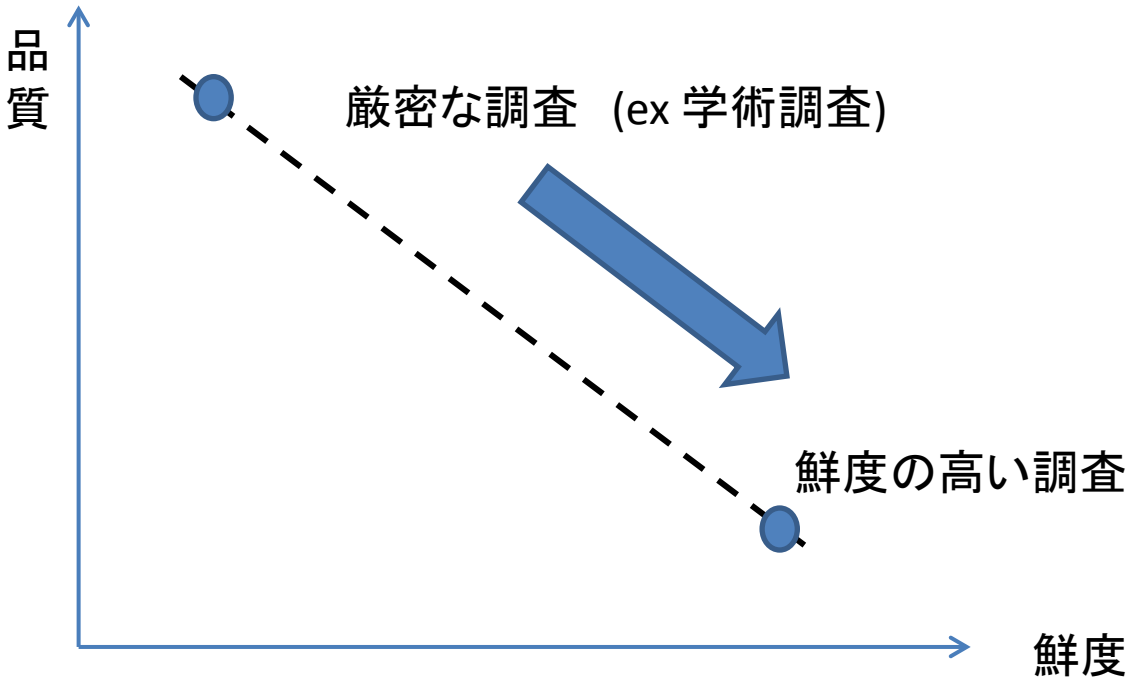
- リガー (学術的厳密性) vs. レリバンス(実務的整合性)*



*佐々木宏(2011),「リガー vs レリバンス　ーそのはざまで揺れ動く情報経営研究ー」,日本情報経営学会第62回全国大会統一論題セッション(神戸大学),ニューセオリーホライゾン(予稿集),pp.1-8.
H.Sasaki(2012),"IS research and its standpoint -Revisiting the "reference discipline" problem from Japan-", Tokyo Keizai University Information Systems Symposium 2012.

データマイニングはどう変わったか。
4. 学術的厳密性と実務的整合性

- （影響のない範囲で）厳密性を犠牲にして鮮度を優先
  - 不完全データの許容　ex)Web上のゴミ
  - 厳密な統計的有意性の検定はあまり意味をなさない
- 品質と鮮度のトレードオフ
  - 今何が起きているか、将来どうなるかについてヒントや確信を与える



データマイニングはどう変わったか。
5. ビッグデータ・ニーズ

ベンダー・サイド：下記を支援するところにビジネスチャンス

市場＼製品	既存	新規
既存	*市場浸透* 既存商品の売上向上を図りたい マーケティング(4P)に活用したい 顧客浸透と新規顧客を獲得に役立たせたい 自社、商品の評判の変化の兆しを即座に知りたい	*製品開発* 新商品のヒントが欲しい 世の中のトレンドを知りたい 競合他社の動向を知りたい
新規	*市場開発* 市場開発のためのマーケティング・リサーチ SNSなどを利用して、新しい市場・販路を拡大したい	*多角化* ビッグデータをからめた新しいビジネスモデルを開発したい SNSをプラットフォームにした事業者とコラボレートしたい

意思決定に役立たせる

顧客・社会とのインタラクション

アンゾフ(1965)「企業戦略論(Corporate Strategy)」の『成長ベクトル』を活用して作成

Ⅱ　ビッグデータをキーワードにした テキストマイニングの事例

1. ビッグデータ(新聞記事)
2. eWOM(学術誌)
3. 普及曲線とスパイラル曲線
4. 過去のIT関連ブーム

Key Question: 新聞記事から、ビッグデータについて何がわかるか？

「ビッグデータ」関連の新聞記事検索

- 2012/09/15　日経テレコン
  - キーワード:ビッグデータ
  - 全期間、全新聞を対象に抽出：314件
- 新聞記事の特徴
  - 日本語は完璧
  - 新聞記事に深いコンテキストは少ない(**次ページ参照**)
- ビッグデータの多様な表現:「BIGDATA」に統一
  - 膨大で雑多なデータの集合体「ビッグデータ」
  - 大量(の)データ「ビッグデータ」
  - 膨大な(量の)データ「ビッグデータ」
  - 爆発的に増大する企業内データ「ビッグデータ」
  - 爆発的に増える(コンピュータ)データ「ビッグデータ」
  - 爆発的に増え(続け)るデータ「ビッグデータ」
  - ビッグデータ(爆発的に増えているデータ)
  - バイト級の巨大なデータ(ビッグデータ)
  - コンピューター情報「ビッグデータ」
  - 膨大なデジタルデータの塊「ビッグデータ」

「ビッグデータ」関連の新聞記事検索

- 不要な単語を除外
  - (要旨を電子版に) (編集委員　XXXX) (執筆者名)など
- 類似語を統一
  - 企業名：日立製作所と日立、日本IBMとIBM　など
  - 外部記憶装置(ストレージ)とストレージ
  - ハードディスク駆動装置(HDD)と　　HDD
  - IT(情報技術)とIT
  - データベース(DB)とDB
  - 交流サイト(SNS)とSNS
  - スマートフォン(高機能携帯電話＝スマホ)とスマホ
  - 人工知能(AI)とAI
  - 基本ソフト(OS)「リナックス(Linux)」とLinax　など
- 単語フィルター：一般用語の「データ」を除く

単語頻度分析

名詞トップ30件

	単語	品詞	頻度
1	ＢＩＧＤＡＴＡ	名詞	247
2	データ	名詞	232
3	企業	名詞	167
4	活用	名詞	152
5	分析	名詞	149
6	膨大	名詞	144
7	開発	名詞	128
8	技術	名詞	114
9	情報	名詞	114
10	サービス	名詞	109
11	システム	名詞	108
12	提供	名詞	107
13	必要	名詞	107
14	今後	名詞	101
15	大量	名詞	95
16	ソフトウェア	名詞	93
17	顧客	名詞	93
18	発表	名詞	92
19	利用	名詞	90
20	従来	名詞	88
21	解析	名詞	87
22	日本	名詞	81
23	東京	名詞	80
24	サーバ	名詞	79
25	ＩＣＴ	名詞	77
26	クラウドコンピューティング	名詞	74
27	日本ＩＢＭ	名詞	72
28	富士通	名詞	71
29	同社	名詞	70
30	収集	名詞	69

係り受け頻度分析

- 話題一般
  - 頻度10回超

	係り元単語	係り元品詞	係り先単語	係り先品詞	頻度
1	ＢＩＧＤＡＴＡ	名詞	活用	名詞	52
2	データ	名詞	分析	名詞	48
3	爆発的	名詞	増える	動詞	48
4	ＢＩＧＤＡＴＡ	名詞	分析	名詞	28
5	データ	名詞	保存	名詞	25
6	ＢＩＧＤＡＴＡ	名詞	呼ぶ	動詞	24
7	サービス	名詞	提供	名詞	24
8	データ	名詞	解析	名詞	24
9	ＢＩＧＤＡＴＡ	名詞	解析	名詞	23
10	技術	名詞	開発	名詞	21
11	サービス	名詞	始める	動詞	20
12	データ	名詞	活用	名詞	20
13	注目	名詞	集める	動詞	20
14	システム	名詞	構築	名詞	18
15	動き	名詞	広がる	動詞	17
16	ＩＣＴ	名詞	活用	名詞	15
17	システム	名詞	開発	名詞	15
18	データ	名詞	集める	動詞	15
19	開発	名詞	発表	名詞	15
20	情報	名詞	分析	名詞	15
21	データ量	名詞	増える	動詞	14
22	日立製作所	名詞	発表	名詞	14
23	分析	名詞	役立てる	動詞	14
24	データ	名詞	処理	名詞	13
25	情報	名詞	収集	名詞	13
26	発売	名詞	発表	名詞	13
27	スマホ	名詞	普及	名詞	12
28	データ	名詞	収集	名詞	12
29	企業	名詞	蓄積	名詞	12
30	分析	名詞	生かす	動詞	12
31	ＢＩＧＤＡＴＡ	名詞	処理	名詞	11
32	ＩＣＴ	名詞	進化	名詞	11
33	クラウドコンピューティング	名詞	使う	動詞	11

注目語:ビッグデータ・キーワード、主要プレーヤー

〇キーワード

【共起ルール抽出】
10 回以上
【注目語を含む表現】
10 回以上

	単語	頻度
11	システム	108
24	サーバ	79
26	クラウドコンピューティング	74
36	スマホ	57
45	ストレージ	52
52	センサー	49
110	データセンター	30
154	Ｈａｄｏｏｐ	24
189	データベース	21
221	ＧＰＳ	19

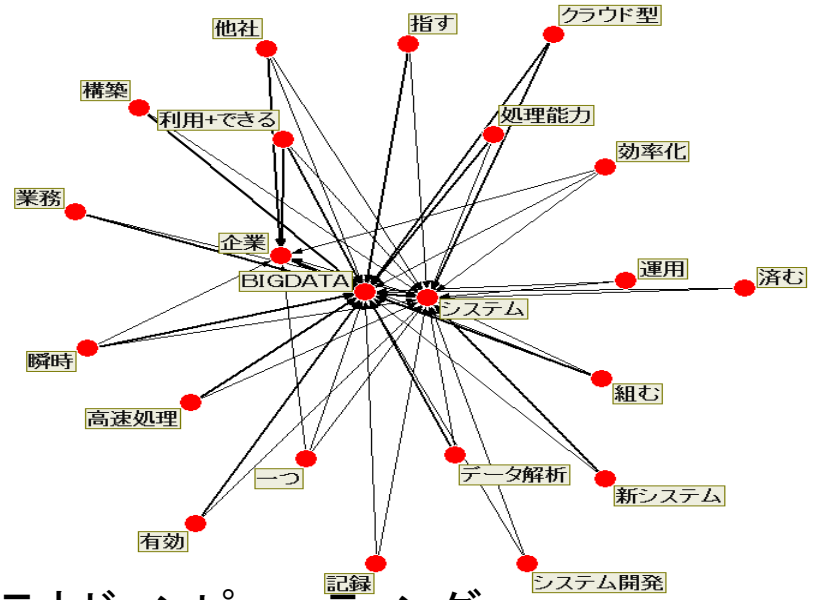
〇主要プレーヤー

【共起ルール抽出】
5 回以上
【注目語を含む表現】
5 回以上

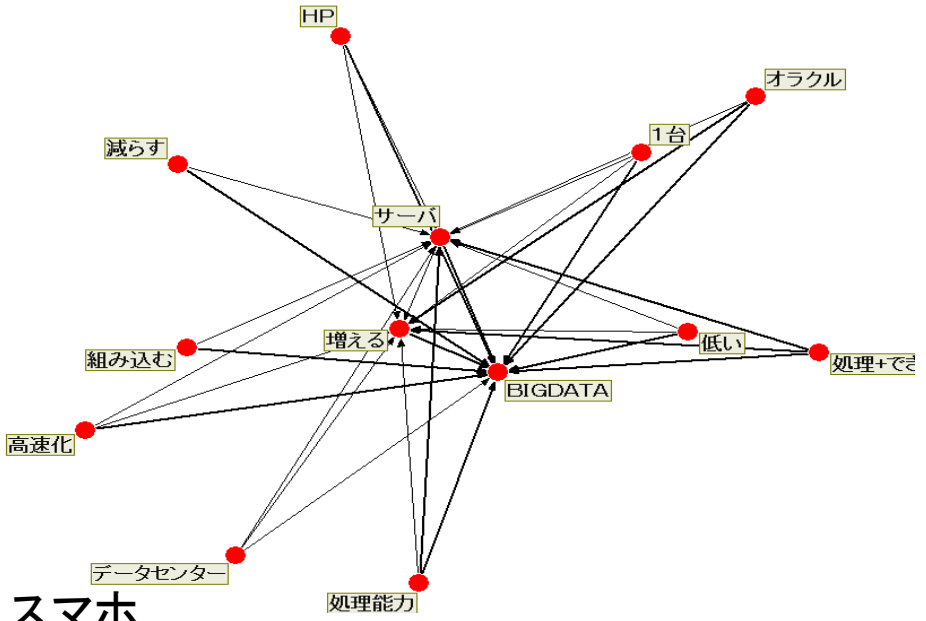
単語順位(名詞)	単語	出現頻度
27	日本ＩＢＭ	72
28	富士通	71
58	日立製作所	46
80	ＮＥＣ	37
145	オラクル	25
174	ＥＭＣ	22
200	NTTデータ	20
222	SAP	19

注目語：システム、サーバ、クラウドコンピューティング、スマホ

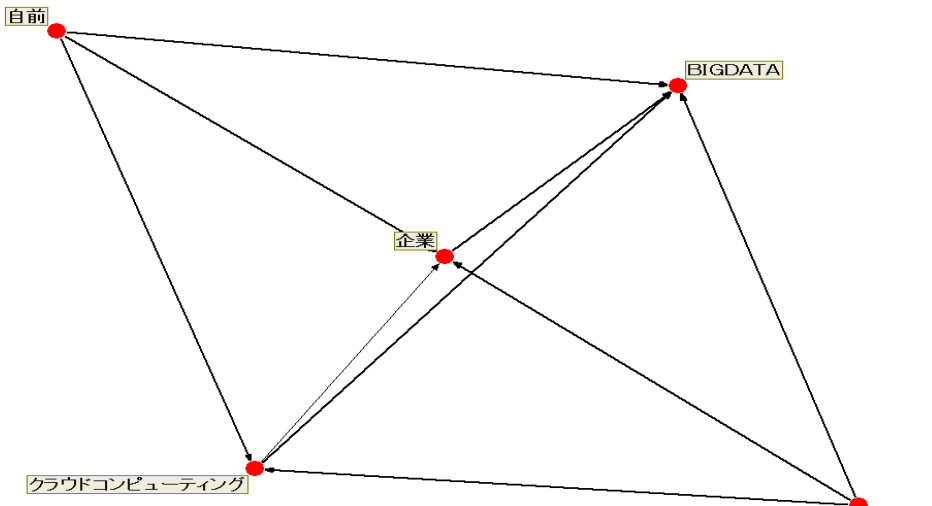
システム



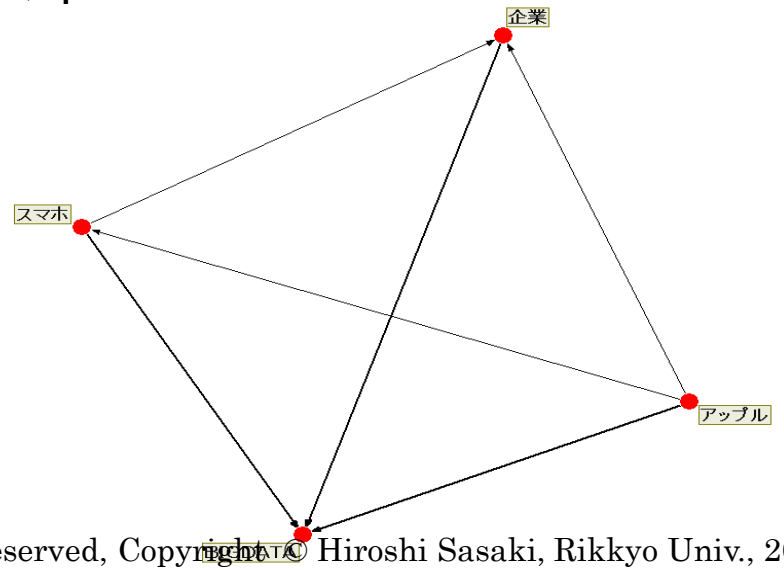
サーバ



クラウドコンピューティング

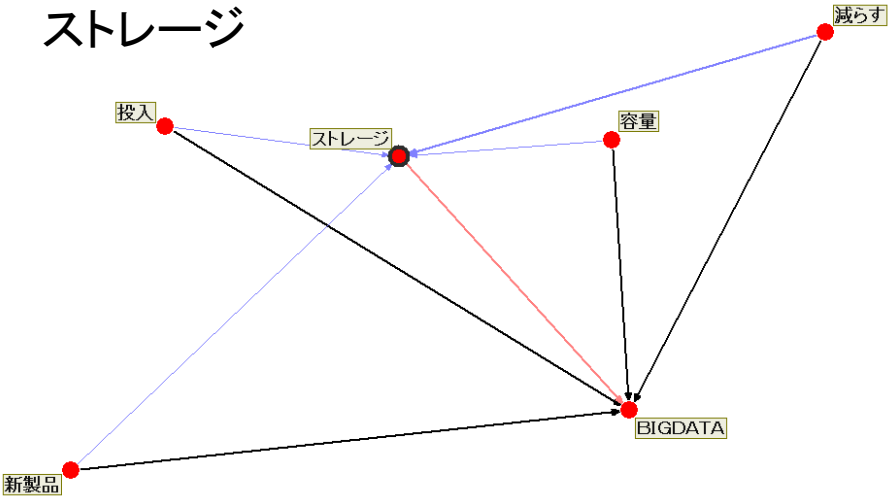


スマホ

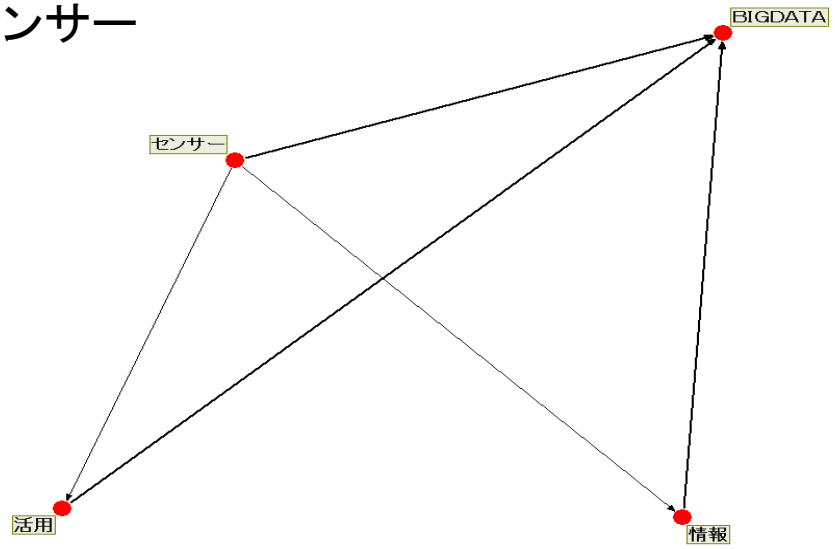


注目語：ストレージ、センサー、データセンター、Hadoop

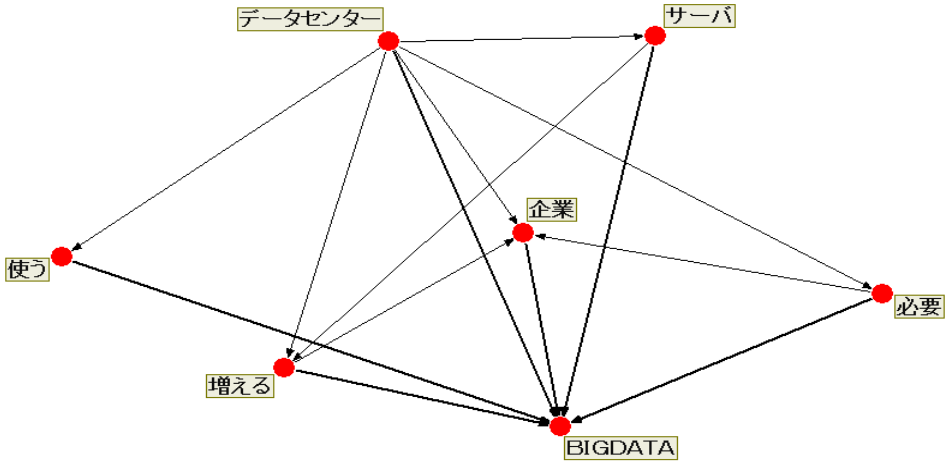
ストレージ



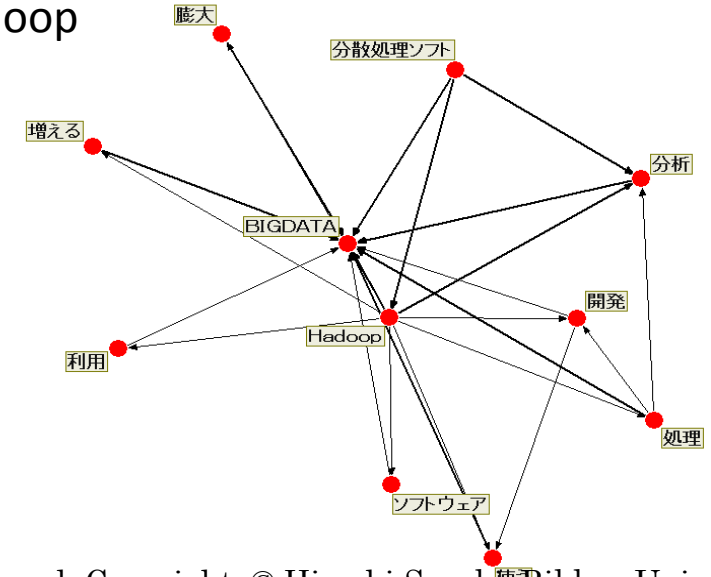
センサー



データセンター

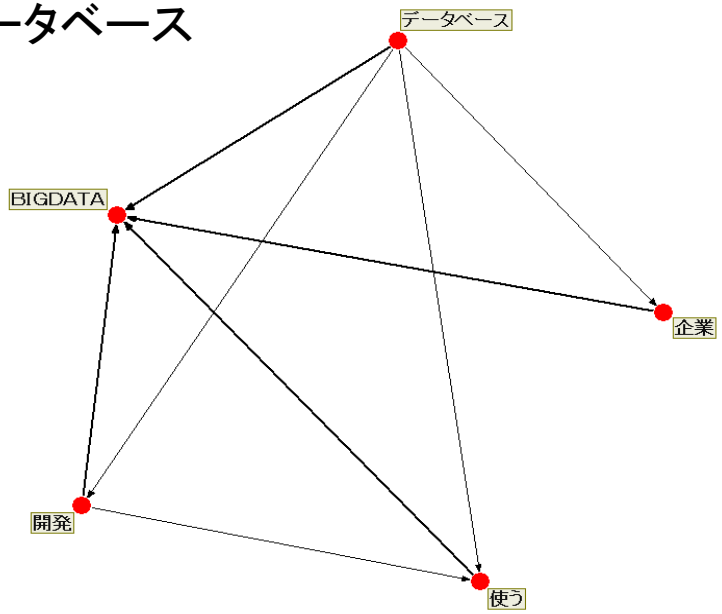


Hadoop

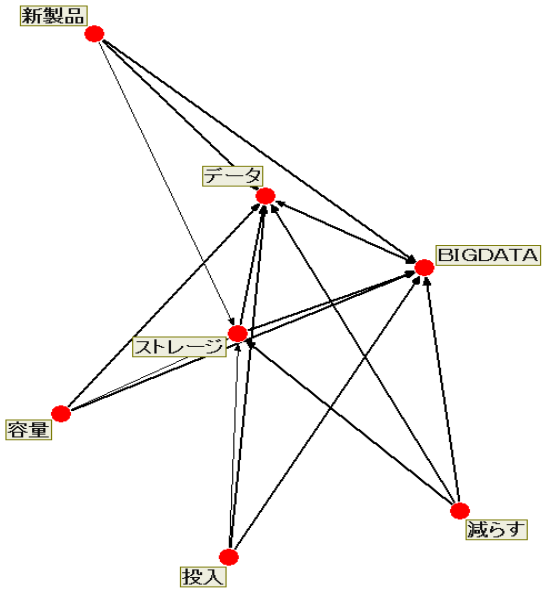


注目語：データベース、GPS

データベース

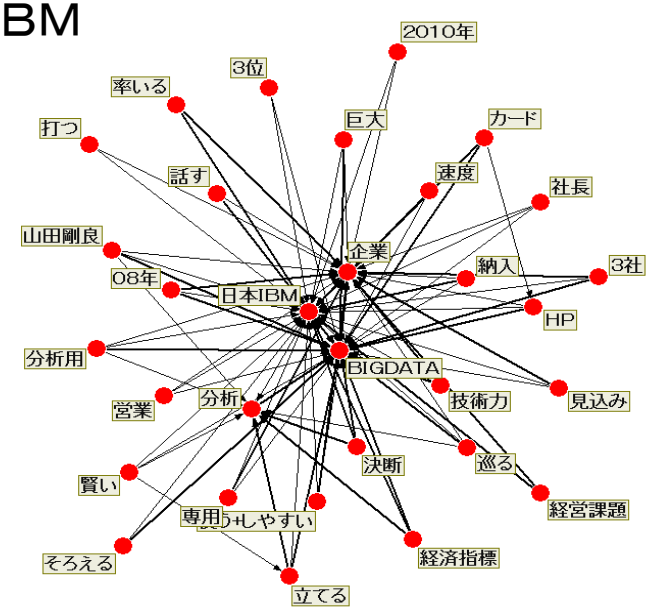


GPS

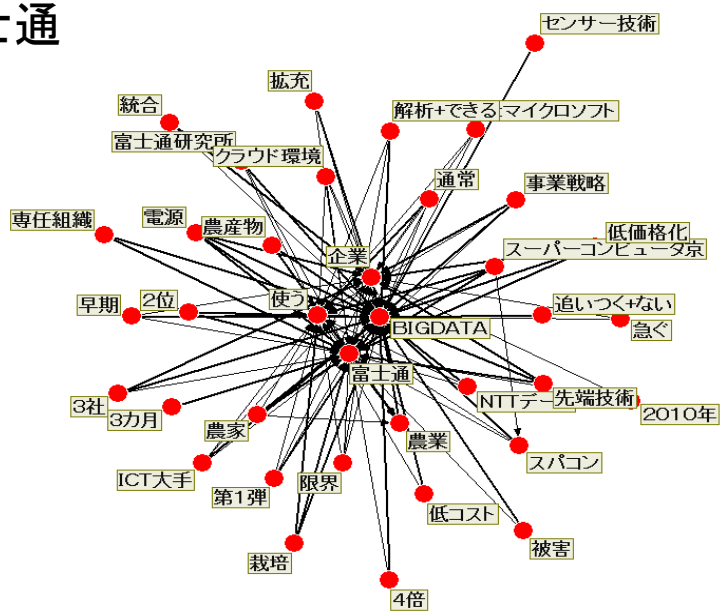


注目語:主要プレーヤー

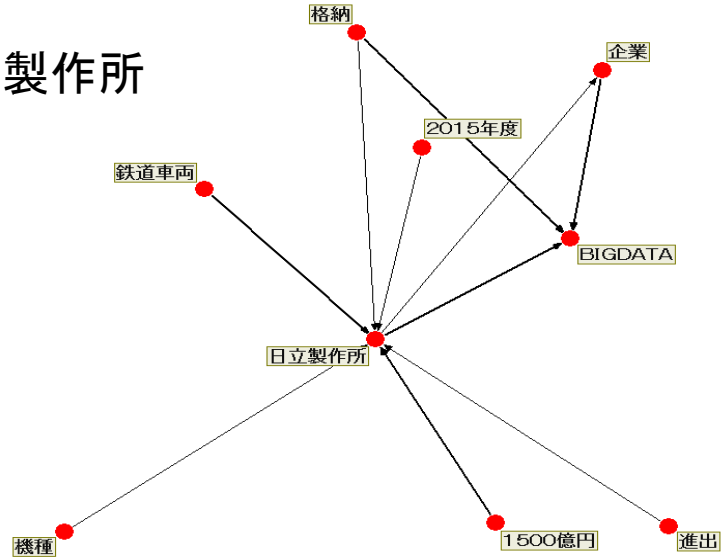
日本IBM



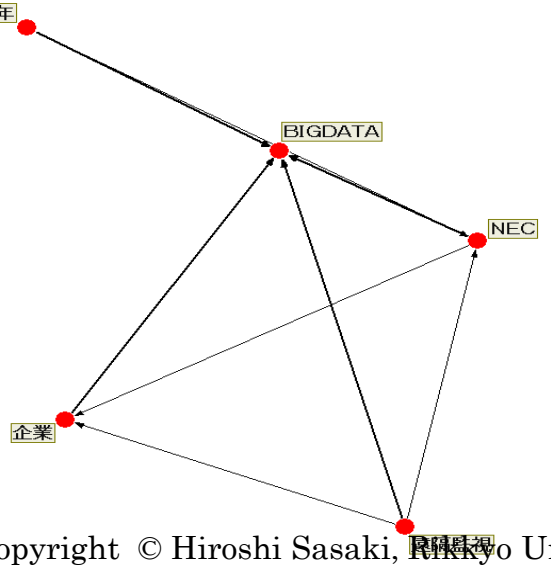
富士通



日立製作所

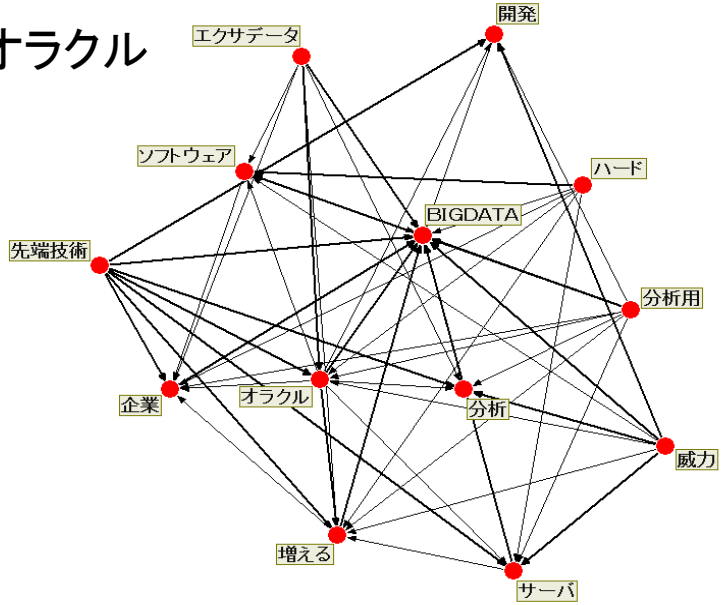


NEC

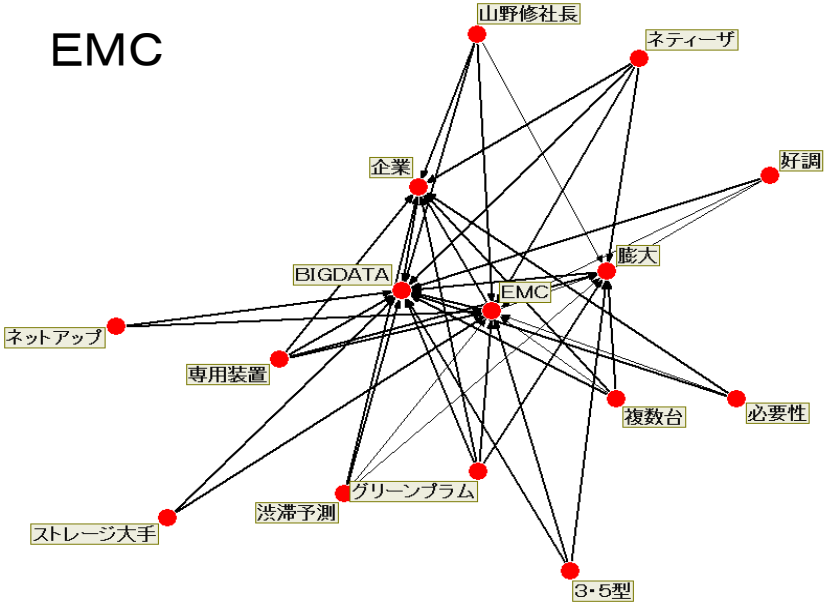


注目語:主要プレーヤー

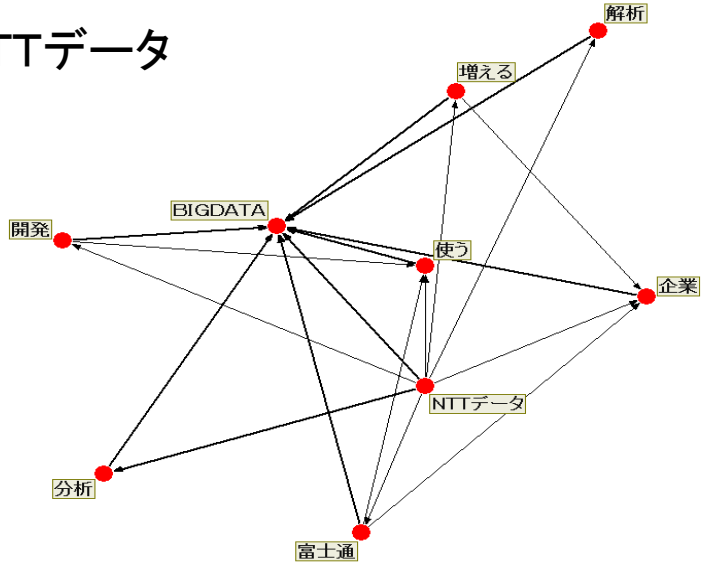
オラクル



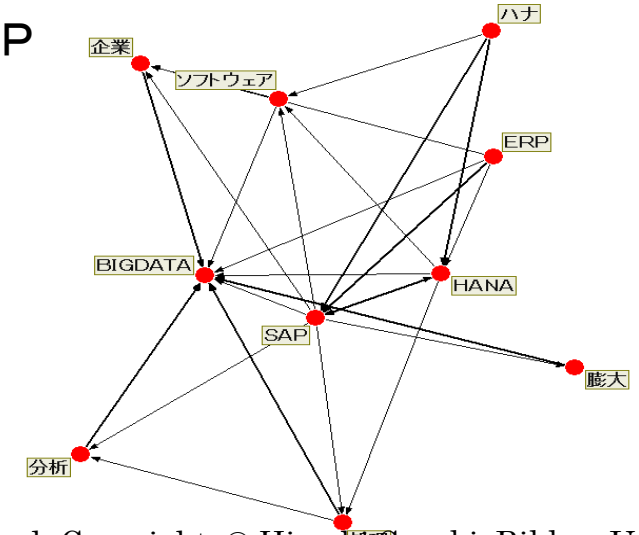
EMC



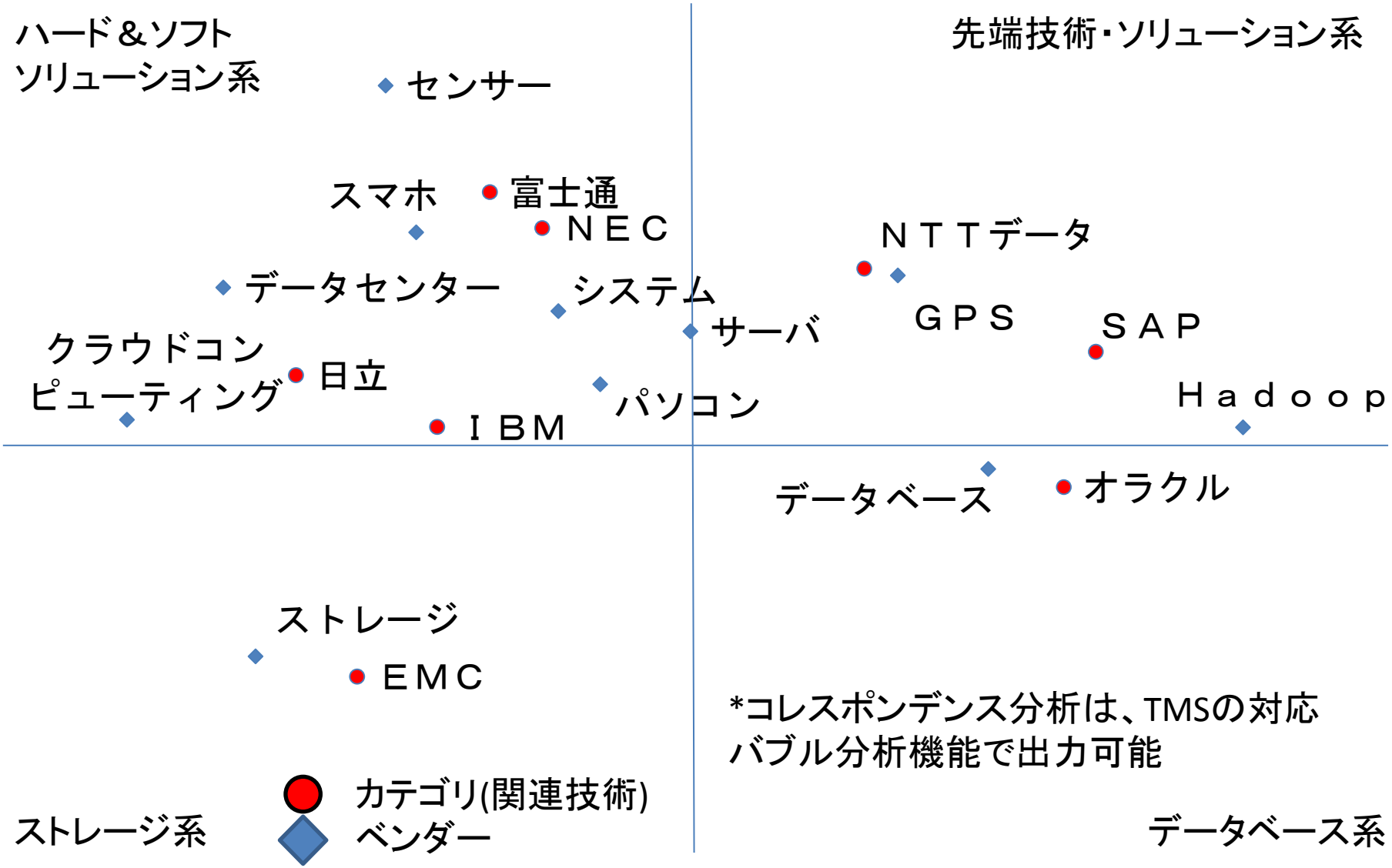
NTTデータ



SAP



コレスポンデンス分析結果*
ビッグデータ・キーワード＆主要プレーヤー



Ⅱ　ビッグデータをキーワードにした テキストマイニングの事例

1. ビッグデータ(新聞記事)
2. eWOM(学術誌)
3. 普及曲線とスパイラル曲線
4. 過去のIT関連ブームとの比較

Key Question: ビッグデータのコアとなる理論は何か？
(eWOM関連リサーチの検証)

ここでちょっとアカデミック！ WOM研究からeWOM研究へ

- 2012/06/09　EBSCOデータベース(世界中の文献データを収集)からWOMとeWOMの文献(査読付き学術誌)をすべて抽出する
  - WOM：250件
  - eWOM：17件

- eWOMで参照の多い著者はだれかを特定する(サイテーション部分のテキストマイニング)

eWOM文献から 参照の多かったもの(5件超)

著者	Reference文献数
Hennig-thurau	15
Dellarocas	13
Kats Elihu	6
Rogers	6
Senecal	6
Hung	6
Godes	6
Brown	6

1. Hennig-Thurau et al. (2004) ： eWOMのことばを定着させた。‘any positive or negative statement made by potential, actual, or former customers about a product or company, which is made available to a multitude of people and institutions via the Internet’.
2. Katz &Lazarsfeld (1955)：WOMのことばを定着させた。’WOM is defined as the act of exchanging marketing information among consumers, and plays an essential role in changing consumer attitudes and behaviour towards products and services’.
3. Dellarocas(2003)：従来のWOMとeWOMの相違点を分析した。
4. **Rogers(1983)：WOMの普及。普及曲線のS-shapedカーブ。**
5. その他：まだ参照は少ないが、Breazeale(2009): eWOM研究の包括的レビューした。

WOM：250件＋eWOM：7件の文献のAbstractを分析

名詞頻度：マーケティング関連の単語がかなり多い

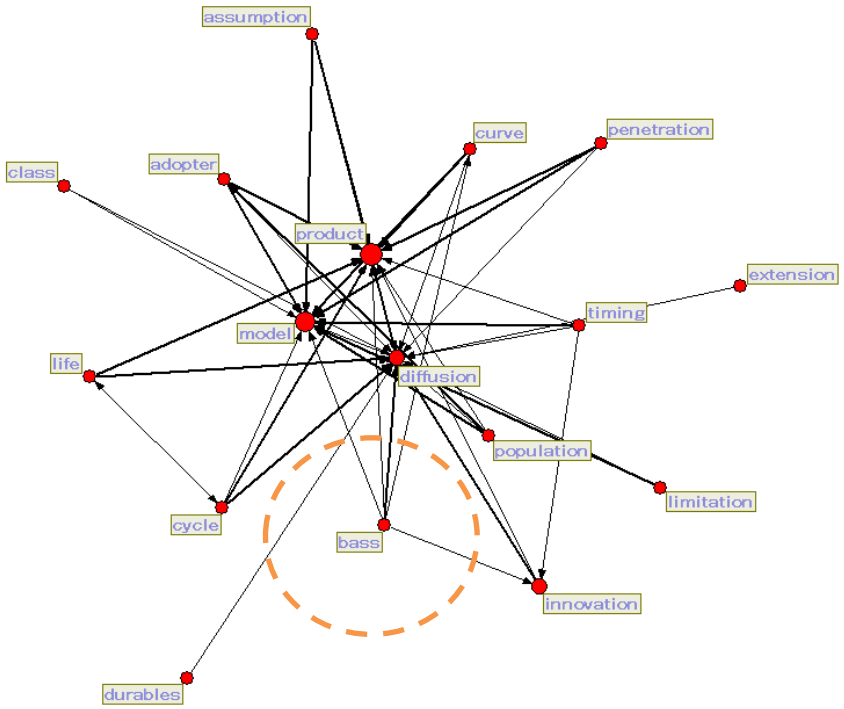
単語	頻度
consumer	116
study	95
article	91
product	91
communication	82
information	82
mouth	76
research	74
marketing	73
advertising	68
result	68
model	66
word	65
customer	59
service	53
effects	51
process	50
influence	46
implication	44
behavior	43
factor	43
paper	42
market	40
purchase	40

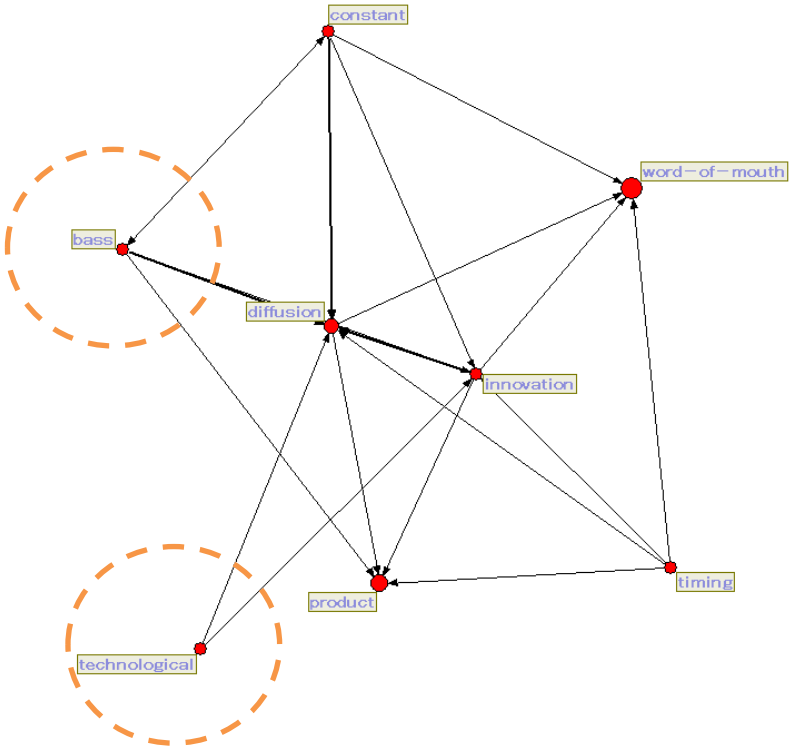
単語	頻度
source	40
brand	38
one	38
decision	37
datum	36
response	35
two	35
effect	34
findings	34
strategy	34
experience	33
role	33
sales	33
time	33
company	32
firm	32
present	32
relationship	32
services	32
diffusion	31
impact	31
satisfaction	31
analysis	30

WOM：250件＋eWOM：7件の文献のAbstractを分析

注目語： Diffusion, Innovation

- 条件：名詞、共起、3回以上

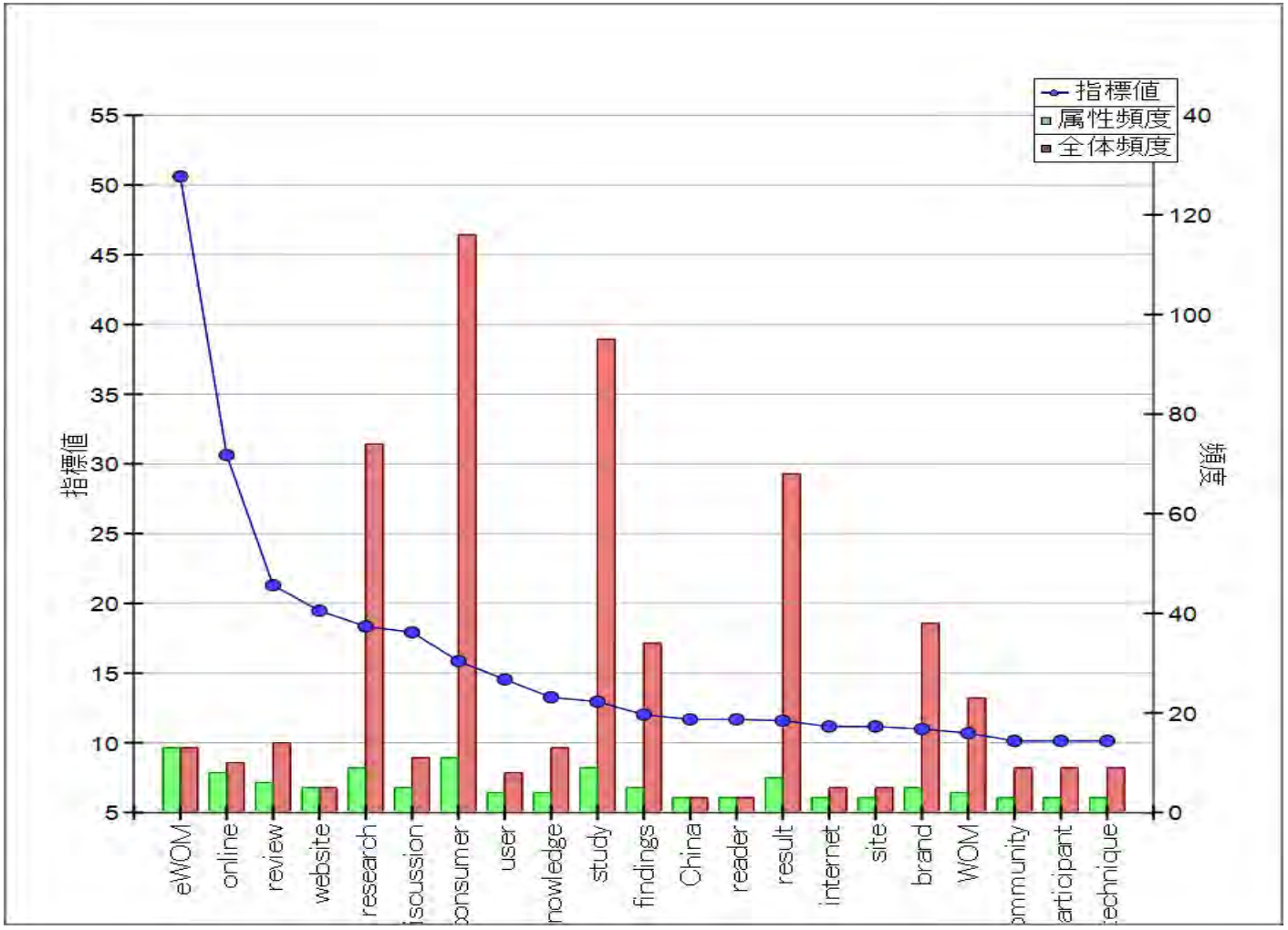




WOM：250件＋eWOM：7件の文献のAbstractを分析

eWOMの特徴語

- 条件：名詞



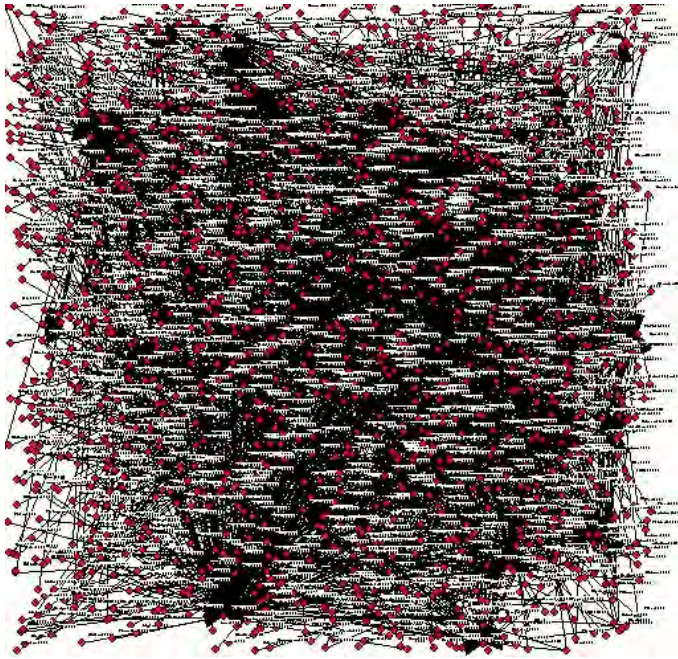
サイテーション分析と実務への応用

- アカデミックな論文のサイテーション分析
  - ナレッジチェーンの把握が目的（下図参照*）
  - 分析手法
    - 有向グラフ(directed graph)
    - 社会ネットワーク分析
- 実務への応用例
  - 普及（流行）の元を辿る

例）ビッグデータと4V

大量=Volume、多様=Variety、速度=Velocity

正確さ=Veracity　だれがほんとの言いだしっぺ？



*拙稿(2007)「入門ビジネスリーダーシップ第14章グローバル・リーダーシップと研究ストリーム,日本評論社,pp.267-286

Ⅱ　ビッグデータをキーワードにした テキストマイニングの事例

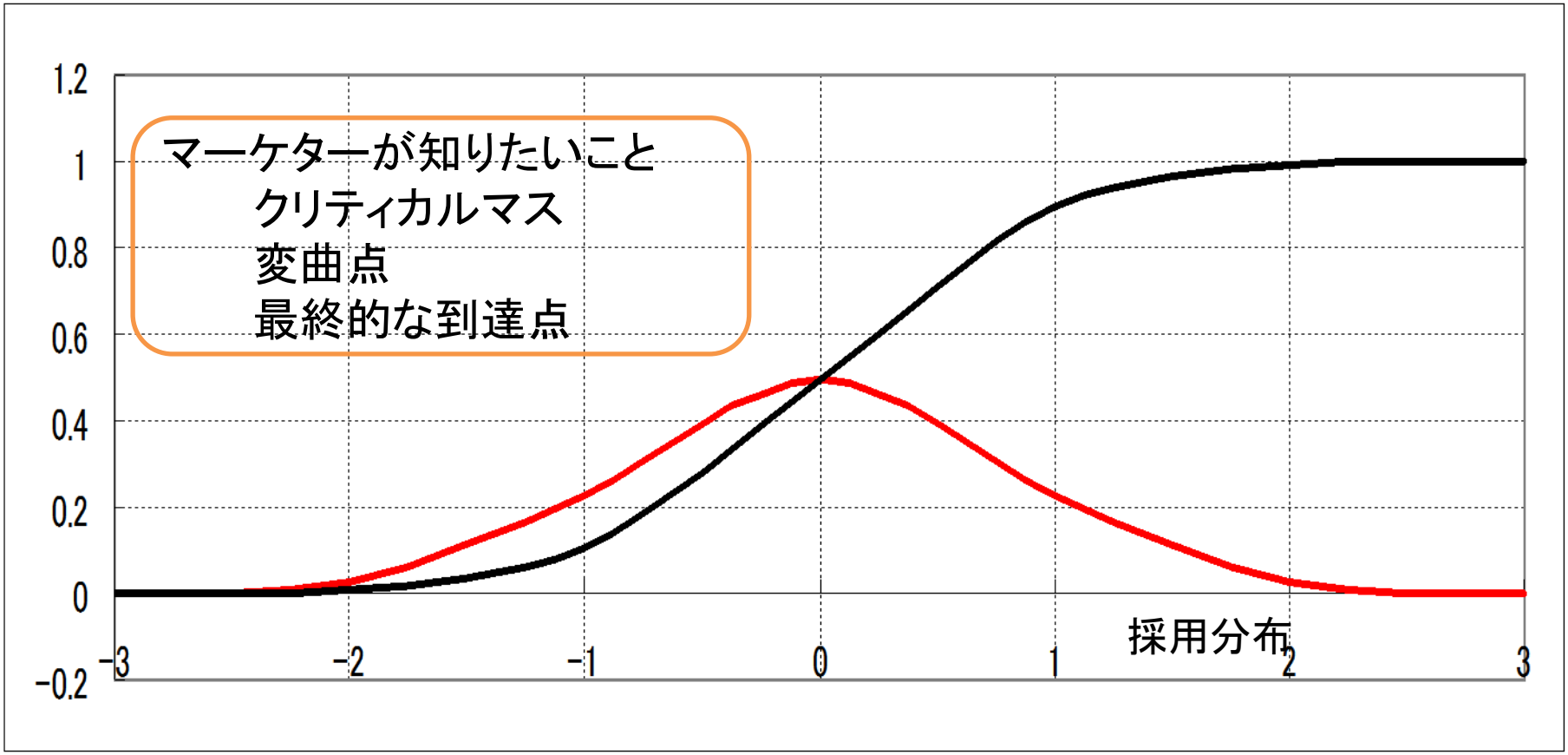
1. ビッグデータ(新聞記事)
2. eWOM(学術誌)
3. 普及曲線とスパイラル曲線
4. 過去のIT関連ブームとの比較

Key Question:　S字カーブとスパイラル曲線はどのように出現するか？　Follow仮説の提示と検証例を示します。

S字型曲線(S-Shaped Curve)

Rogers

- S字型曲線を作るモデル
  - 正規分布に基づくS字カーブ、ロジスティック・モデル、Bassモデル、・・・



正規分布の場合： σ:標準偏差=0.8, μ:平均値=0

S字型曲線とスパイラル曲線

- 流行や普及をどこで感知するか
  - Googleなどの検索キー
  - 新聞記事
  - SNSなどのeWOM
- eWOMから連続した曲線を構成
  - S字型曲線：時間軸を動かし、eWOMデータから曲線(波動)を構成する*

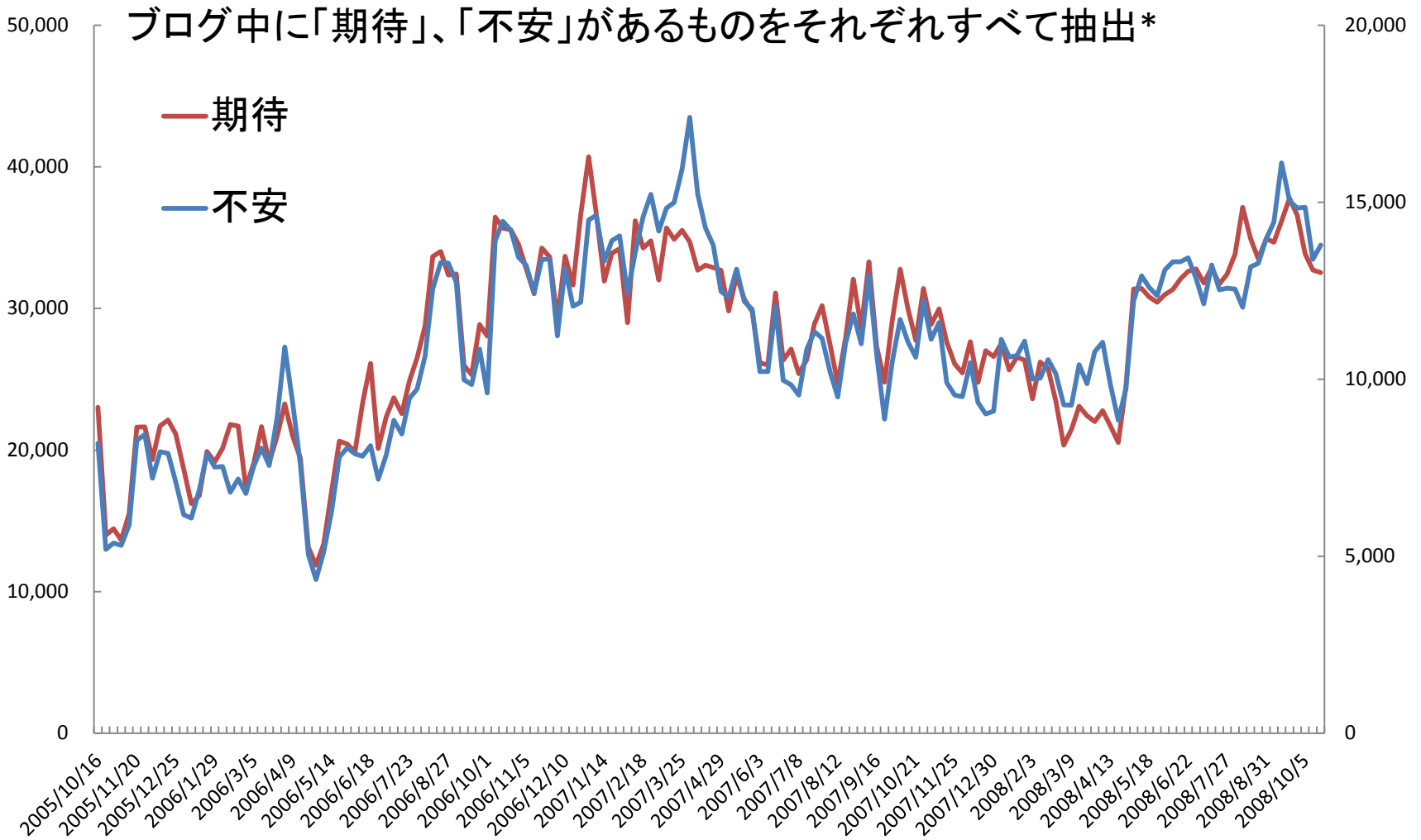
    *検索キーやマーケットバスケット分析(アソシエーション・ルールの導出)では時間軸固定で同時出現する単語、商品に注目している
  - スパイラル曲線：2つの位相のずれた波動の重なりで出現
    - 事業戦略と組織
    - マーケティング活動と成果
      - CM、プロモーション、メディア発表
    - eWOMと販売

Follows仮説

Follows仮説
X follows Y by Z

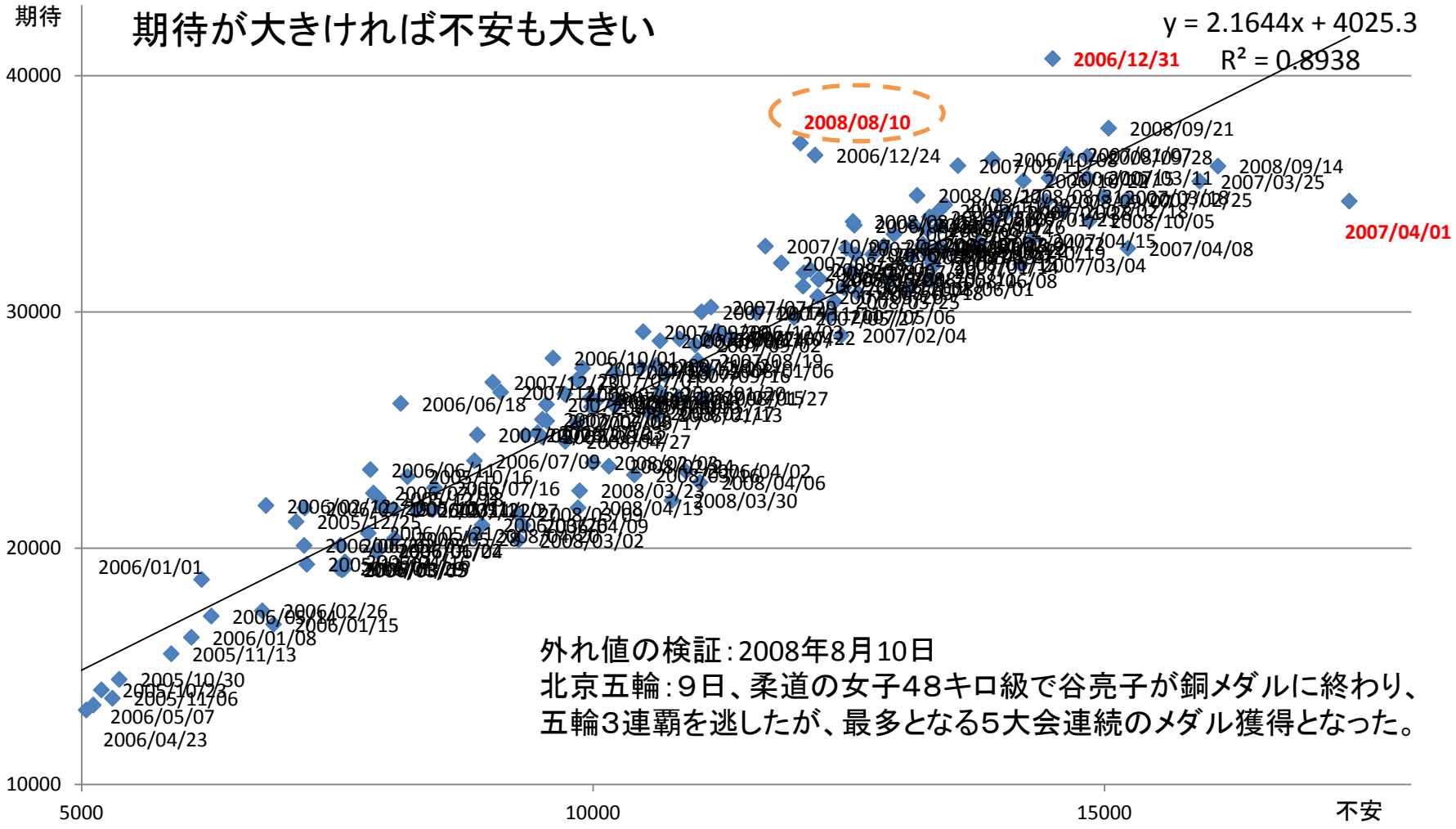
- XがYの動きに連動し、位相のずれた普及曲線(S-shaped curve)を描くことを予想し、検証する
  - Y: Y_1, Y_2, ・・・Y_nに拡張可能
  - Z: 時間のずれ(オプション)
- XとY、2つの位相のずれた波動の動きが継続すればスパイラル曲線が出現する

波動の位相が同期化している例

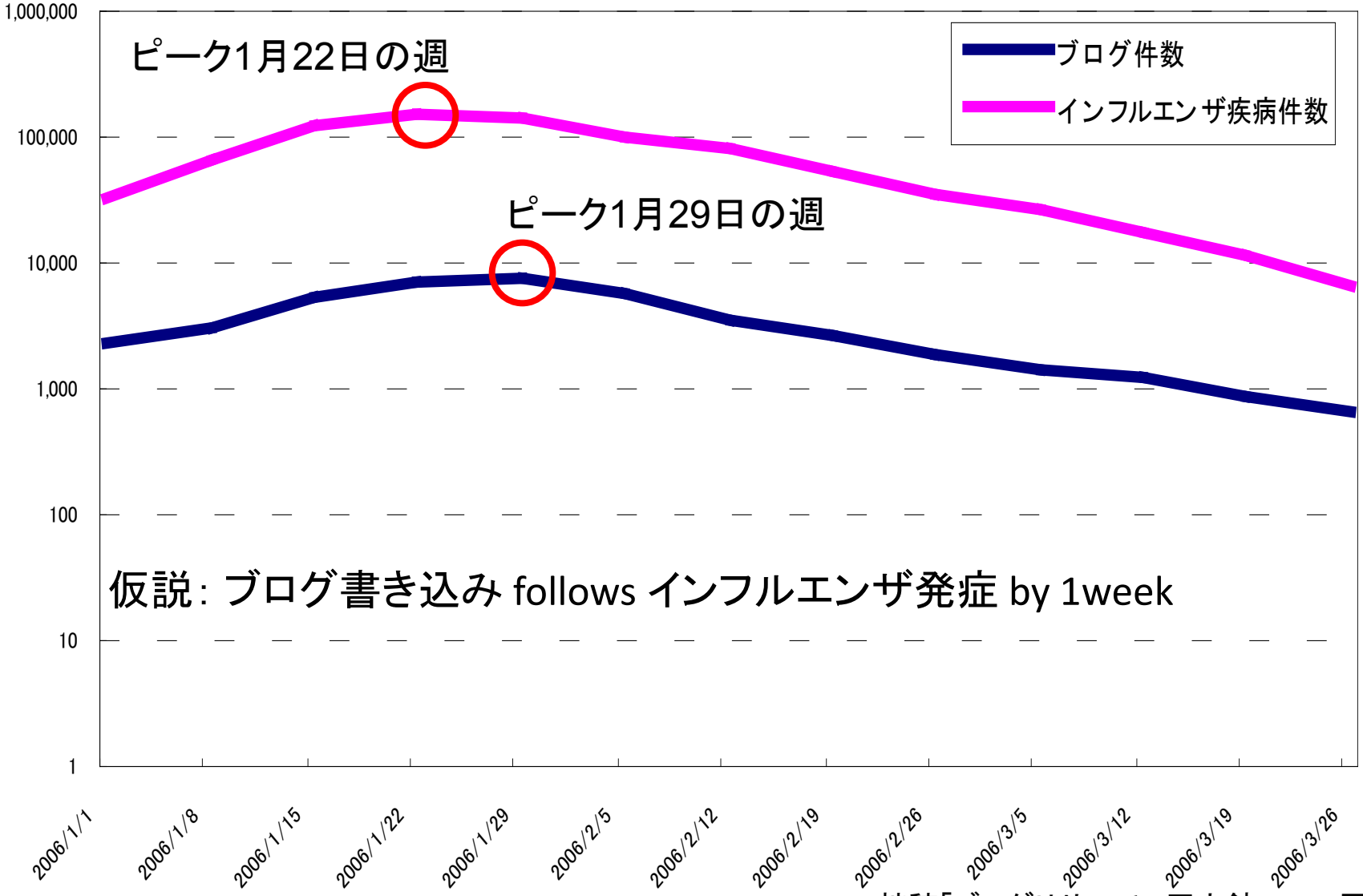


*拙稿「ブログリサーチ」,同文舘,p.27,図2-4

波動の位相が同期化している例

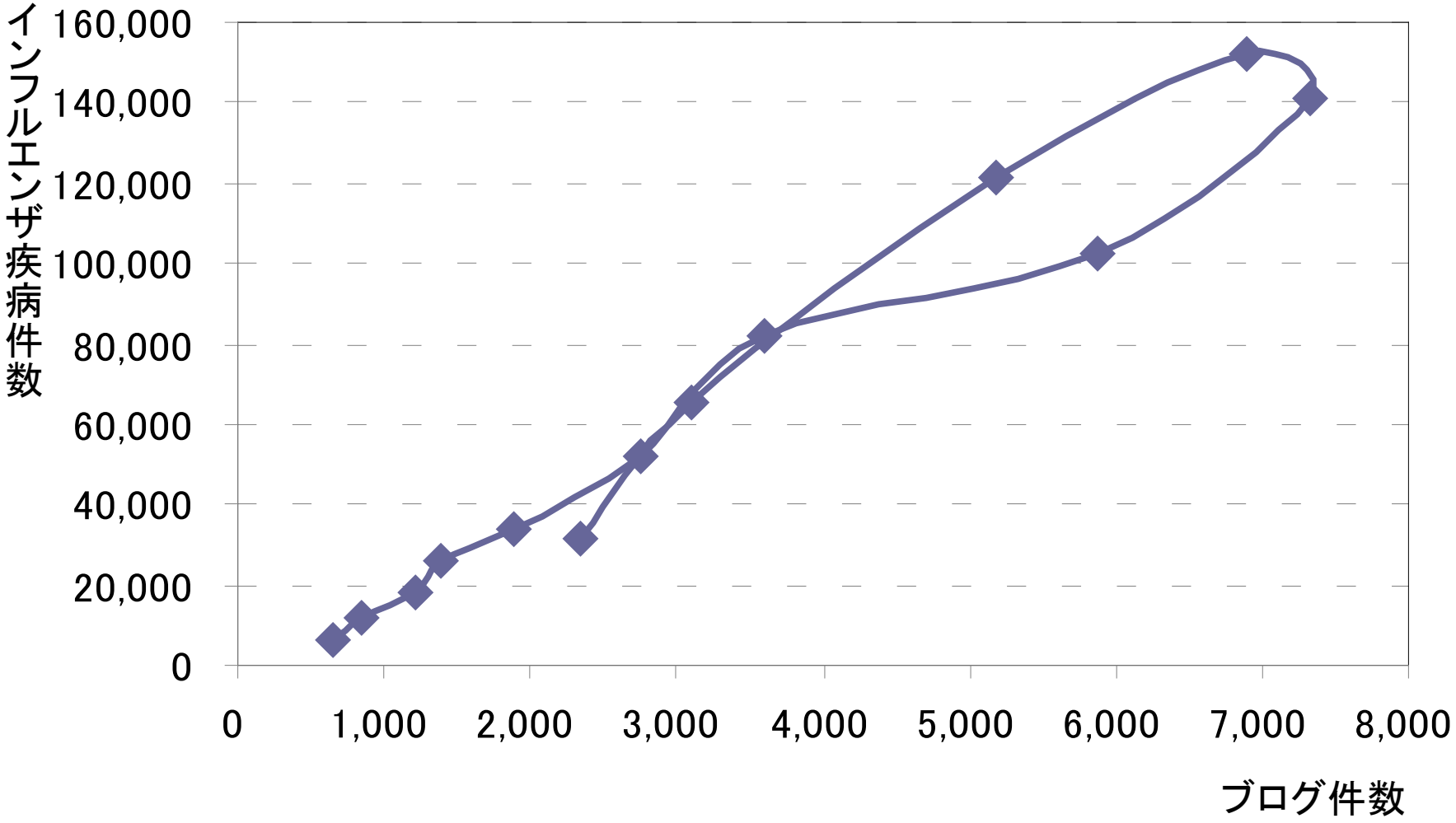


位相がずれている例：インフルエンザの流行*



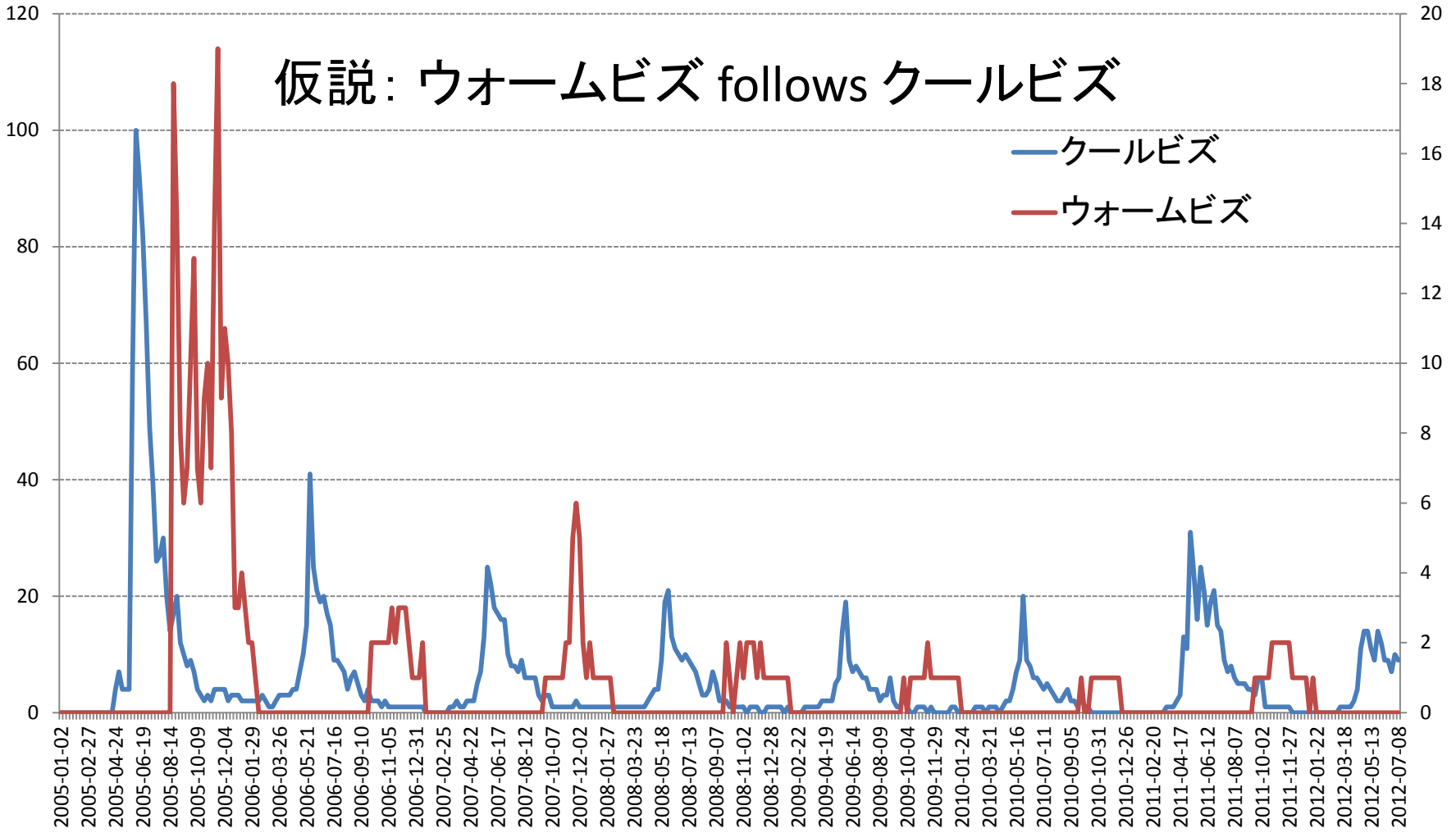
*拙稿「ブログリサーチ」,同文舘,p.140,図5-17

位相がずれている例：インフルエンザの流行*



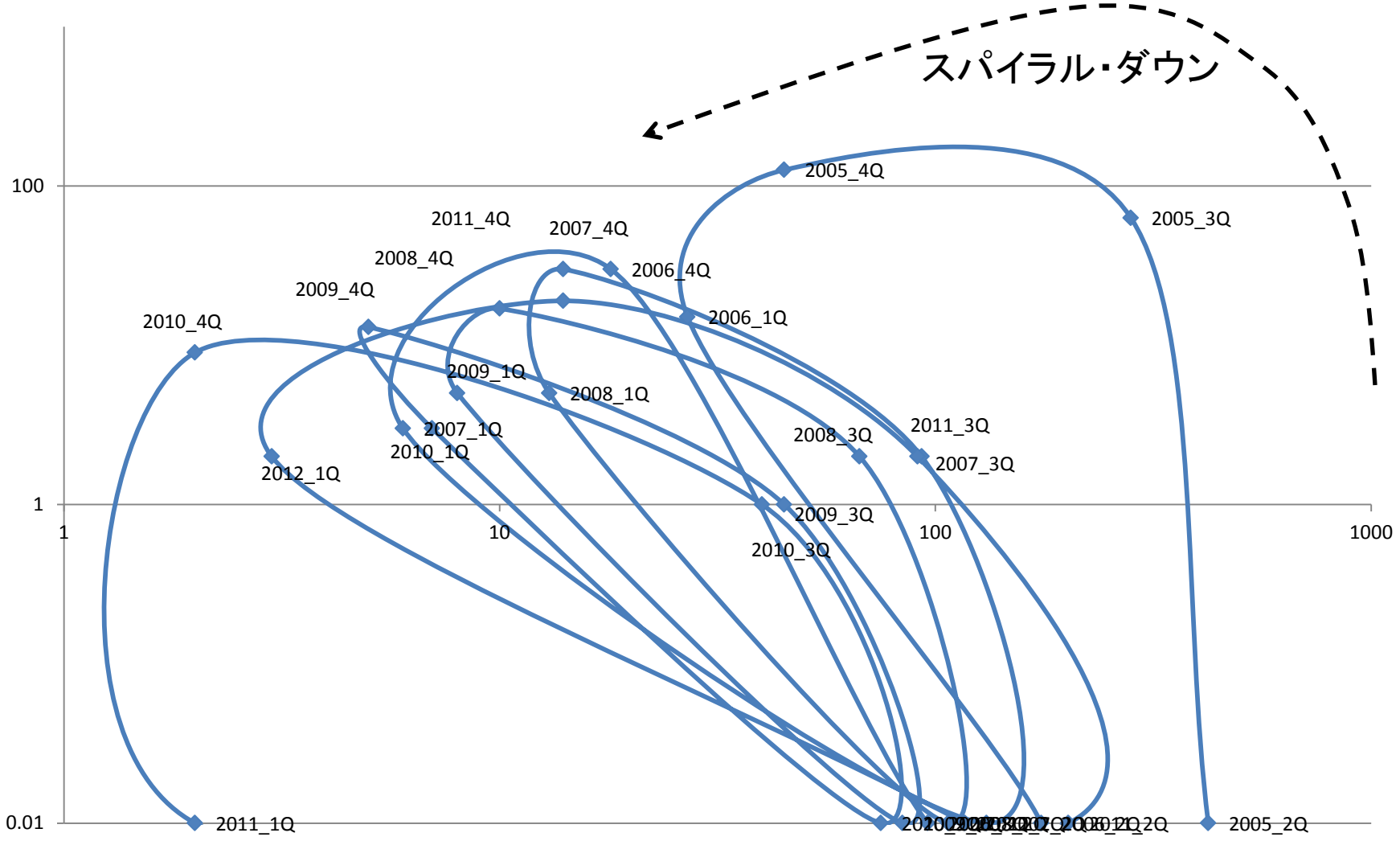
*拙稿「ブログリサーチ」,同文舘,p.141,図5-18

位相がずれている例:
クールビズ とウォームビズ



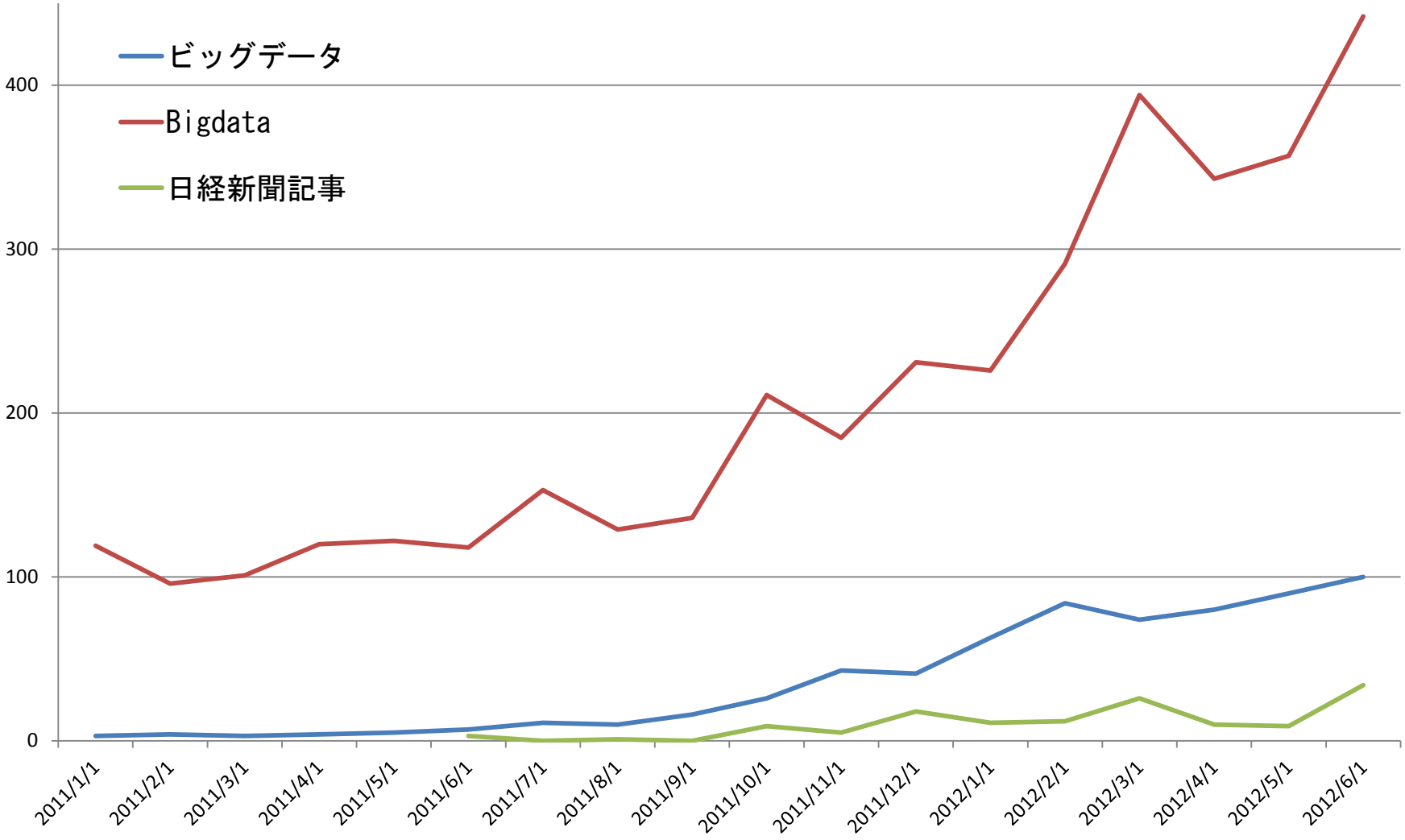
マーケットバスケット分析が時間的に同期化したファクターの共起関係に注目するのに対し、他のファクターからの波及、時間のずれた共起(相関)関係に注目する

位相がずれている例：クールビズとウォームビズ



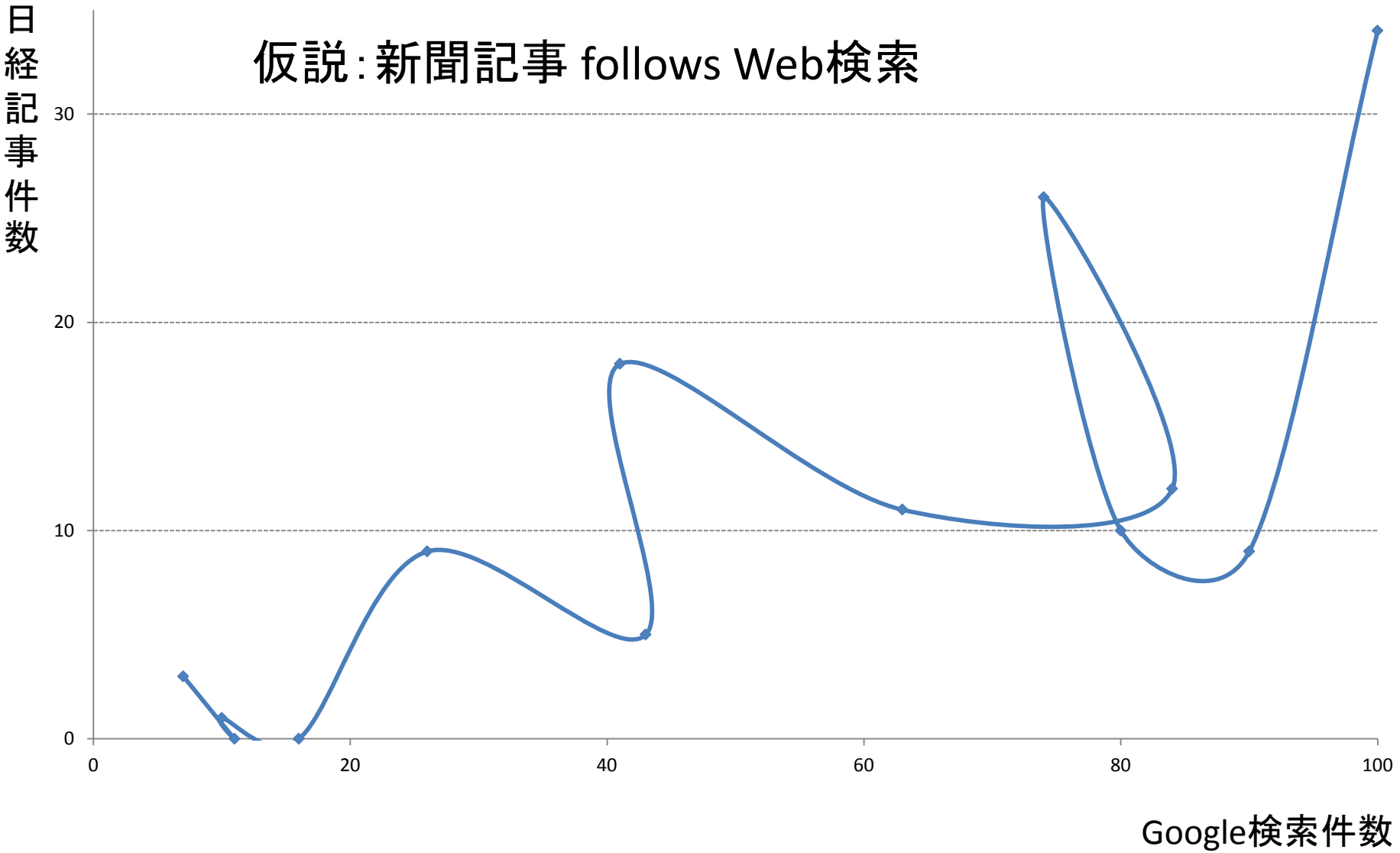
注)対数軸のため、0の値を0.01に変更している

ビッグデータのトレンド
Google Insights と 日経新聞の同期化



ビッグデータのトレンド

Google Insights と 日経新聞の同期化



Ⅱ　ビッグデータをキーワードにした テキストマイニングの事例

1. ビッグデータ(新聞記事)
2. eWOM(学術誌)
3. 普及曲線とスパイラル曲線
4. 過去のIT関連ブーム
5. POSとeWOM

Key Question:　ビッグ・データのトレンドは、どう広がるか？　　＝＞
過去のIT関連ブームは、どのように普及し収束していったか？

これまでのIT関連ブーム

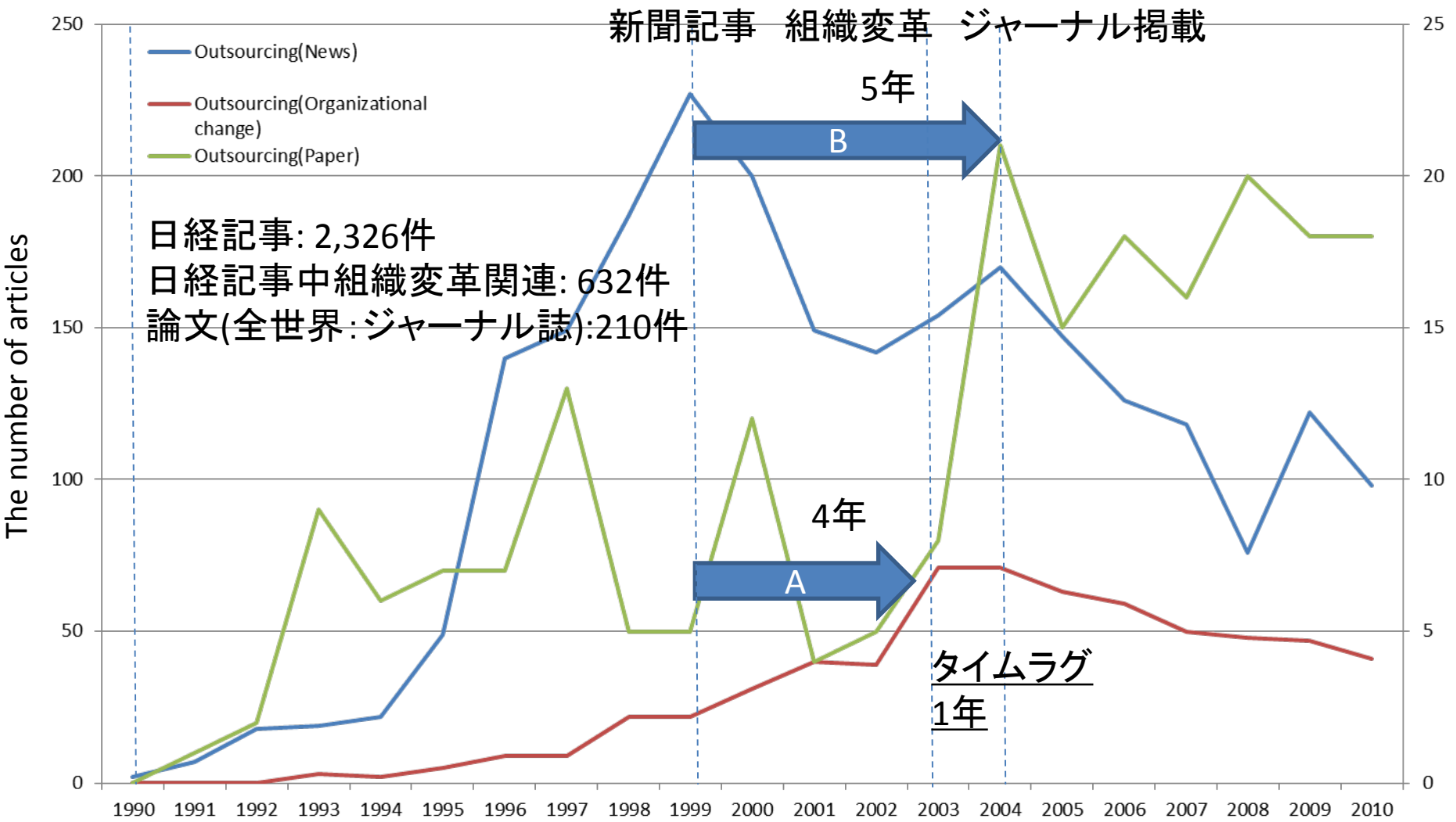
- BPR
  - ハマー＆チャンピー：リエンジニアリング
- ERP
  - 業務統合ソフト
- データウェアハウス
  - リレーショナルデータベース
  - コッド博士：OLAP
- SCM
  - ゴールドラット：ザ・ゴール、制約理論
  - I2などのソフトウェア
- ブームの共通点
  - IT関連企業が、ソリューションやソフトウェアを準備
  - 日本企業は欧米の優れた事例を調査して、自社に適用可能性を判断

タイムラグ分析*

- 異なるデータソースを用いて同期化
  - 新聞記事
  - アカデミック文献
- それぞれを包括的にレビューし、データを抽出、テキストマイニングを実施、関連性を分析する
- ITドリブンのイノベーション普及とFollow仮説(H.Sasaki)
  - Structure follows strategy (Chandler,1962) by 2-4 years.
  - The diffusion of publication is slower than the diffusion of innovation.

*以下、次の報告資料を利用：H.Sasaki(2012),"IS research and its standpoint -Revisiting the "reference discipline" problem from Japan-", Tokyo Keizai University Information Systems Symposium 2012.

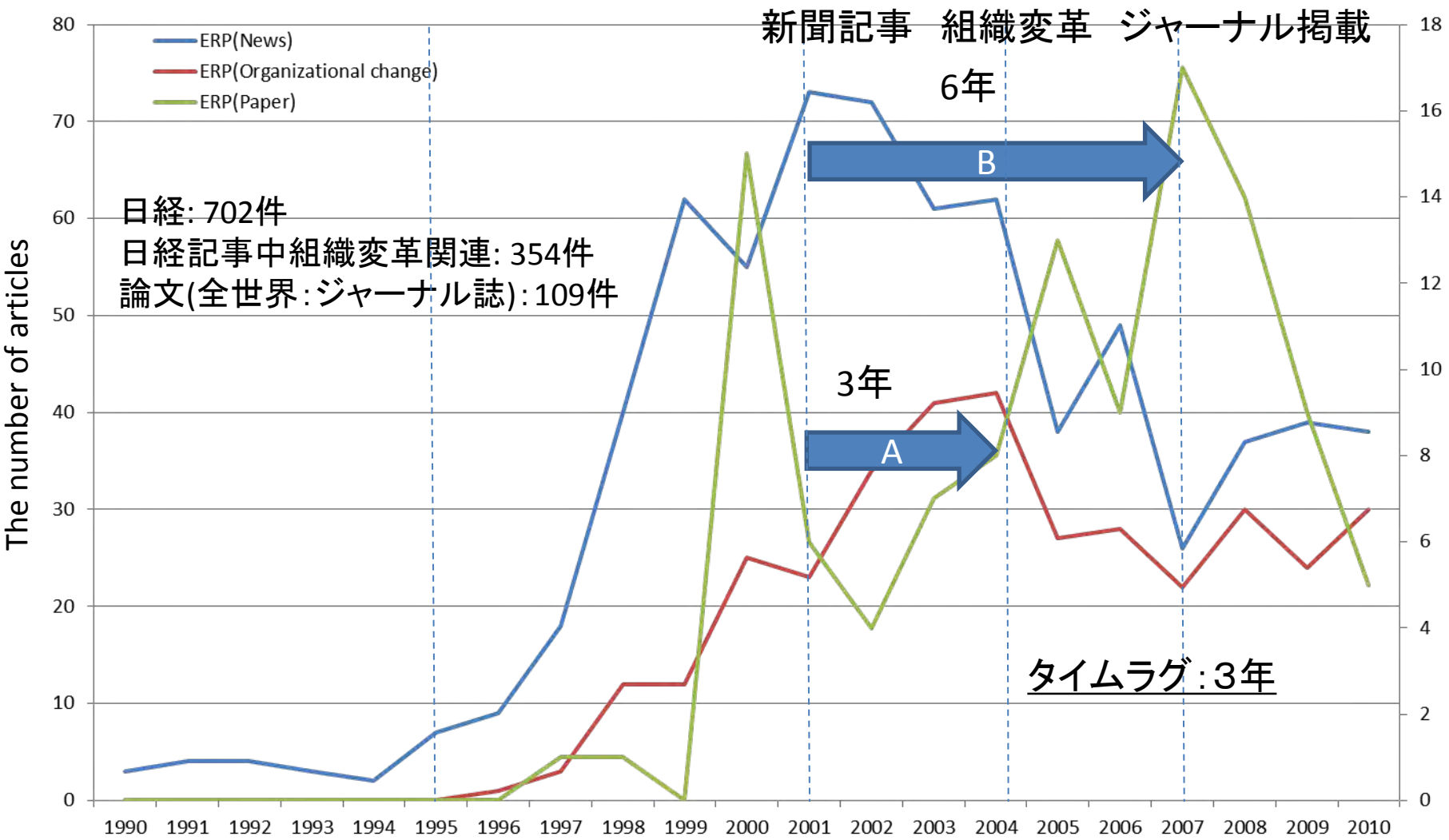
タイムラグ(文献と新聞)*: Outsourcing



(A)イノベーション普及:新聞記事＝＞組織変化

(B)文献発刊　　:新聞記事＝＞ジャーナル掲載

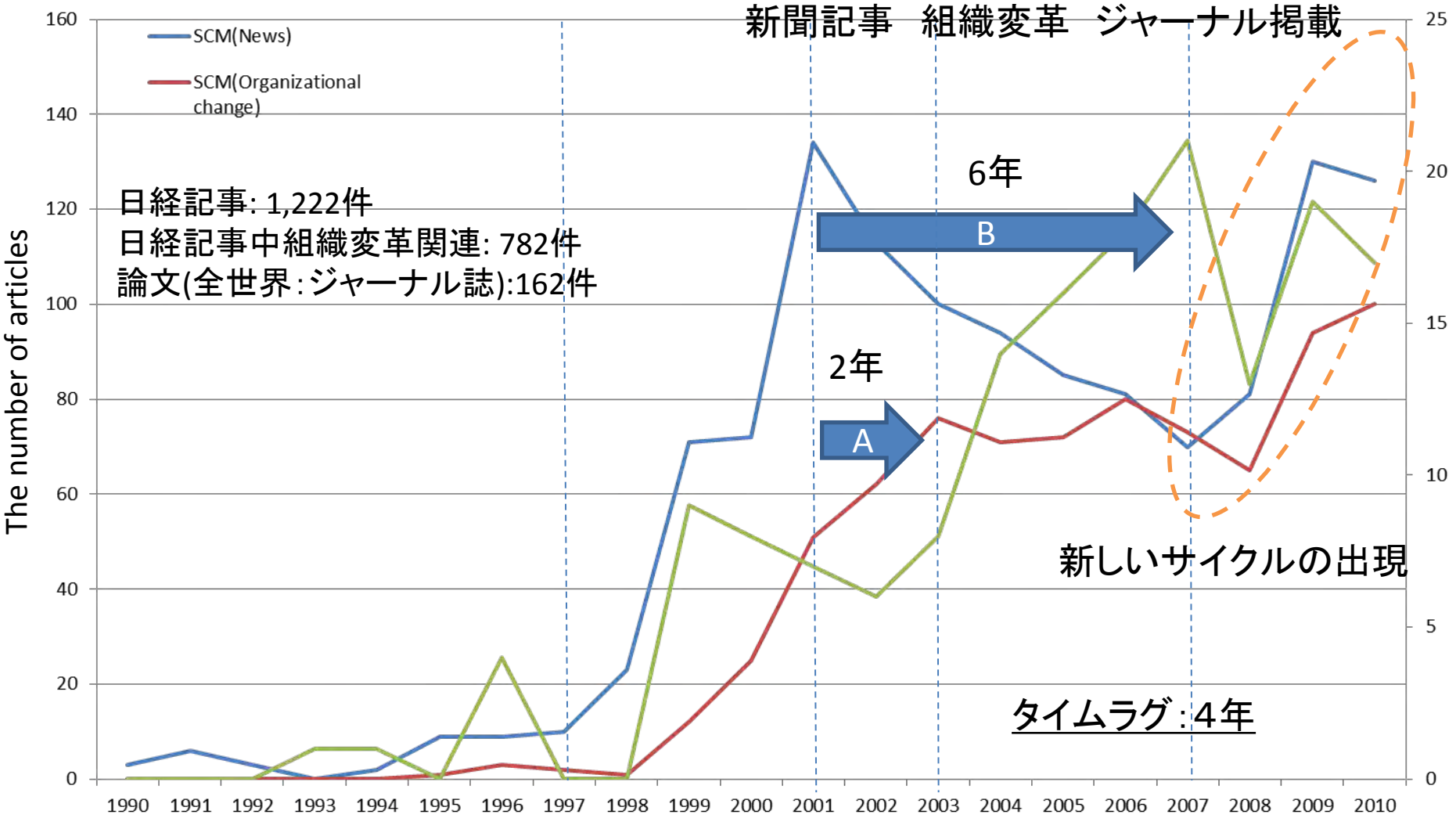
タイムラグ(文献と新聞)*: ERP



(A)イノベーション普及:新聞記事＝＞組織変化

(B)文献発刊　　:新聞記事＝＞ジャーナル掲載

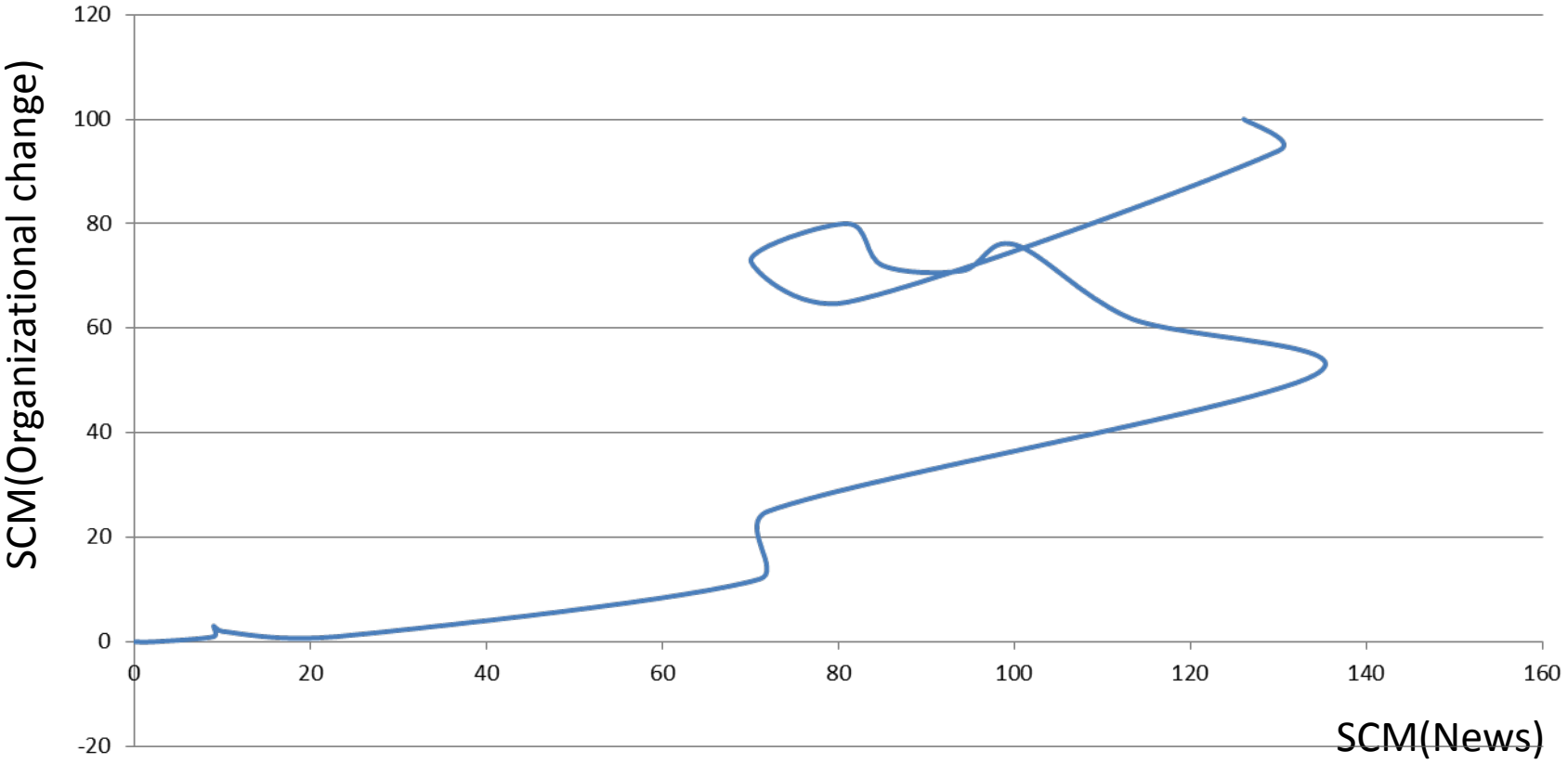
タイムラグ(文献と新聞)*: SCM



(A)イノベーション普及:新聞記事＝＞組織変化

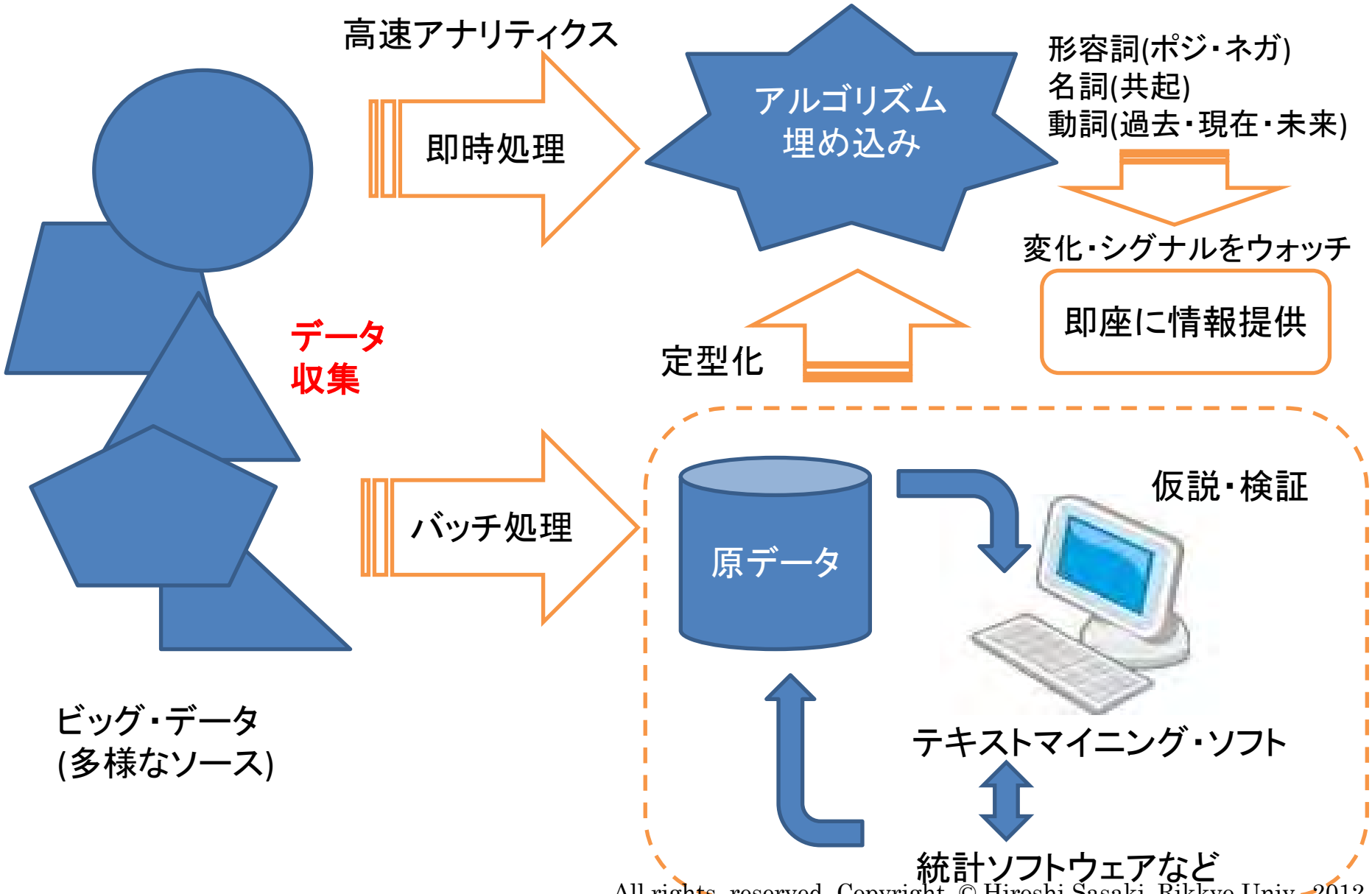
(B)文献発刊　　:新聞記事＝＞ジャーナル掲載

スパイラル曲線: SCM



Ⅲ　テキストマイニングの効率化

テキストマイニングと仮説検証プロセス



おわり

ご清聴、ありがとうございました

sasaki-h(a)rikkyo.ac.jp