大規模データからのデータマイニングと工学への応用
～知識創出学の確立を目指して～

北海道大学　吉岡真治

本発表の構成

- 北海道大学グローバルCOEプロジェクトの紹介
  - 知の創出を支える次世代IT基盤拠点
- テキストマイニングによる工学的活動の支援
  - 創造的設計支援環境UAS
  - プログラムマネジメントのための新聞からの利害関係者の抽出

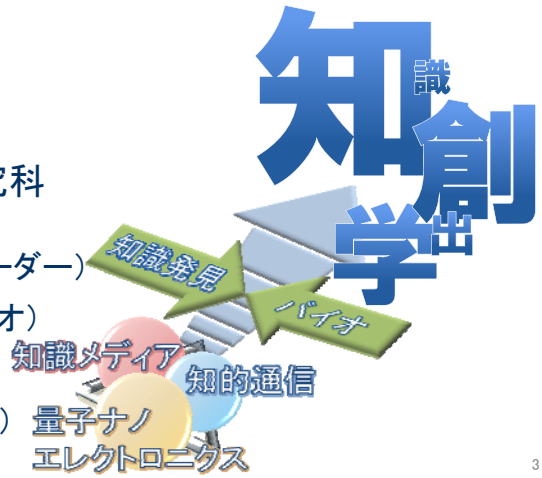
北海道大学
大学院情報科学研究科

平成19年度グローバルCOEプログラム

知の創出を支える次世代IT基盤拠点

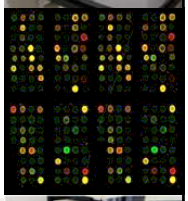
北海道大学
大学院情報科学研究科

有村 博紀（拠点リーダー）
渡邉 日出海（バイオ）
末岡 和久（ナノ）
宮永 喜一（メディア）

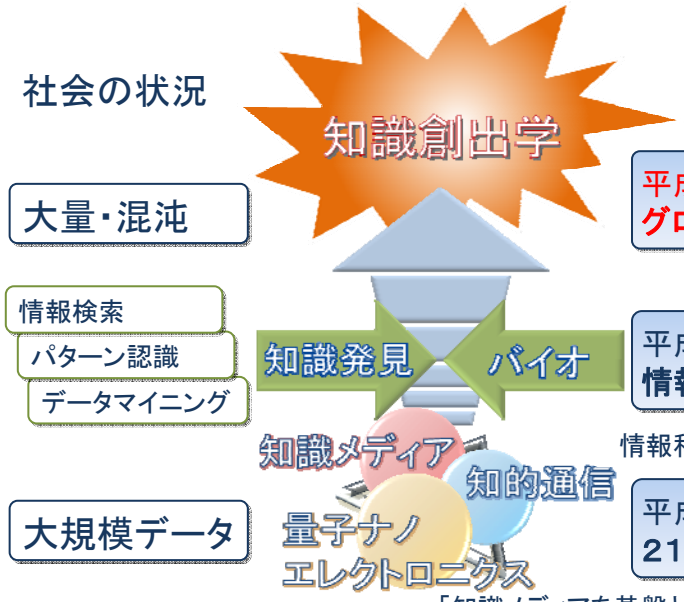


背景 ー 大量データと知の創出

- 実世界と情報世界の大量・混沌データ
  - 新しい観測手段と自動計測技術の飛躍的進歩
  - 情報通信技術の急速な発展
  - サイバー世界：巨大な情報の海
- 科学と技術
  - 実世界と情報世界からの**知の創出**を目的
  - 知識創出が困難になりつつある
  - 膨大なデータからの知の創出を支えるための新しい**情報技術・学問基盤・人材**が必要
- 本グローバルCOEの目標
  - **「知識創出学」**の確立
  - **世界的教育研究拠点形成**



将来計画 ― 21世紀COEからの出発



拠点形成計画　―　計画を支える研究基盤

科研費特別推進研究(平成17－19)
半構造マイニング

科研費特別推進研究(平成18－22)
半導体ナノワイヤ
エレクトロニクス

NatureとScienceに6つの論文
多数の引用(一つは3600回以上)

産学共同シーズイノベーション化事業(育成ステージ)
極限計測バイオセンサー

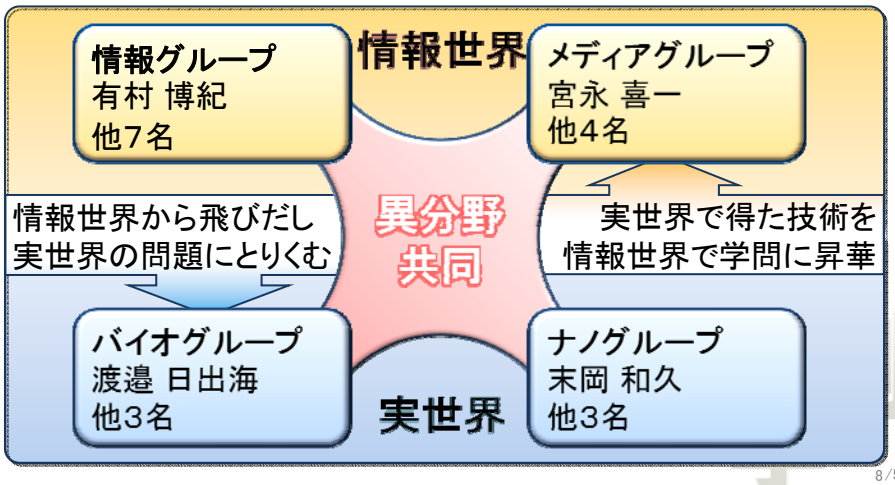
世界最高速の無線通信
600メガbpsスループットの無線LSIチップ

21世紀COEプログラム(H14－18)
知識メディアを基盤とする次世代ITの研究

拠点形成計画 ― 「知識創出学」の創成

実施体制 ― 異分野共同教育研究

- 異分野の結合 → 異分野共同研究
- 教育と研究の連携 → 共同プロジェクト制

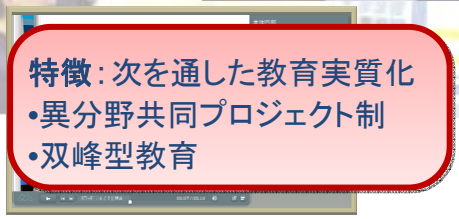


若手人材育成

情報科学研究科

双峰型教育

大学院教育実質化
- 先端的研究に基づく教育
- 学位の副指導教員制
- 外国人教員による教育・研究
- E-learning（独自開発）



特徴：次を通した教育実質化
- 異分野共同プロジェクト制
- 双峰型教育

グローバルCOE

国際化
- 国際会議発表支援
- 海外インターンシップ
- 研究交流プログラム
- 海外大学との交流協定
- 著名な研究者の招聘

共同プロジェクト制

自立支援
- RA雇用
- 競争的若手研究資金

拠点形成実施体制 ― 事業推進担当者一覧

情報
- 有村 博紀 (41)
- T. Zeugmann (51)
- 吉岡 真治 (38)
- 湊 真一 (41)
- 工藤 峰一 (48)
- 原口 誠 (53)
- 佐藤 義治 (61)
- 田中 譲 (57)

バイオ
- 渡邉 日出海 (43)
- 遠藤 俊徳 (39)
- 岡嶋 孝治 (38)
- 山本 克之 (61)

メディア
- 宮永 喜一 (50)
- 長谷山 美紀 (43)
- 金子 俊一 (51)
- 栗原 正仁 (51)
- 小柴 正則 (58)

ナノ
- 末岡 和久 (41)
- 齋藤 晋聖 (33)
- 福井 孝志 (56)
- 末宗 幾夫 (57)

若く活力あるメンバーで
長期的な拠点形成を目指す

テキストマイニングによる工学的活動の支援

- 研究事例紹介
  - 創造的設計支援環境UAS
  - プログラムマネジメントのための新聞からの利害関係者の抽出

創造的設計支援のための仮説的知識生成支援環境 Universal Abduction Studio

- 背景
  - 創造的な設計解
    - 以前の設計で使った知識をそのまま組み合わせて利用した設計では、創造的な設計解が提案できない
    - 以前に使ったことのない知識が必要
      - 科学的な発見に基づく知識の増加
      - 問題の再定義による異なる領域知識の組み合わせ
  - 知識生成の方法
    - 新しい科学的な発見による知識発見のモデル化は困難
    - 異なる領域の知識の統合により、創造的設計解に至る可能性を確認
- 目的
  - 設計者にとって役立つ異なる領域の関連知識の発見と融合により新たな仮説的知識の生成を行う手法の提案

創造的な設計の支援

- TRIZ
  - 発明のパターンを類型化
    - 問題が持つ要因を整理して構造化する手法を提案
    - 問題構造に応じて類型化された解法の提示
- 等価変換理論
  - 既存の解決済の問題(Ao) と問題解決領域 (B τ) の間に本質的に含まれる同じもの(c) を見いだし、、問題解決を行う方法を提案

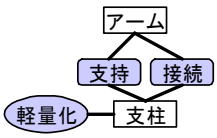
$$A_o \overset{\sum a}{\underset{\sum b}{\Uparrow = \Uparrow}} B_\tau \quad c\varepsilon$$

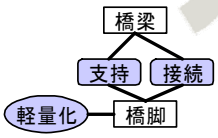
異なる領域知識の利用

- 類似した事例の利用
  - 機械要素の摩擦・ばね・ダンパー
  - 電気回路の抵抗・コンデンサ・コイル
- 異なる領域の知識の統合により、創造的設計解に至る可能性を確認
  - ポータブルCDの音とび防止
    - ピックアップの振動制御：機械ドメインの問題
    - データの継続的読み出し：通信や電子回路ドメインの問題

問題構造の類似性に基づく仮説的知識生成と利用

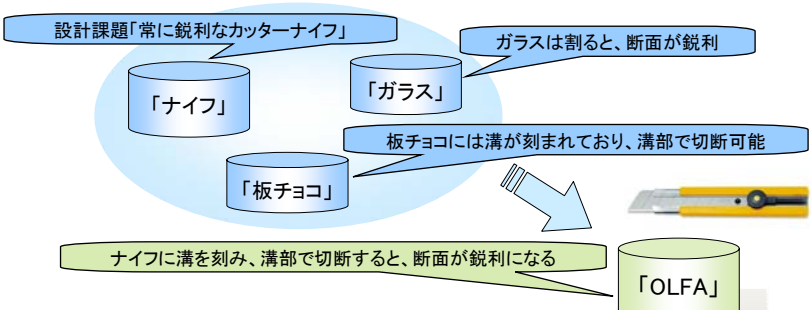
- 知識単独での類似性
  - 数式表現が似ている
    - 例：摩擦・ばね・ダンパーと抵抗・コンデンサ・コイル
- ○問題構造の類似性
  - 使われる状況が似ている知識
    - 問題定義において同じような役割を果たす概念間に類推に基づく対応関係を想定





創造的設計の具体例

- 「折る刃式カッターナイフ」[オルファ株式会社]

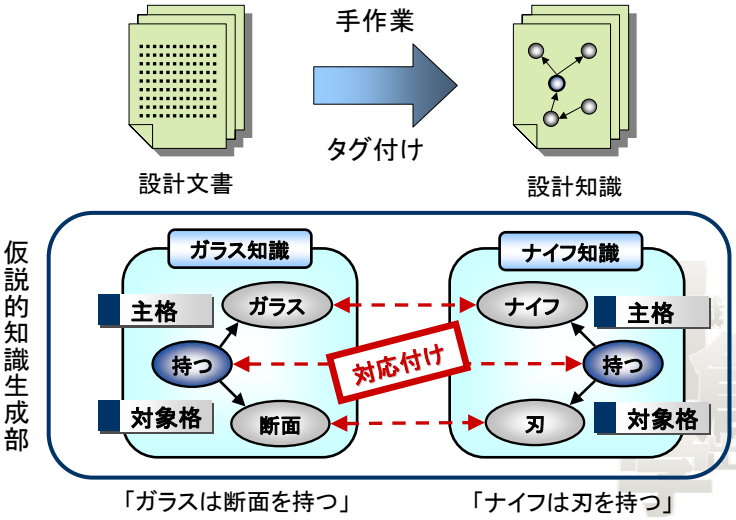


- 本研究で対象とする創造的設計のための知識生成

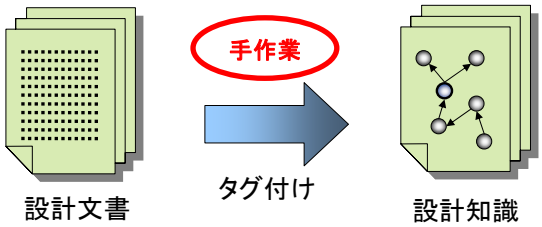
異なる領域で行われている設計事例あるいは当初無関係と思われていた領域知識から新たな知識を生成する

Universal Abduction Studio (UAS)

- 創造的設計支援のための仮説的知識生成システム



UASの問題点



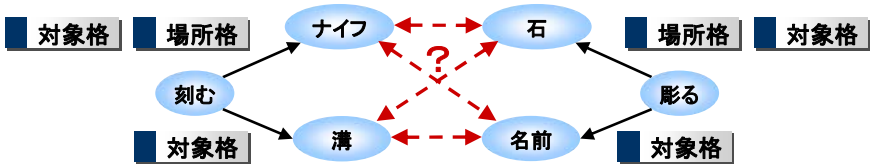
- 手作業によるタグ付けのため、知識の量が不足
- 大量の文書に一貫性のあるタグ付けは困難

自動化

知識生成に用いる自動文書タグ付け手法と生成された知識の利用法の提案

タグ付けによる文書の構造化

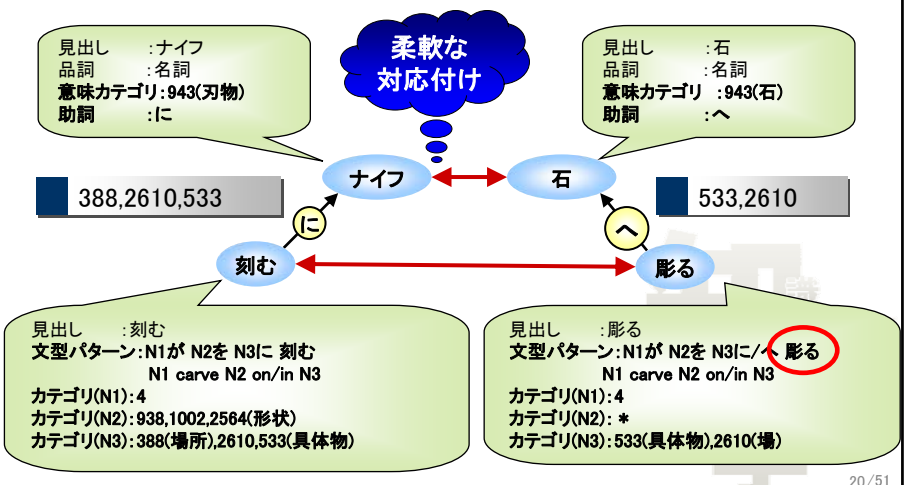
- 「ナイフに溝を刻む」&「石に名前を彫る」



- 述語と名詞の意味的関係をタグ付けする方法
  - **格文法**：全ての言語に共通する比較的少数の「深層格」の存在を仮定し、それを付与する
- 問題点
  - 深層格の体系は曖昧であり格を一意に決めることは困難
  - 計算機上に表現できる情報（係り受け情報など）だけではさらに困難

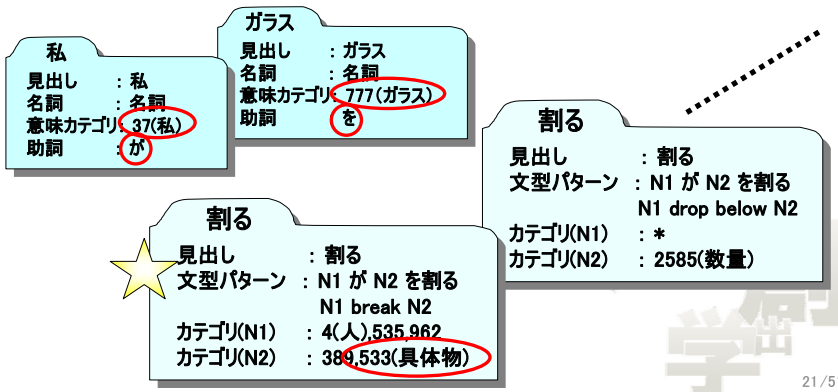
述語と名詞の意味的関係

- 結合価文法：「深層格」の存在を仮定せず、「表層」で述語と名詞の結合関係が定義できる

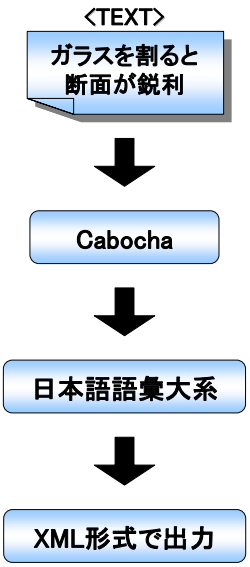


文型パターンの選定

- 動詞が多義の場合、複数の文型パターンを持つ
  - 動詞に対応する複数の文型パターンから、動詞に係る名詞と助詞の組で選定
  - 例:「私がガラスを割る」

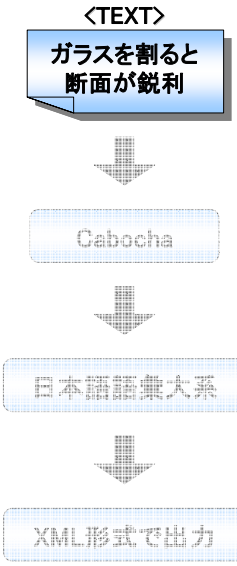


自動文書タグ付けシステム(全体像)



- 入力部
  - テキスト形式の日本語文書を入力
  - if-then型文書
- 構文解析部
  - 形態素解析・係り受け解析を行い、文書の構造を抽出
- 辞書情報取得部
  - 語彙に辞書情報を付加
  - 文型パターンの選定
- 出力部
  - XML形式で出力

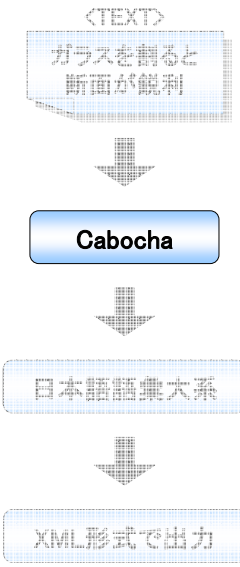
自動文書タグ付けシステム（入力部）



- if–then型知識
- パターンによる分類の自動化
  - 「～の場合、・・・。」
  - 「～すると、・・・。」

のようなパターン10種類を作成し、分類の自動化

自動文書タグ付けシステム（構文解析部）




```
＊0 -1D
ガラス      ガラス      ガラス      名詞–一般
を          ヲ          を          助詞–格助詞–一般
＊1 -1D
割る        ワル        割る        動詞–自立 五段・ラ行 基本形
と          ト          と          助詞–接続助詞
・・・・・・・・・・・・
```

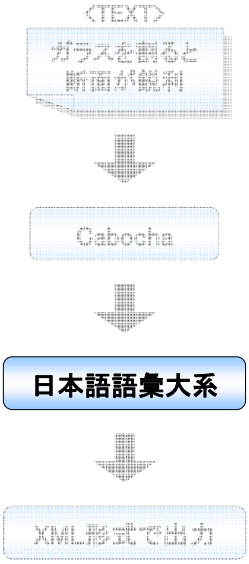
手作業で修正（文節・品詞）

```
ガラスを–D
  割ると––D
    断面が–D
      鋭利
```

修飾関係　語彙情報

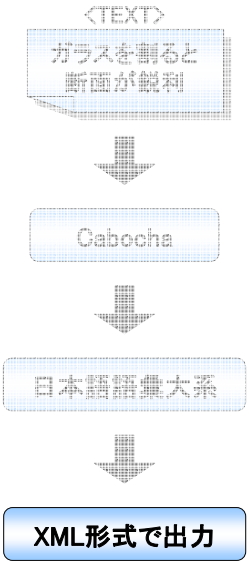
```
＊0 1D 0/1 1.18556152
ガラス      ガラス      ガラス      名詞–一般
を          ヲ          を          助詞–格助詞–一般
＊1 3D 0/1 0.61950975
割る        ワル        割る        動詞–自立 五段・ラ行 基本形
と          ト          と          助詞–接続助詞
・・・・・・・・・・・・
```

自動文書タグ付けシステム(辞書情報取得部)



- 名詞に辞書情報を付加
  - 辞書未登録語には全ての意味カテゴリと対応づく意味カテゴリを付加
  - 複合語未登録の場合は主辞(多くの場合は複合語の末尾の名詞)の意味カテゴリを用いる
- 述語の文型パターンを絞込む
  - 省略された情報がある場合には補完
    - 主体となる名詞(主語)
    - 同一文中での対象の省略

自動文書タグ付けシステム(出力部)




```
<?xml version="1.0" encording="UTF-8"?>
<!DOCTYPE template>
<taggedDocument>
<document>ガラスを割ると、断面が鋭利</document>
<rule>
<CWord pos="名詞-一般" base="ガラス">ガラス</Cword>
<CWord pos="助詞-格助詞-一般" base="を">を</Cword>
<CWord pos="動詞-自立" base="割る">割る</Cword>
      ・・・・・・・・・・・・・・・・・・
<condition>
<word wordID="ガラス-1" pos="名詞-一般" goiType="777">ガラス</word>
<word wordID="割る-1" pos="動詞-自立"
          sentenceType="N1が N2を 割る  N1 break N2">割る</word>
<modifier type="を格" nGoiType="389,533,2582">
<fromWord>ガラス</fromWord>
<toWord>割る</toWord>
</modifier>
</condition>
<consequence>
      ・・・・・・・・・・・・・・・・・・
</consequence>
</rule>
</taggedDocument>
```

実験概要

- 実験：自動文書タグ付けシステムの評価
  - 生成された知識が必要な情報を保持できているかの検証
- 実験手順
  - 「続・機械設計心得ノート」から、if-then知識と考えられる270文(文節数3594)・346個を入力データとする
  - 作成したシステムに入力データを1文ずつ入力し、出力結果を以下の観点で検証
    - 手作業をどのくらい必要としたか
    - 辞書情報をどのくらい付加することができたか

実験結果

- 手作業による修正

手作業による文節区切り修正数	265(7.4%)

- 構文解析の精度

修飾関係の誤解析	121(3.4%)

- 辞書情報の付加

	のべ語彙数	カテゴリ付与（再現率）	適切なカテゴリ（精度）
名詞	1163	1053（90.5%）	952（90.4%）
述語	849	411(48.4%)	345(83.9%)

実験結果

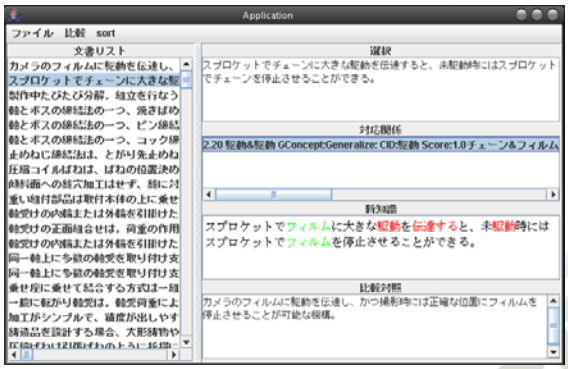
- 不適切なカテゴリを付加した原因
  - 専門用語（自重、内輪、仕上げしろ、～座）
- 複合語の処理で不適切なカテゴリを付加した原因
  - 形式名詞（～よう）の取り扱い
  - 接尾辞（～部）

考察

- 名詞は再現率・精度ともに良いと考えられる
- 述語の結果の悪さは係り受け解析の精度・辞書の詳細度にも依存
- 設計に関する文書には、複合語が多く、主辞を用いたタグ付けにある程度効果が得られた
- 助詞は異なるが意味カテゴリのみが対応とれた場合で絞った文型パターンに適切なものが多かったことで、助詞の表記のゆれにある程度対応できる

UASのシステムを用いた検証

- 実験で生成した知識を用いて、UASで導出されている仮説的知識が導出できるかを検証



- 実際にUASで導出されている知識の導出に成功

まとめと今後の課題

- まとめ
  - 創造的設計に有効と考えられる、文書に自動で結合価文法を利用したタグ付けを行う手法を提案した
  - 生成された知識を用いて，UASの仮説的知識導出手法を改良し，創造的な仮説的知識を導出する可能性を示すことができた
- 今後の課題
  - 仮説的知識検証を設計の専門家に依頼する
  - 既存研究と、本研究の比較をしながら、タグ付けの詳細度と知識の量のトレードオフの問題について、そのバランスを検討する

背景

- **プログラムマネジメント**
  - いくつものプロジェクトが絡み合った大規模・複雑な事業(プログラム)を統合的に協調させて管理し、事業実施の効率を高める手法。
  - プログラムの全体を把握する必要がある。

起こりそうな問題は？　　どんな企業と関わるのか？

どんなプロジェクトが必要？

プログラム
・マンション建設
・駅前再開発

過去に実施された類似プログラムにおけるステークホルダー(利害関係者)を新聞記事などでチェックすることで、見落としを防ぐことが大事

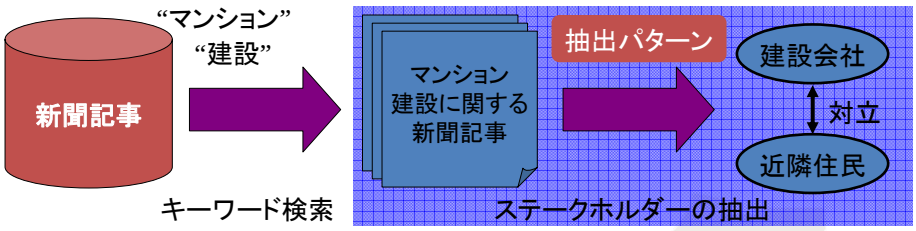
しかし・・・　人手によるチェック　＝　膨大なコスト

背景 – プログラムマネジメント支援

- **先行研究(吉田 '05)**

ステークホルダー自動抽出システム

マンション建設の場合・・・



抽出パターン　～が・・・として～を訴えた訴訟で　→ マッチング (係り受けも考慮)

新聞記事　近隣住民が・・・として建設会社を訴えた訴訟で・・・

問題点　▪ あらかじめ手作業でパターンを用意　→　抽出漏れ