適合的汎化に基づく情報検索システムの実験的研究（第2報）
－検索語の網羅性に注目した検索インターフェースの作成－

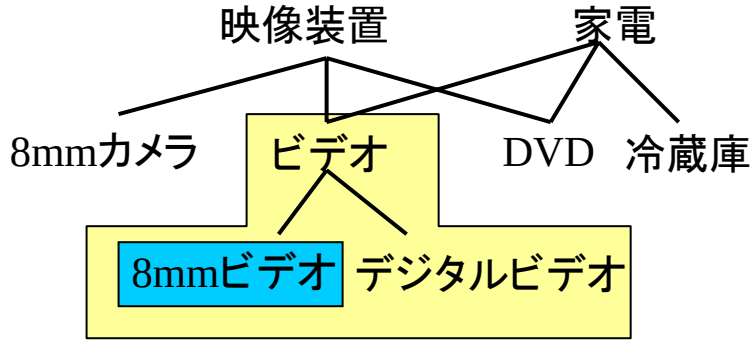
北海道大学　吉岡真治
原口誠

背景と目的

- 背景
  - IT技術の発展に伴い、大量の文書情報が利用可能
    ⇒適切な情報検索のためには、検索者の多様な検索意図を適切に表現する事が必要
  - 検索者の検索意図を表現する検索式
    - 検索語は検索意図を十分に反映しているのか？
    - 検索意図の推定
      - レレバンス・フィードバック、ユーザモデルの利用
    - 様々な詳細度の語が混在し、検索式の理解が困難
  - 電子化辞書・シソーラスの利用
    - 検索意図と異なる検索式の拡張を行う可能性
- 目的
  - 検索意図を反映した検索語の概念階層の汎化を行う情報検索システム
    - 検索語が持つ概念の抽象度を、検索者の検索意図に応じて適合的に設定することにより、高効率な情報検索を実現

適合的汎化に基づく情報検索システム

- 適合的汎化
  - データマイニングのための概念の汎化（工藤、原口, 2000）
    - 抽象化による情報量の損失を最小にする属性の汎化
    - 精度を維持したままで意味のある少数のルールを作成
  - 辞書が持つ概念階層構造を利用した検索語の汎化
    - 検索精度を維持するために、検索意図に応じた汎化
    - 検索者にとって理解しやすい検索式の拡張
    - 検索語の汎化による影響
      - 再現率の向上が期待される。
      - 過度に汎化を行うと、検索意図が曖昧になり、関係のない文書を検索してしまう。

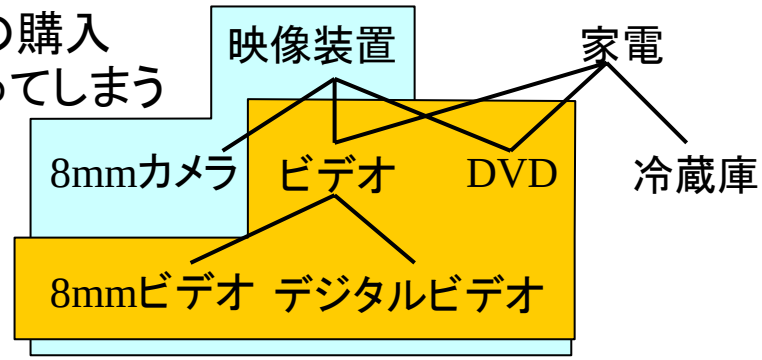


概念階層の一部

電子化辞書・シソーラスの概念階層について

- 一般的な電子化辞書・シソーラスの階層構造の特徴
  - シソーラスの概念階層が一般的な分類を目標としているため、中間階層の抽象概念が粗い。
  - 一段の抽象化が過剰に抽象度を上げる。
- 目的指向の概念階層の修正
  - 特定の目的を正確に表すためには、目的に応じた中間階層の概念を作る必要がある。

検索意図：レンタルビデオのための機械の購入
DVDまで拡張したいが、8mmカメラが入ってしまう

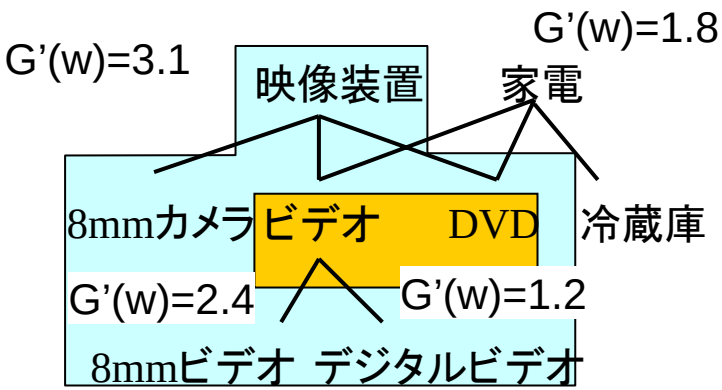


適合的汎化のための指標

- 今回のシステムで利用する検索語の特徴量
  - 検索語の存在が文献の適合性を判断するのにどれだけ役に立つかを示す指標を考える
    - 相互情報量に基づくキーワード選択

$$G'(w) = p(w|r)\log_2 \frac{p(w|r)}{p(w)}$$

    - 指標に基づく概念の汎化と適合文書に基づく、中間階層概念の作成



適合的汎化と概念カテゴリーの生成

適合文書

レンタルビデオ
映画「.....」
....

DVDタイトルリスト
映画「.....」
....

検索式の作成

- 適合的汎化は検索語と補完関係にある語を探す手法
- 適合フィードバックによる検索拡張
  - 検索語と共起性の強い語など、補完関係にはない語なども利用
- 両者を融合した形で検索拡張

1. 形態素解析による初期検索式の作成
2. 関連文書5件を利用し、検索語と関連文書中に含まれる語の汎化
3. 汎化に成功した語は、汎化概念に対応する語を拡張した検索語として追加
4. 検索語とは関係なく、G’(w)の値が大きいもの上位10件を検索語に追加

プロトタイプシステムの作成

- 通信総合研究所で開発された情報検索パッケージ（内山, 2001）（以下ベースパッケージと呼ぶ）をもとに作成
  - Okapi BM25（Robertson and Walker, 2000）をベースとする
    - 文書のスコアリングの関数

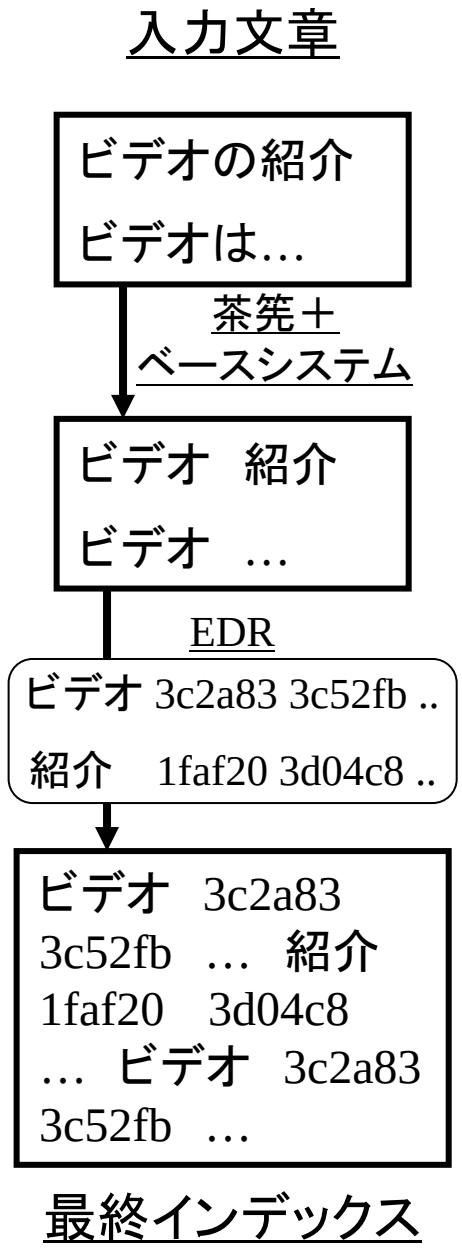
$$\sum_{T \in Q} w^{(1)} \frac{(k_1+1)tf}{K+tf} \frac{(k_3+1)qtf}{k_3+qtf} \qquad K = k_1((1-b)+b \times \frac{document\ length}{average\ document\ length}$$

tf : 文書中のTの頻度　　$w^{(1)}$: 文献中の語Tの重み（Roberson Sparck Jones）
qtf : 検索式中のTの頻度　k_1, k_3, b: パラメータ（デフォルトk_1=1,k_3 =1000,b =1）

  - ベースラインシステムとしての性能を有する
    - NTCIR（NII-NACSIS Test Collection for IR）のデータに対して、上位グループと同等の性能を持つ
- 電子化辞書データとしてEDRを利用
- 形態素解析ツールとして茶筅を利用

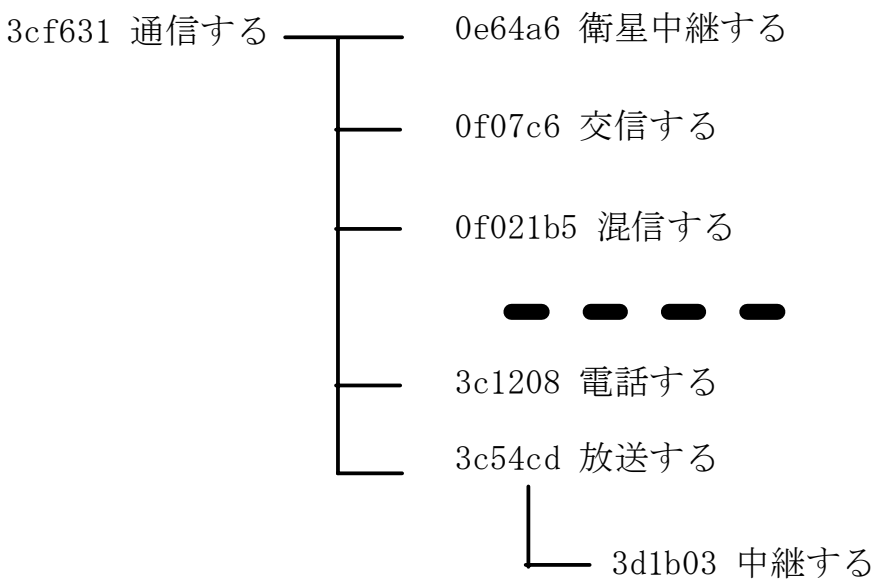
情報検索用インデックスの作成

1. 文書データに対し、茶筅で形態素解析を行い、ベースパッケージと同様に初期インデックス候補となる語（主に名詞）を抜き出す。
2. インデックス候補となった各語に対して、EDRの辞書と照合し、各々の語が所属する概念ID（抽象度の高い概念も含む）のリストを作成する。
   - インデックスサイズを抑えるために、ある一定以上の割合で、文献に現れる概念IDについては、一般的な概念であるとして、汎化の対象から除外した。
3. 形態素解析の結果の語のリストと、各々の語に対応するEDRの概念IDのリストの全てを文献に対するインデックスとして設定する。



作成された中間階層の例

- 検索課題：次世代インターネット
  - キーワード：「通信する」
    - 文書中に含まれる語
      - 通信する、放送する、中継する



検索モデルの特徴

- 確率型モデル・ベクトルモデルの特徴
  - ユーザは検索に必要な言葉をうまく選べない
    → ユーザが選んだ言葉が入っていなくても関連文書になる可能性がある。
  - 関連する多くの語（直接関係なくても、共起性の高い語など）により検索意図を表現
    → 検索式の視認性が低い
- ブーリアンモデルの特徴
  - ユーザは検索に必要な言葉をきちんと選べる
    → ユーザが選んだブーリアン式を満たさないものは関連文書ではない
  - 検索に必要な語により検索意図を表現
    → 検索式の視認性が高い

人間に理解しやすい検索拡張

- ブーリアン式による検索意図の表現
  - 仮説：欲しい文書には、必須の語（概念）が含まれているはず
  - 検索語が含まれなくても関連文書とみなす文書がある。
    → 検索語選択の問題
    → 検索語を補完する語が文書中に含まれるはず
- 適合的汎化の特徴
  - 検索結果をクラスタリングするのではなく、関連文書クラスタのみに注目
    - 検索語と補完関係にある語
    - 網羅的に関連文書に含まれる語（概念）

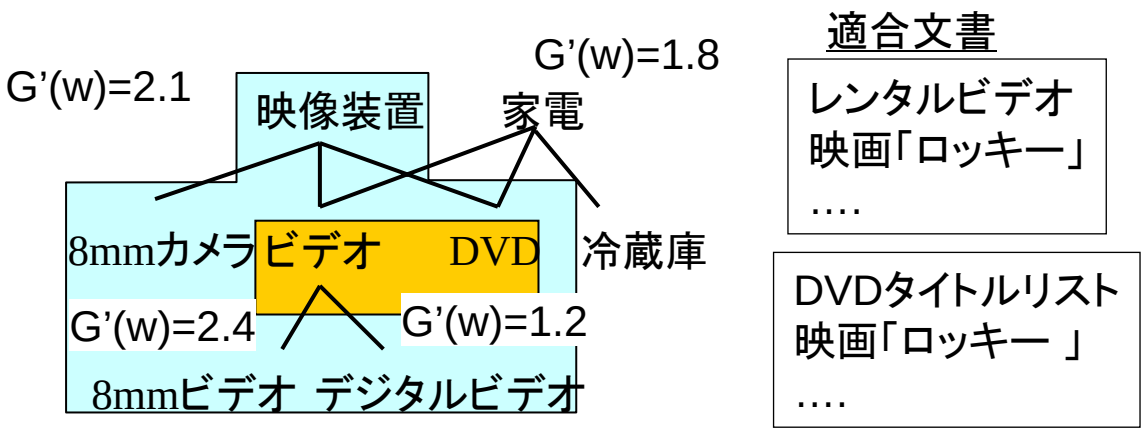
適合的汎化に基づくインタラクティブ検索システム

- 確率型モデルとブーリアンモデルの融合
  - 初期検索式：検索語に関する考察が不十分
    → 確率型モデルによる検索
  - ブーリアンモデルへの切り替え
    - 適合文書が上位に見つからない場合
    - ある程度クリアな検索式が記述できた場合
- 適合文書から得られるキーワードに関する特徴的な情報を提示

ブーリアン式の構築

- 適合文書の網羅性に注目してブーリアン式を作成
  - 適合文書の網羅性に注目：G’(w)の指標が下がっても（一定値で足きり）、網羅性があがる場合に汎化し、概念カテゴリーを作成（ブーリアンのor式として扱う）

初期検索：ロッキー, ビデオ

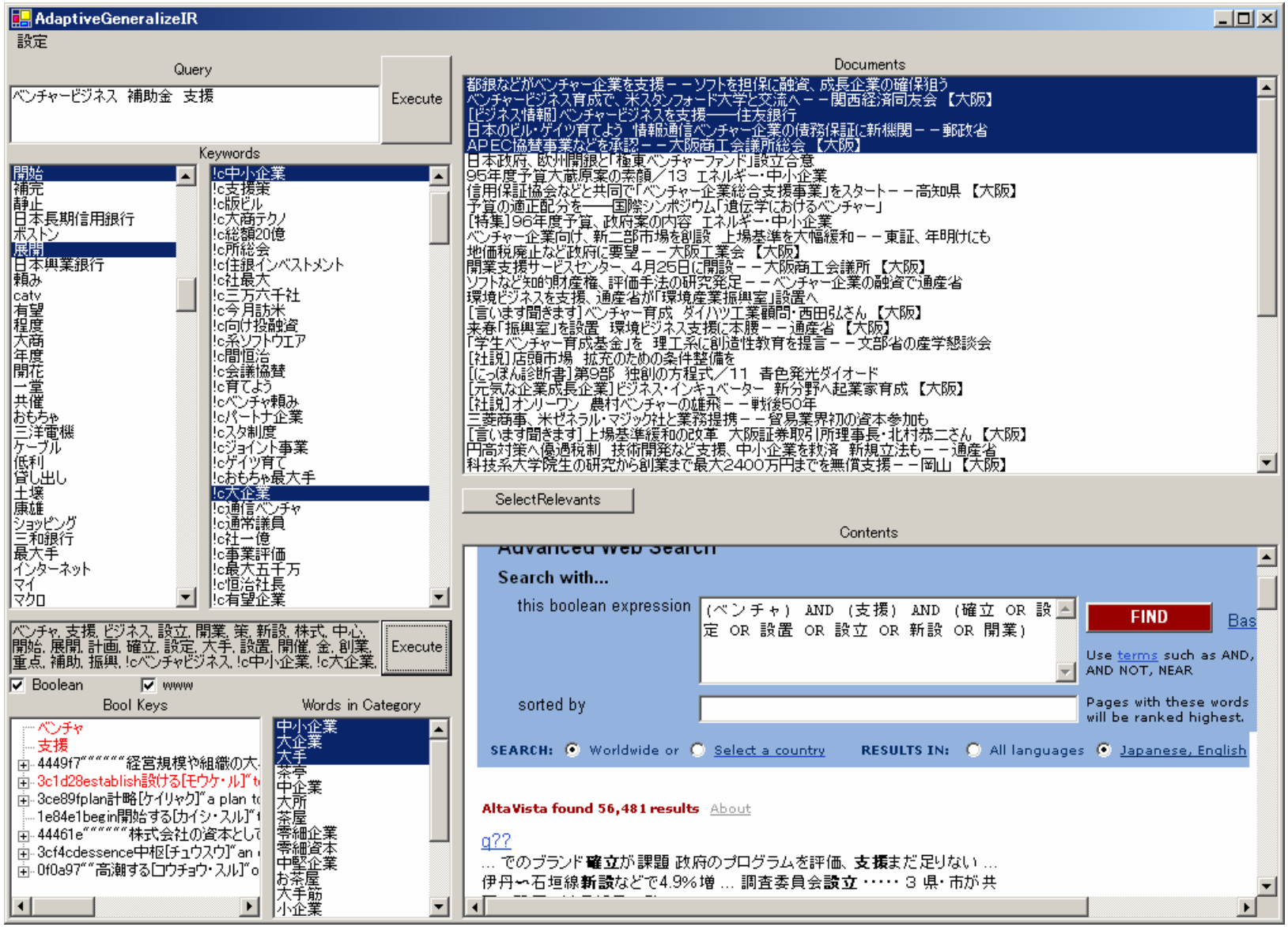


→ Boolean式： ロッキー and (ビデオ or DVD)

検索語の網羅性に注目した検索インターフェース

- Web検索への展開
  - ブーリアン検索式の利点
    - 他のデータベースに対する問い合わせにも利用可能
  - 手元にあるデータベースによる検索結果に基づき、Web検索用のQueryを作成し、検索
    - イントラネット内のデータベースで検索式を明確化して、インターネット上の情報検索へ応用

システムの外観


AdaptiveGeneralizeIR
設定
Query
ベンチャービジネス 補助金 支援
Execute
Keywords
開始
補完
静止
日本長期信用銀行
ボストン
展開
日本興業銀行
頼み
catv
有望
程度
大商
年度
開花
一堂
共催
おもちゃ
三洋電機
ケーブル
低利
貸し出し
土壌
康雄
ショッピング
三和銀行
最大手
インターネット
マイ
マクロ
!c中小企業
!c支援策
!c版ビル
!c大商テクノ
!c総額20億
!c所総会
!c住銀インベストメント
!c社最大
!c三万六千社
!c今月訪米
!c向け投融資
!c系ソフトウエア
!c間恒治
!c会議協賛
!c育てよう
!cベンチャ頼み
!cパートナ企業
!cスタ制度
!cジョイント事業
!cゲイツ育て
!cおもちゃ最大手
!c大企業
!c通信ベンチャ
!c通常議員
!c社一億
!c事業評価
!c最大五千万
!c恒治社長
!c有望企業
ベンチャ, 支援, ビジネス, 設立, 開業, 策, 新設, 株式, 中心, 開始, 展開, 計画, 確立, 設定, 大手, 設置, 開催, 金, 創業, 重点, 補助, 振興, !cベンチャビジネス, !c中小企業, !c大企業,
Execute
Boolean
www
Bool Keys
ベンチャ
支援
Words in Category
中小企業
大企業
大手
Documents
都銀などがベンチャー企業を支援－－ソフトを担保に融資、成長企業の確保狙う
ベンチャービジネス育成で、米スタンフォード大学と交流へ－－関西経済同友会 【大阪】
[ビジネス情報]ベンチャービジネスを支援――住友銀行
日本のビル・ゲイツ育てよう 情報通信ベンチャー企業の債務保証に新機関－－郵政省
APEC協賛事業などを承認－－大阪商工会議所総会 【大阪】
日本政府、欧州開銀と「極東ベンチャーファンド」設立合意
95年度予算大蔵原案の素顔／13 エネルギー・中小企業
信用保証協会などと共同で「ベンチャー企業総合支援事業」をスタート－－高知県 【大阪】
予算の適正配分を――国際シンポジウム「遺伝学におけるベンチャー」
[特集]96年度予算、政府案の内容 エネルギー・中小企業
ベンチャー企業向け、新二部市場を創設 上場基準を大幅緩和－－東証、年明けにも
地価税廃止など政府に要望－－大阪工業会 【大阪】
開業支援サービスセンター、4月25日に開設－－大阪商工会議所 【大阪】
ソフトなど知的財産権、評価手法の研究発足－－ベンチャー企業の融資で通産省
環境ビジネスを支援、通産省が「環境産業振興室」設置へ
[言います聞きます]ベンチャー育成 ダイハツ工業顧問・西田弘さん 【大阪】
来春「振興室」を設置 環境ビジネス支援に本腰－－通産省 【大阪】
「学生ベンチャー育成基金」を 理工系に創造性教育を提言－－文部省の産学懇談会
[社説]店頭市場 拡充のための条件整備を
[にっぽん診断書]第9部 独創の方程式／11 青色発光ダイオード
[元気な企業成長企業]ビジネス・インキュベーター 新分野へ起業家育成 【大阪】
[社説]オンリーワン 農村ベンチャーの雄飛－－戦後50年
三菱商事、米ゼネラル・マジック社と業務提携－－貿易業界初の資本参加も
[言います聞きます]上場基準緩和の改革 大阪証券取引所理事長・北村恭二さん 【大阪】
円高対策へ優遇税制 技術開発など支援、中小企業を救済 新規立法も－－通産省
科技系大学院生の研究から創業まで最大2400万円までを無償支援－－岡山 【大阪】
SelectRelevants
Contents
Advanced Web Search
Search with...
this boolean expression
(ベンチャ) AND (支援) AND (確立 OR 設定 OR 設置 OR 設立 OR 新設 OR 開業)
FIND
Use terms such as AND, AND NOT, NEAR
sorted by
Pages with these words will be ranked highest.
SEARCH: Worldwide or Select a country
RESULTS IN: All languages Japanese, English
AltaVista found 56,481 results About
q??
... でのブランド確立が課題 政府のプログラムを評価、支援まだ足りない ...
伊丹～石垣線新設などで4.9%増 ... 調査委員会設立 ····· 3 県・市が共

まとめ

- 適合的汎化に基づく情報検索システム
  - 適合文書に基づき、検索語の抽象度を設定
  - 既存の概念カテゴリーではなく、適合文書群に即した一時的な概念カテゴリーを形成
    - 概念の粒度が異なる概念体系辞書においても、同等の検索性能
- 今後の展望
  - 自動構築されたシソーラス（特定の意味のあるキーワード集合でも可）の利用