意味構造に基づく検索システム

背景

- 大規模な機械可読テキストデータの流通・計算機性能の向上・統計的アプローチ

  ⇒ 高性能な構文解析器が研究レベルで手軽に利用可能

- 知識(獲得)や推論に関しては、まだ「手軽に利用可能」というレベルではない。

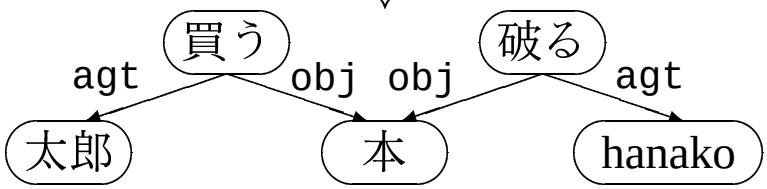
⇒ **人間と計算機との協調　(人間: 推論 ⇔ 計算機: 探索)**

Global Document Annotation

XML のインスタンスであり、タグの構造と属性によって文書の統語や意味に関する構造を明示するための枠組

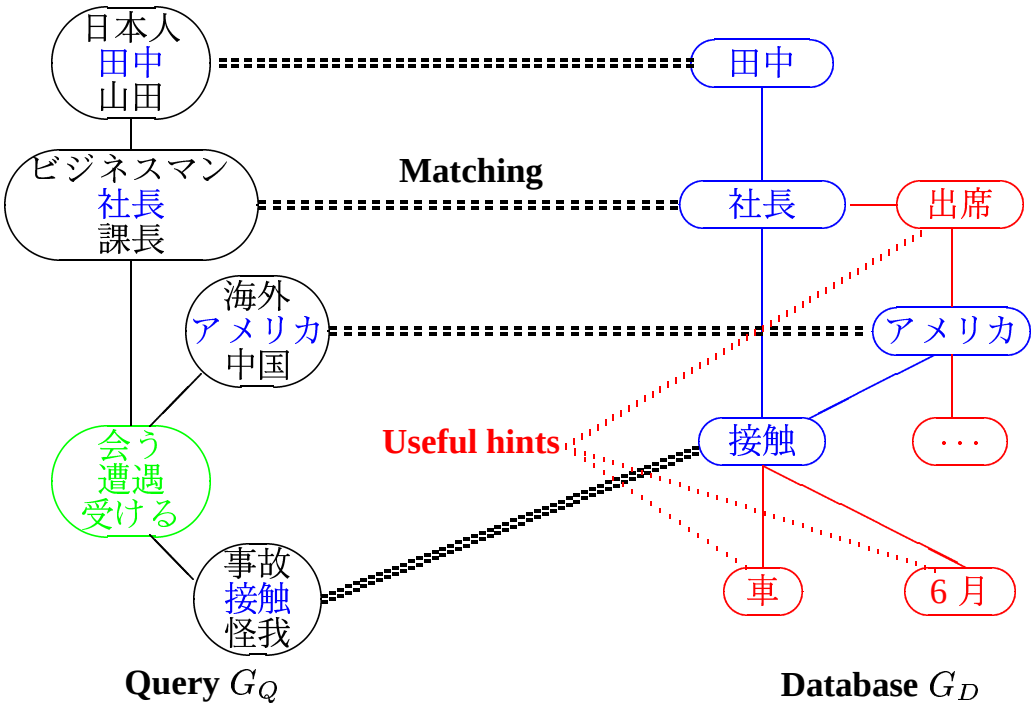
```
<su><adp opr="obj">
        <np><adp opr="agt">太郎が</adp>
            <v obj="mcn">買った</v>
            <n>本</n></np>を</adp>
    <v agt="hanako">破った</v></su>
```

⇓



「太郎が買った本を破った」に対するアノテーション (上) とその意味構造 (下)

情報検索における意味構造



検索に意味構造を使う利点

高い検索精度：

入力したキーワードがたまたま出現しているだけで内容としては無関係な文書が排除されるので、より正確な検索を行なえる。

より細かいヒントの提示：

文書群を予め解析しておけば、その情報をもとに「入力した語と内容的に共起しやすい語」といった、より細かいヒントを提示することができる。

ユーザの意図の適切な表現：

グラフ構造なら、述語論理程度の内容が表現できる。

検索に意味構造を使う欠点

低い再現率：

条件が厳しくなるので、そのままでは再現率が下がる。ユーザ・システム間のインタラクションが重要。

‘正しい’意味構造：

評価実験では首を傾げるような構造がかなりあったが、これはむしろシステムの問題

解析コスト・インデックスサイズ：

現在のところ、プレーンテキストで数KB×100万文書程度なら問題なさそう。

グラフの埋め込み

- NP-hard 問題 [Zhang et al 1996]
- 問合せグラフは十分小さい。
  ⇒ 動的計画法で候補を数え上げ、一致度を再計算
  (必ずしも厳密な解が得られるわけではない)
- 簡単のため、無向グラフで辺にはラベルがないと仮定

検索例

1994 年の毎日新聞約 10 万記事を自動解析したものを検索対象として、**「ロボットを使って住宅を安く作る」**という文書を検索

1. 「ロボットを使って住宅を安く作る」という文を入力
   ⇒ 質問文を構文解析し、文書群に対してグラフ照合
2. システムが提示した類義語を参照して、「利用」「家屋」「建設」といった類義語・関連語を追加
   ⇒ 目的の文書に到達

質問と解候補の構造を使った類似度計算の効果

	順位	時間 (分)	操作数
キーワードのみ	32.71 (36.64)	18.01 (12.00)	27.62 (10.76)
構造も利用	1.50 (0.71)	7.62 (4.46)	13.62 (6.32)

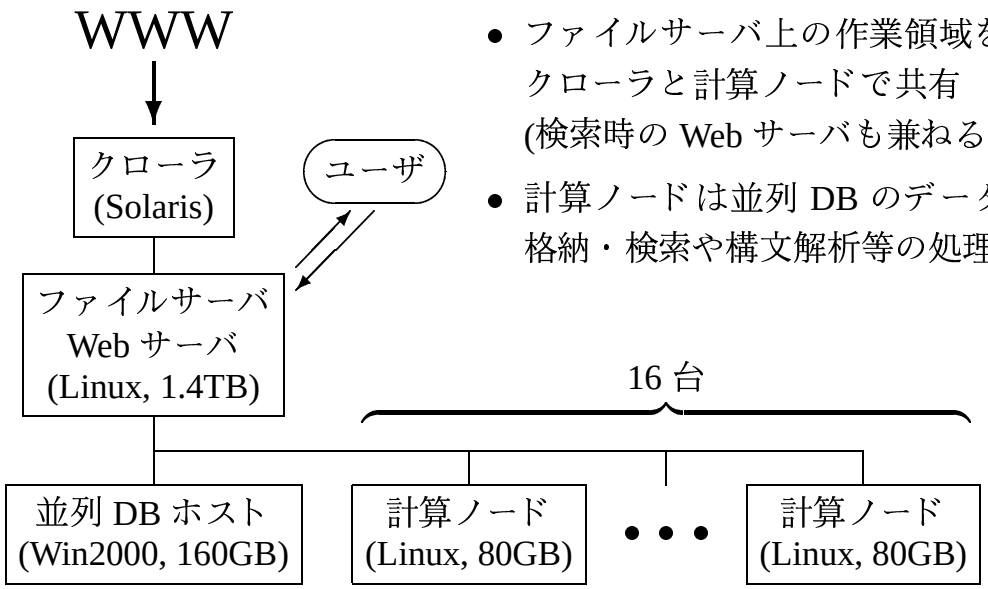
(平均と標準偏差)

- 提示した条件に合致する記事を 1994 年毎日新聞記事約 10 万件の中から 1 件以上探すという課題 4 題を、8 人の被験者で評価
- 各課題には時間制限を設けず、被験者が「この文書だ」と解答した時点で課題終了
- 正解に達した人数はそれぞれ 4 人、6 人でほぼ同数

大規模化に伴う課題

- クローリング
- フィルタリング・クリーニング
- 形態素解析・構文 (係り受け) 解析
- インデキシング
- 並列化

システム構成



- ファイルサーバ上の作業領域をクローラと計算ノードで共有 (検索時の Web サーバも兼ねる)
- 計算ノードは並列 DB のデータ格納・検索や構文解析等の処理

クローラ

- GNU `wget` を使用
- リトライ・差分ダウンロードなどの機能をそのまま利用
- 機械処理しやすい形でログを出力するように若干修正

フィルタ・クリーニング

- 文字コードを euc-jp に変換した後、正規表現で記述したパターンに基づいて加工 (`flex`)、euc 文字の数および比率で ‘日本語’ かどうかを判断
- 現在のところ、プレーンテキスト・HTML・XML のみに対応としているが、PDF やワードファイルがかなり多い。

解析・インデキシング

- 形態素・文節係り受け解析: 統計的な解析器
- 類義語・隣接語の抽出: 二分探索を基本とする独自の DB
- キーワード検索・グラフ照合: 並列 DB「高性能並列情報検索システム」(三菱電機)
  一文書の情報を一レコードとして DB に格納し、グラフ照合アルゴリズムをユーザ定義関数として実装

前処理に必要な時間

7万個のURLを起点として、リンクを5段まで辿ってページを収集 (いずれも新規登録時)

クローリング・解析		
取得ページ数:	150万	
有効ページ数:	126万	
時間:	7日	(400 URL/h)
インデキシング		
独自DB (類義語・隣接語)		3.5日
並列DB (キーワード検索・グラフ照合)		3日

まとめ

- 意味構造に基づく検索では、(再現率を補うために)ユーザとのインタラクションが重要
- 検索精度よりも、検索・インタラクションの効率向上の方が有望
- 計算機性能や構文解析技術の向上によって、数KB×100万文書程度の規模ならば実用的に運用可能

今後の課題

- 被験者による、さらなる評価実験
- PDFやワードファイルなど、使用頻度の高いファイル形式への対応
- システムのパッケージ化・配布 (逐次版)

(研究協力: アリゾナデザイン、三菱電機)