第2週B4 ゼミ

Wikipediaマイニングによるシソーラス辞書の構築手法

出典

- 中山 浩太郎 ,原 隆浩 ,西尾 章治郎 Wikipediaマイニングによるシソーラス辞書の構築手法(情報検索) 情報処理学会論文誌 Journal Article (2006) Vol. 47 No. 10 pp. 2917-2928

はじめに

- シソーラス辞書
  - 言葉を同義語や意味上の類似関係、包含関係などによって分類した辞書
- 「全文検索システム」
  - 「日本」、「Japan」、「JPN」

背景・概要

- 現在のシソーラス
  - 作成時と利用時のタイムラグにより最新の語への対応が困難
- WikipediaにWebマイニングの手法を用いる
  - シソーラス辞書を自動構築

関連研究

- Webマイニング
  - Web内容マイニング
    - Webページの内容
  - Web利用マイニング
    - 利用者の行動履歴、利用ログ
  - Web構造マイニング
    - Webサイトの構造、ページ間の関係

関連研究

- 自然言語処理によるシソーラス辞書構築
  - 共起関係、フィルタリング、クラスタリング
  - 多義性、同義語、曖昧性、精度
- Webマイニングによるシソーラス辞書構築
  - リンクテキスト
  - 同義語、多義語

WebコーパスとしてのWikipedia

- ハイパーリンクによる記事同士の参照
- 高密度なリンク構造
- 辞書更新の即時性
- コンテンツの網羅性

WebコーパスとしてのWikipedia

- ハイパーリンクによる記事同士の参照
- Wikipedia
  - 多数のリンクで構成される
- 高密度なリンク構造
  - 65万ページで約715万の内部リンク

WebコーパスとしてのWikipedia

- 辞書更新の即時性
  - リアルタイムで記事が公開・アップロード
- コンテンツの網羅性
  - リンク構造は閉じられている
  - 幅広い分野の記事が網羅されている

リンク構造の解析

- Forward Link
  - 現在ページから別のページへのリンク
- Backward Link
  - 別のページから現在ページへのリンク

- これらが存在すると2つのページは関連がある
  - 最も簡単なアプローチ

リンク構造の解析

- 問題点１
  - リンクの有無だけでは関連度が計算できない
- 問題点２
  - 記事の作成者の主観に依存する
- 問題点３
  - 隠れた関係を発見できない

- 語をノード、リンクをエッジとする有効グラフを生成
- 距離n以内のノードを再帰的に探索し、関係の強さを計算

距離

- Forwardリンクの総数を|F_{pi}|関連度を除算してリンク先のページの関連度として加算

- S:関連度
- p:ページ

Algorithm $RE(p_i, weight, depth)$

1 **if** $depth > n$ **then return**;
2 $F_{p_i} = GetForwardLinks(p_i)$;
3 **for each** $(p_j) \in F_{p_i}$ **do**
4 $score = weight/|F_{p_i}|$;
5 $S_{p_j} = S_{p_j} + score$;
6 $RE(p_j, score, depth + 1)$;

同義語･多義語の抽出

- 同義語
  - 現在ページへのBackwarリンク

- 多義語
  - B:Backwardリンク　w:検索語

$$CS(p_i, w) = \frac{|B_{p_i,w}|}{\sum_j |B_{p_j,w}|}$$

シソーラス辞書の更新

- 1 更新日付を比較し、更新されたページを抽出
- 2 旧シソーラス辞書の中であるページに関連語を持つページを抽出
- 3 抽出したページに対し関連度を再計算
- 4 リンク構造解析によりあるページから距離n以内のページを抽出
- 5 再度関連度を計算

評価

- 探索距離が増加するごとに計算量はO(n^2)オーダで増加
-

単語	1 ホップ	2 ホップ	3 ホップ
Nintendo	0.05 sec., 328 ノード	6.63 sec., 53981 ノード	1129.03 sec., 9973687 ノード
apple	0.03 sec., 208 ノード	3.66 sec., 24217 ノード	380.27 sec., 3022035 ノード
iPod	0.04 sec., 159 ノード	1.71 sec., 11645 ノード	205.36 sec., 1647562 ノード

評価

- 語から連想できる語を「関係ある語」

$$CP = \frac{\text{発見された，関係が深い概念の数}}{\text{発見された，すべての概念の数}}$$

探索距離	トップ 10	トップ 20	トップ 30
1 ホップ	66.7%	64.2%	61.2%
2 ホップ	93.2%	86.2%	83.1%
3 ホップ	91.4%	89.4%	85.9%

今後の展開

- 即時性の高い語彙の抽出
- 多言語Wikipediaによるシソーラスの構築
- N-gram解析によるドメイン特有概念の発見、リンクの共起性解析による精度の向上