

ベンチャー企業検索システムの開発

三澤 佳裕



背景と目的

- ・ インターネットに検索エンジンは必要不可欠
- ・ ドメイン(ベンチャー企業)に特化したロボット型検索エンジンの作成
- ・ Webサイトから効率的にデータ収集する手法の検討
- ・ Webサイトを評価する手法の検討

開発環境

言語

Python

データベース

MySQL

- 構文解析ソフト

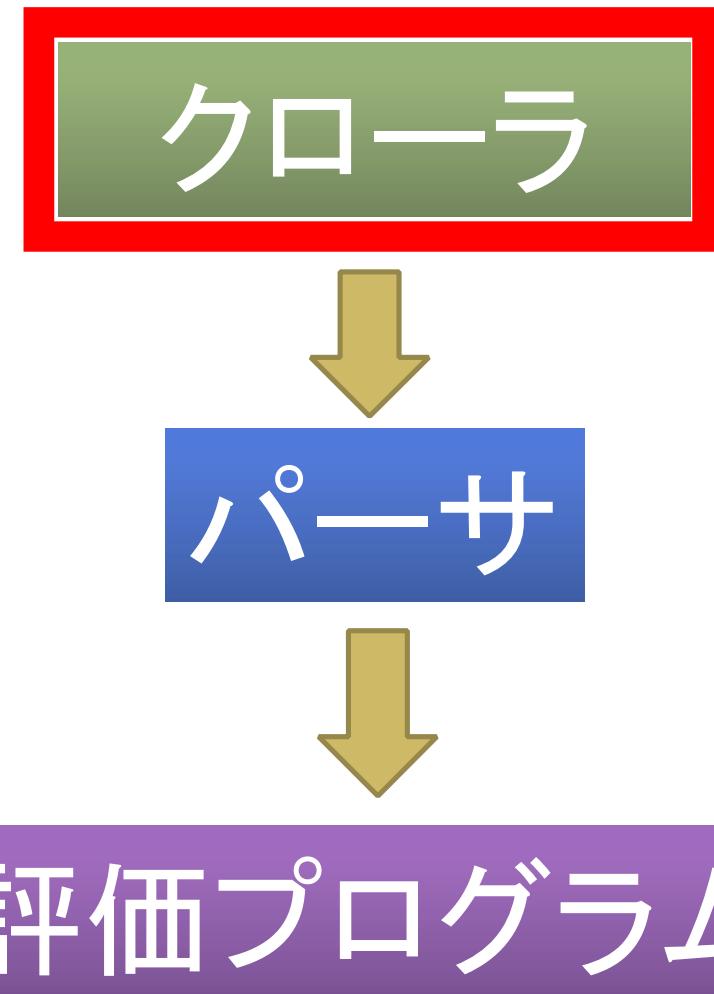
MeCab

NAIST(奈良先端科学技術大学院大学)のMeCab用辞書データ

- サーバスペック

| | CPU | RAM |
|------|-------------------------|-------|
| サーバA | Pentium4 3.2GHz(1C/2HT) | 3GB |
| サーバB | Atom 330 1.6GHz(2C/4HT) | 2GB |
| サーバC | i3-2100 3.1GHz(2C/4HT) | 12 GB |

検索エンジンの概要



クローラについて

- Webページ(HTML)を解析してリンクをたどるもの
- たどったWebページはデータストアに保存
- 処理効率化の工夫
 - 複数サーバに処理を分散
 - ソケット通信を用いてディスパッチャとワーカで情報をやりとり
 - サーバスペックに最適化
 - サーバのCPUコア数に応じてクローラプロセスを生成

ディスパッチャ

クロール対象のサイト情報を共有

ワーカ
(サーバ)

ワーカ
(サーバ)

ワーカ
(サーバ)

クローラを生成/起動しデータ収集を開始

クローラ
プロセス

検索エンジンの概要

クローラ



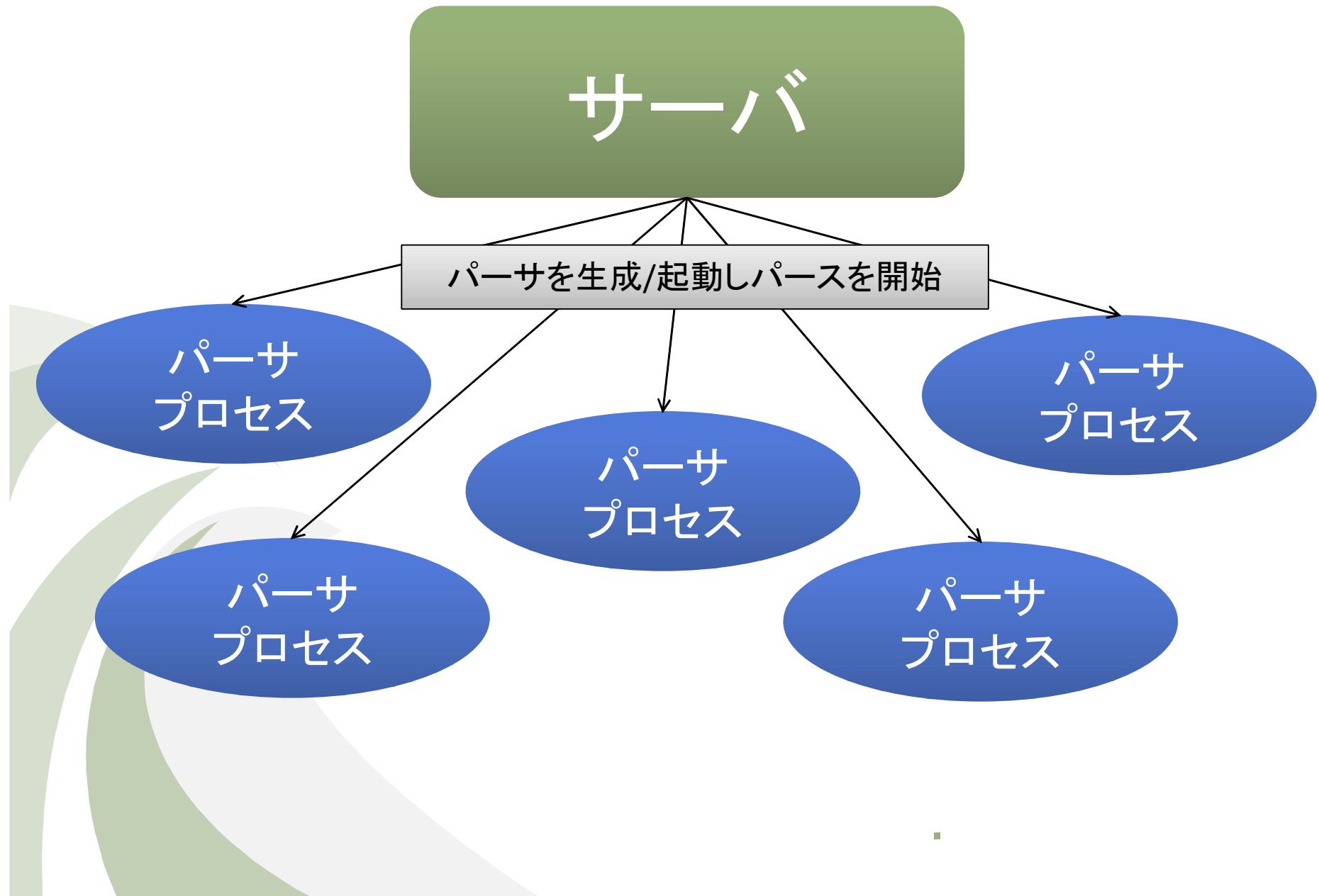
パーサ



評価プログラム

パーサについて

- HTMLからタグ情報を除去した本文を抽出
- 構文解析器を用い本文を品詞毎に分類
- 処理効率化の工夫
 - ストップワードの除去
 - 大文字/小文字の統一
 - サーバスペックに最適化
 - サーバCPUコア数に応じてパーサプロセスを生成



検索エンジンの概要

クローラ



パーサ



評価プログラム

評価プログラムについて

- ・ ベンチャー企業を推定する
- ・ 企業サイトを1つのドキュメントとしコーパスを生成する
- ・ コーパスを元に以下の手法で推定評価する
 - TF-IDF(Term Frequency-Inverse Document Frequency)
 - LSA(Latent Semantic Analysis)
 - LDA(Latent Dirichlet Allocation)

TF-IDFについて

- 値が大きいほど重要
- 代表キーワードを求めるのに最適
- TF(Term Frequency)
 - ドキュメント内のキーワードが使用されているかの指標
 - キーワードが多く含まれるに連れてキーワードについてより詳しく説明している
- IDF(Inverse Document Frequency)
 - キーワードがどれだけのドキュメントで使用されているかの指標
 - 特定のキーワードが多くのドキュメントに含まれているよりも少ないほうが重要

LSAについて

- 潜在的意味解析
- キーワードやドキュメントの意味を学習し、ベクトルとして表現する
- キーワード間やキーワード、ドキュメント間の類似性を求めることができる

LDAについて

- 潜在的ディレクトリ配分法
- そのトピックをドキュメントから教師なしで推定する手法
- ドキュメントの背景にあるトピックが何かわかる

実験結果

- ベンチャー企業の定義
 - 「リスク」、「挑戦」、「冒険」、「ベンチャー」、「先進」
- 150の企業サイトをクロール
 - 60件をベンチャー企業と明示している企業サイトから抽出
 - 90件をWikipediaの企業リストから抽出

実験結果

- TF-IDFによる結果
 - リスク:
 1. www.prospect.ne.jp
 2. www.arai-medphoton.com
 3. lafla.co.jp
 - 挑戦:
 1. www.motor-solution.co.jp
 2. www.ishii-pt.co.jp
 3. www.bna.jp
 - 冒険:
 1. www.toli.co.jp
 2. www.yamaha-motor.co.jp
 - ベンチャー:
 1. www.kairospharma.co.jp
 2. www.cbt-bio.co.jp
 3. www.sleepwell.co.jp
 - 先進:
 1. www.sleepwell.co.jp
 2. www.healthcare-systems.co.jp
 3. www.hondakinzoku.co.jp

LSAによる結果

リスク:

1. skyplatform.co.jp
2. biohydrogen.co.jp
3. www.arai-medphoton.com

挑戦:

1. machine-intelligence.co.jp
2. www.axelspace.com
3. www.protek.co.jp

冒険

1. www.kao.com
2. kyokuyoamerica.com
3. www.gembu.co.jp

ベンチャー

1. www.unipres.co.jp
2. www.elmo.co.jp
3. www.orthree.jp

先進

1. www.bna.jp
2. www.lecip.co.jp
3. hydrokraft.net

評価

- TF-IDFよりLSAが良いと考える
 - TF-IDFは全文検索に最適である
 - LSAは推論により、より多くのキーワードに対応できる
- 企業サイト150件をクロールし約80件をデータとして扱える形にできた
 - 文字コード識別に不具合があり、扱えないデータがあった
- ストップワードの除去の仕方に問題あり
 - 推論された値に「し」などの、その単語自体では理解不能なものが結果上位に出力された

まとめ

- 検索エンジンのクローラ、パーサの分散処理による実装
- 評価プログラムの実装

今後の課題

- LDAの評価プログラムの改善
- クローラ、パーサだけでなく評価プログラムの分散処理の実装



ご清聴ありがとうございました