

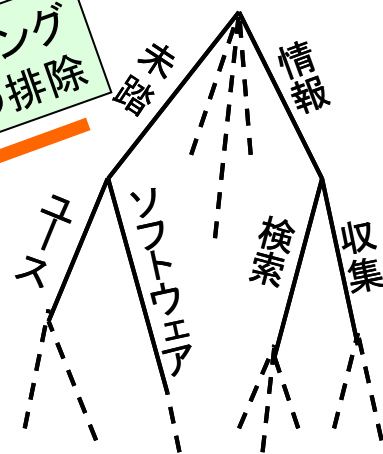
テキストマイニング技術を融合したウェブブラウザの開発

開発者: 上村卓史

高性能なキーワード抽出エンジン + 情報収集を補助するインターフェース

文書要約:
重要部分を抜粋して表示

スコア計算・ランキング
類似キーワードの排除



アドレス <http://www.ipa.go.jp/jinzi/esp/2006youth...>

抽出アルゴリズム
単独 + 0 0 0
関連サイト 要約 ハイ

- ページ (8)
- キーワード (5)
- ウェブブラウザ (3)
- テキスト (3)
- 単語 (3)
- ユーザ (3)
- 高精度 (2)
- 効果的 (2)
- プロジェクト (1)
- 高速 (2)
- 情報 (2)
- エンジンを (2)
- 開発 (4)
- 抽出 (7)
- 機能 (4)
- 検索 (4)
- 本 (2)

6. テーマ概要 (応募時資料を一部分改訂)

ウェブブラウザの世界では激しい開発競争が行われているが、現在その多くは多機能化に焦点が当てられている。そのためウェブブラウザの機能向上は、検索エンジンを駆使して候補を絞り込み、あとは実際にページを見て内容を確認する必要がある。本プロジェクトではテキストマイニングの技術を用いて、ウェブブラウザの機能を向上させる。その中で本プロジェクトではテキストマイニングの技術を搭載することで、ブラウジングの効率を改善する。

1. テキストマイニングエンジンの開発

ユーザが閲覧したページを蓄積し、重要なキーワードは、単語Nグラム(N個の単語のつながり)を表示する。より大きなテキストが扱える。また質の面でも、本来存在しない単語は扱わないため、高精度な単語Nグラムを生成する。

2. ブラウザとの統合

1)のエンジンを用いてウェブブラウザに以下の機能を実装する。

- キーワード表示(自動抽出によるキーワードが表示され、1クリックで検索等が可能)
- 自動リンク(関連サイトを自動的に検索して提示)・自動キーワードハイライト(特に重要なキーワードをハイライト)
- ページの要約表示

ユーザが実際に閲覧したページやクリックしたキーワードがテキストマイニングの重要なデータとなり、ブラウジングとマイニングが相互に影響を与え合う関係を作ることがこの二つの融合だといえる。

7. 採択理由(担当PMからのコメント)

キーワード表示:
ページ内／ウェブ検索、
ハイライトがワンクリックで

「単語Nグラム方式」の索引

検索語入力支援:
文書内の単語を使って予測

関連サイト表示:
キーワードを使って検索し、
関係の深いページを抽出

形態素解析(MeCabを利用)
文章のみを抽出

Webページ
Webユーザ
Web検索

検索 [Web]

前へ 次へ 強調 大文字/小文字を区別 検索履歴表示