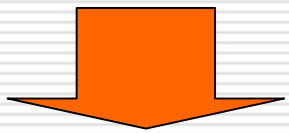
平成15年度未踏ソフトウェア創造事業

blogページの自動収集と監視に基づくテキストマイニング

奥村学，南野朋之，藤木稔明，鈴木泰裕
（東京工業大学 精密工学研究所）

開発の目的

- アメリカでは，hostingサービスを使用したblogが多数を占めるに対し，日本では，サービスとして運用されている ものだけでなく，Webページと変わらず個人が各自書いているものが多数を占める
- 定期的な監視を網羅的に行うことはそれほど容易ではない



- クローリングしたHTML文書を解析し，そのWebページがblogであるかを判定することによってblogの収集を行う
- blogツールなど，特定のシステムを利用していないWeb日記なども広く収集することが可能
- RSSベースの収集では過去の記事を収集できないのに対し，過去のものまで収集することが可能

システムの主な機能，特徴

- 収集したblogの検索が可能
- ホットキーワード抽出でホットな話題をチェック！
- おすすめblogも提案してくれる
- 評価表現抽出を利用した評判情報検索もできます！

システムの評価

- 構築したシステムを2週間運用
- データに偏りはまだまだあるものの…
- blog判定の精度は94.3%
- 推定再現率は83.8%
- 82%がblogツールやhostingサービスを使用していない
- 過去のものも取得できている