Wikipediaカテゴリを用いた ブログ著者の得意分野プロファイリング

野田陽平　清田陽司　中川裕志　(東京大学)

概　要

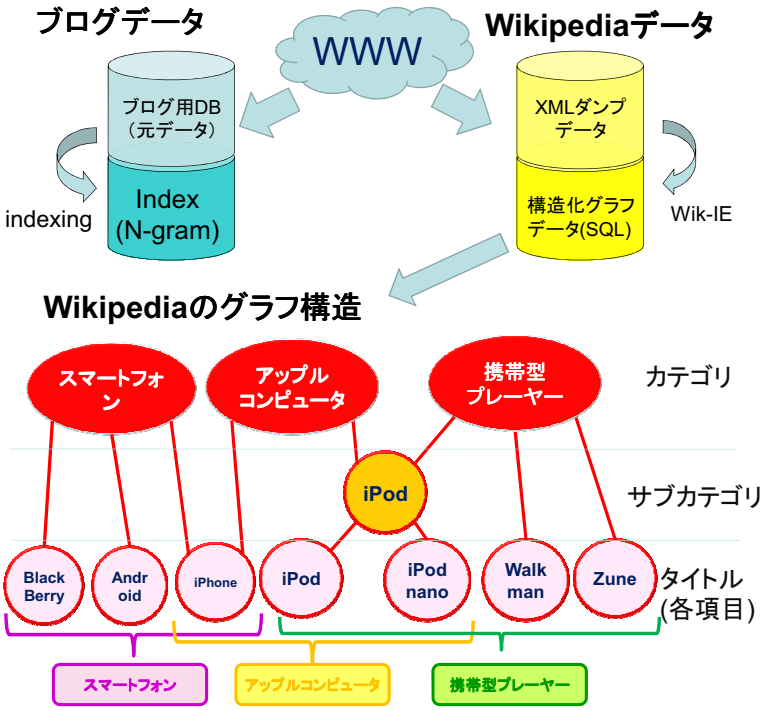
本研究では，ブログ投稿者の得意分野のプロファイリングを行う．ブログ投稿者の中には，特定の分野に精通し，非常に有用なブログ記事を投稿する者も存在する．しかし，膨大なブログサイトの中から特定分野に特化したブログサイトを的確に抽出することは，明確な分類体系が存在しないブログにおいては困難である．また，投稿者の投稿記事がひとつの分野のみに特化しているとは限らないため，投稿者を単一の分類にマッピングすることは困難である．そこで，非常に有用な集合知であるWikipediaの構造に直接ブログ記事をマッピングし，投稿者の投稿行動の傾向を観察することで，あらゆる分野の専門家をブロガーの中から発見することを目指す．

研究の目的

◆特定分野に関して注力して投稿活動を行っているブログ投稿者を発見する(ブログ上での専門家の発見)

◆専門分野の分布情報を用いた評価分析

ブロガーの得意分野推定



従来のブログ分類･αブロガーの発見

=ブログ分類=

●辞書を使って文書から作成した単語ベクトルを元にクラスタリング

●Wikipediaのカテゴリ情報を使っての文書の2値分類

=αブロガーの発見=

●αブロガーを"Agitator"と"Summarizer"に分け、リンク構造や特定の話題への言及のタイミングで分析

BlogとWikipediaの融合

Blog

◆ブログサイト数　1,690万
◆総記事数　13億5000万件

個人による最新情報などが日々投稿されるCGM

■製品、サービスなどに関する評価情報

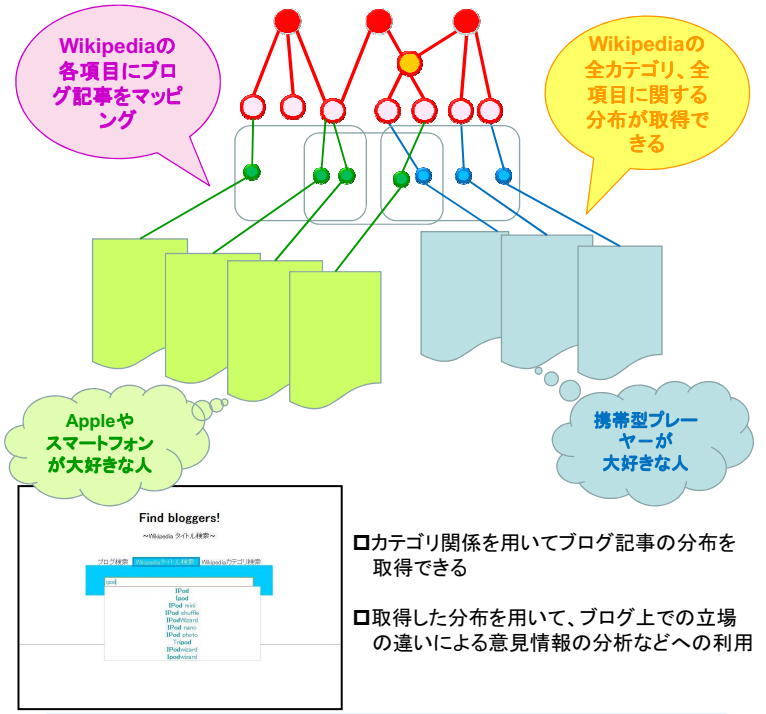
Wikipedia

◆50万項目突破(2008年6月)
⇒日々更新
◆日々更新が続けられている有用な集合知
◆カテゴリとタイトル項目のマップ構造

各項目ごとに意味的なつながりがある

Blog記事をWikipediaにマッピング

ブログの中の"隠れた専門家"を発見する



□カテゴリ関係を用いてブログ記事の分布を取得できる

□取得した分布を用いて、ブログ上での立場の違いによる意見情報の分析などへの利用

今後の予定

●時系列での各項目の言及頻度分析

●コンスタントに特定の項目に関する記事を投稿しているブロガーの追跡

●スプログへの対応

●Wikipediaカテゴリを利用した評価分析