情報アクセス論 第9回
「ソーシャル検索」

ソーシャル検索とは?

- 検索のプロセスに一般利用者のコミュニティが積極的に関与
  - すべての利用者を同一に扱う従来の検索手法とは大きく異なる
- 従来の検索手法との違い
  - 従来の検索は，利用者とシステムとのやり取り
  - ソーシャル検索は，利用者は他の利用者と直接的あるいは間接的にやり取り

ソーシャルメディア

- ソーシャルメディアとは?
  - 利用者が作成するコンテンツ
    - Consumer Generated Media（**CGM**）と呼ばれる
  - 利用者が自分や他人のコンテンツに付与したタグ
  - 利用者がブックマークやタグを他のユーザと共有
- ソーシャルメディアの例:
  - ブログ，動画共有サイト，SNS（Social Networking Service），レビューサイト，Q&Aサイト，電子掲示板，ソーシャルブックマーク

「タグ」と手動による索引付け

- **昔**：**図書館のカード目録**
  - 索引語は検索しやすさを考慮して選ばれる
  - 専門家が索引語を生成
  - 質が非常に高い
  - **統制語彙**から索引語が選ばれる（用語の統一）
- **今**：**ソーシャルメディアタギング**
  - タグは必ずしも検索しやすさを考慮して選ばれない
  - 一般利用者がタグを生成
  - タグにはノイズや誤りが多い
  - **フォークソノミー**からタグの語が選ばれる

統制語彙とフォークソノミー

- **統制語彙**
  - 主題を表す索引語を，あらかじめ決められた閉集合の中から選ぶ
    - 例：「**オートバイ**」→「バイク」「二輪車」「単車」「自動二輪」
- **フォークソノミー**
  - folk（民衆）+taxonomy（分類法）を合わせた造語
  - あるコンテンツに対して多くの利用者が自由にタグを付与
  - 付与されたタグによってコンテンツを分類・検索できる

タグの種類

- 内容に基づくタグ
  - 自動車，立命館，空
- コンテキストに基づくタグ
  - 京都，金閣寺
- 属性を示すタグ
  - ニコン（カメラの種類），白黒（画像の種類），ホームページ（Webページの種類）
- 主観的なタグ
  - かわいい，すごい，見事な
- 整理用のタグ
  - あとで読む，私の写真，README

タグの検索

- 利用者が付与したタグで検索するのは難しい
  - タグが数個しか付与されていない
  - タグ自体が非常に短い
- ブーリアン・確率・ベクトル空間などの検索モデルをそのまま使ってもうまく行かない
- 問合せとタグの**語彙のミスマッチ**の問題を解決する必要がある
  - 問合せ拡張と同様に「タグ」拡張を行うことで解決できる
    - シソーラス，Web検索結果，問合せログなど

タグの推定

- タグ拡張の問題点
  - タグがまだ付与されていない場合は拡張ができない
  - タグには誤りやノイズが多い
- 人気が高いものは見つけやすいが，人気がないものは見つけにくい
  - タグが少ない，あるいはまったく無い
- タグが少ない，あるいは無いものに対して，タグを自動で付与できないか？

タグを推定する手法

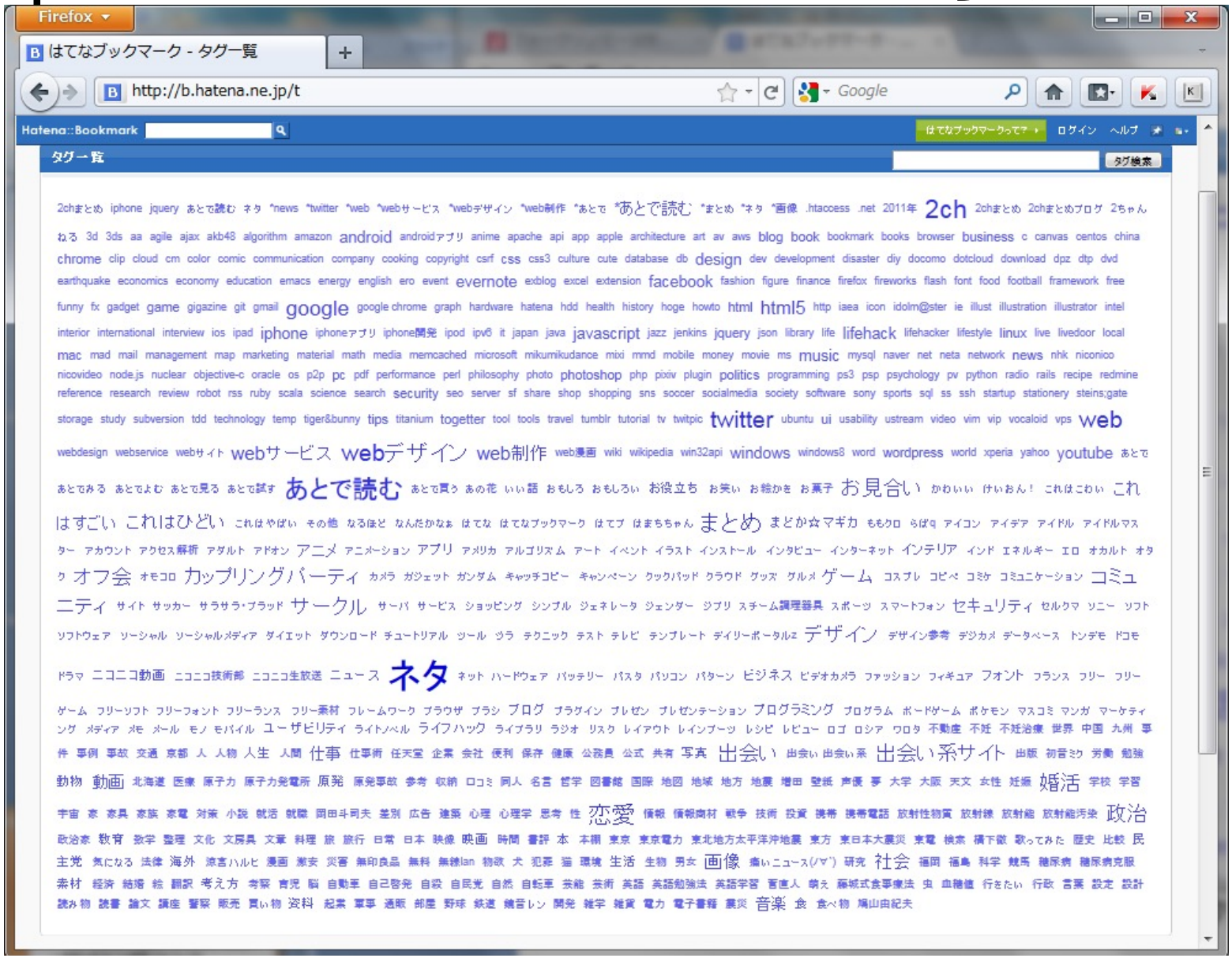
- **TF･IDF**
  - 対象の中でTF･IDFの値が高いタグを提案
  - テキストで表現できる対象に限られる
- **分類**
  - 各タグに対して2値の分類器を学習
  - 良く使われるタグには有効
- **MMR**（Maximal Marginal Relevance）
  - 既存のタグに対して新規性があり，かつ対象に適合しているタグを見つける

$$MMR(t;T_i) = \left(\lambda Sim_{item}(t,i) - (1-\lambda)\max_{t \in T_i} Sim_{tag}(t_i,t)\right)$$

タグクラウド

- 検索は，関心がある対象を見つけるのに有効
- ブラウジングは，タグがつけられた対象の集合を探索するのに有効
- タグの集合を可視化する様々な方法がある
  - タグのリスト
  - タグクラウド
  - アルファベット順
  - カテゴリ別
  - 人気順

タグクラウドの例
（はてなブックマークのタグ一覧）



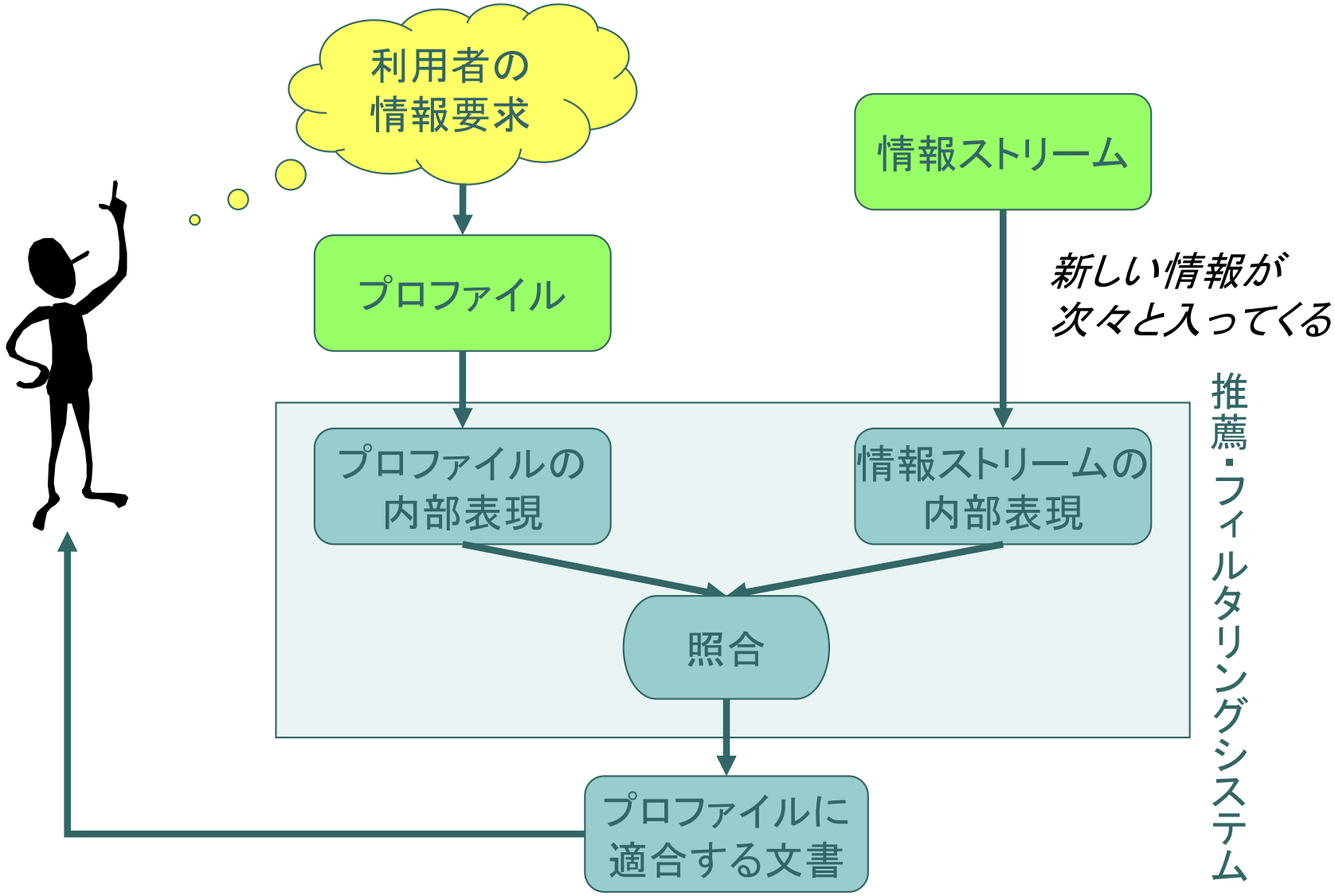
情報推薦・情報フィルタリング（1）

- **情報過多／情報洪水／情報爆発**（情報の氾濫）
  - すでにある膨大な情報の蓄積だけではなく，新しい情報が次々に入ってくる
    - 蓄積された情報は検索システムで検索できる
  - 日々新たに入ってくる情報のすべてに目を通すのは不可能
  - 有用な情報は見落とすことなく手に入れたい

情報推薦・情報フィルタリング（2）

- 次々と入ってくる新しい情報に対して，利用者の興味のある情報のみを選別して提示
  - 例：ハードディスクレコーダの自動録画機能
- 興味のない情報（有害情報など）の除外も含まれる
  - 例：スパムメール（迷惑メール）のフィルタリング
- 利用者の興味（情報要求）は「**プロファイル**」と呼ばれる（情報検索における問合せに相当）
  - 例：RSSリーダに登録したキーワード

情報推薦・情報フィルタリングのモデル



情報推薦・情報フィルタリングと情報検索

- 情報検索の技術を応用可能
  - 興味をキーワードで表現すれば，それと照合することで実現可能
- 情報検索との違い
  - 情報検索では，問合せは**動的**（短期的）で文書集合は**静的**
  - 情報推薦・フィルタリングでは，プロファイルは**静的**（長期的）で情報ストリームは**動的**
  - あらかじめ索引付けができない

情報推薦・情報フィルタリングの分類(1)

- **内容に基づくフィルタリング**(cognitive filtering)
  - 情報それ自身の内容と利用者のプロファイルを比較し，その関係を基にフィルタリング
    - 情報検索の手法が適用できる
    - 例：図書館のSDIサービス，Googleアラート

使用中の Google アラート

検索用語	タイプ	頻度
"Dublin Core"	ニュース、ウェブ	1 週間に 1 回
セマンティックウェブ	ニュース、ウェブ	1 週間に 1 回
"ダブリンコア"	ニュース、ウェブ	1 週間に 1 回
"ディジタルアーカイブ"	ニュース、ウェブ	1 週間に 1 回
"情報図書館"	ニュース、ウェブ	1 週間に 1 回
"情報検索"	ニュース、ウェブ	1 週間に 1 回
ディジタル図書館	ニュース、ウェブ	1 週間に 1 回

情報推薦・情報フィルタリングの分類(2)

- **社会的フィルタリング**(social filtering)
  - 情報の送信者の特徴や受信者(あるいは所属組織)との関係に基づくフィルタリング
    - 自分の上司からのメールは重要
    - スパムメールのフィルタリングにも使われる
- **経済的フィルタリング**(economic filtering)
  - 情報を得ることによる利益と，それに必要な対価の比に基づくフィルタリング
    - 課金などだけでなく，情報の長さなど心理的な要因も含む

情報推薦・情報フィルタリングの分類(3)

- **協調フィルタリング**(collaborative filtering)
  - 「同じ興味を持つ人は同じ情報を求めている」
  - 自分が読んだ情報の印象などを記録し，他のユーザのフィルタリングを手助けする
    - 例：Amazonの「おすすめ度」「カスタマーレビュー」「この商品を買った人はこんな商品も買っています」
  - スパムフィルタリングへの応用
    - 他のユーザがスパムと判定したメールの送信元IP，ドメイン名，本文などをブラックリストに登録

協調フィルタリングの手法

- ユーザAとユーザBの嗜好に高い相関がある場合
  - ユーザAが高く評価したものをユーザBに推薦

商品	Ken	Lee	Meg	Nan
a	1	4	2	2
b	5	2	4	4
c			3	
d	2	5		5
e	4	1		1
f	?	2	5	?

相関係数法(1)

- 評価値ベクトルの**相関係数**でユーザ間の類似度を測る
  - Ken(K)とLee(L)の相関係数

$$r_{KL} = \frac{\sum_i (K_i - \overline{K})(L_i - \overline{L})}{\sqrt{\sum_i (K_i - \overline{K})^2} \sqrt{\sum_i (L_i - \overline{L})^2}}$$

  - $\overline{K}, \overline{L}$はそのユーザの評価値の平均

相関係数法(2)

- 各ユーザの評価値に類似度で重み付けして足し合わせる

$$predict(K, f) = \overline{K} + \frac{\sum_{J \in user} (J_f - \overline{J}) r_{KJ}}{\sum_{J} |r_{KJ}|}$$

- 正の相関があるユーザの評価がより反映される

情報推薦・情報フィルタリングの研究課題

- プロファイル（利用者の嗜好）の抽出
  - 既に読んだ文書そのものを使う
  - 情報の注視時間から興味を推測
- 重要な情報を見落とさない工夫
  - ベイジアンフィルタリングによる機械学習
    - 単純ベイズ分類器をフィルタリングに応用
    - 利用者が過去に判定したスパム／非スパムメール中の単語の出現頻度から，新しいメールがスパムである確率を計算（使っていくうちに誤判定が減っていく）

まとめ

- 近年利用者が増加しているソーシャルメディアを用いた検索技術について説明した
- 一般利用者が付与する「タグ」と，その利用法について述べた
- 情報推薦および情報フィルタリングの技術について説明した