

2009/9/13 SBM勉強会

言語表現に基づく

ブックマークコメントの分類とフィルタリング



山田 剛一, 本 晶夫, 高田 彰, 絹川 博之
東京電機大学 未来科学部



自己紹介

- 山田 剛一（やまだ こういち）
- 自然言語処理の研究者
 - 意味解析 → 情報検索 → Web ...
- ソーシャルブックマーク関連では...
 - スパマー検出もやっています
 - 「はてバーぶろぐ」はスパマー判定されるかも



ソーシャルブックマークは熱いか？（１）

- 今現在の話題性としては断然 twitter
 - 一般人にわかりやすい
 - 「つぶやき公開」は説明不要
 - 対して「ブックマーク公開」は意味不明
 - 速報性のメリットもわかりやすい
 - 本体の機能が単純
 - 西谷さんも研究会を立ち上げ



ソーシャルブックマークは熱いか？（２）

- はてブが 2.0 になる時の熱気はどこに？
 - なにかが変わるという期待感があった
 - ユーザ数は増えているが？
 - お気に入りを使いこなして情報を集めようとする層は限られている
 - それにお金を出す層はさらに少ない
 - 「ソーシャルブックマーク」の範疇を越えない限り、一般人が使うサービスにはなり得ないのかも...



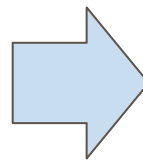
ソーシャルブックマークは熱いか？（3）

- ただ、認知度は大きく向上している
 - 「ああ、意見が集まっているところでしょう？」
 - 2ch と同じで、
書く人は少なくとも読む人は多い
- アカウントを持っていない人にもアピール
→ メディアとしての存在感は増大

認知されているのは...

■ ブックマーク一覧ページ

- すべてのWebコンテンツに対するコメントが集積する裏ページとして



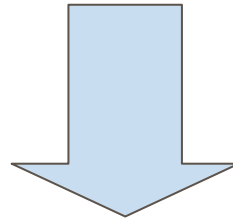
一般人(=非ユーザ)による利用

- 人気エントリーを見る際に先にコンテンツではなくコメントを確認する
- ニュースやブログを読んでなにかおかしいと感じたときにコメントを確認して安心する/がっかりする
- 自分のブログに対するコメントが気になって通勤中に確認をする



問題点

- ブックマーク数の多いコンテンツ
→ コメントが多すぎて読みきれない
- コメントの内容はなんでもあり
→ 読む意味のないコメントが多い
 - コンテンツを見た上でないとわからないコメントと
見る前でも参考になるコメントが混在



コメントのフィルタリングが必要

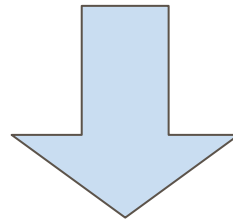


コメントのフィルタリングができれば

- もちろん時間の節約に
- 狭い場所への埋め込み
 - ブログの記事の下
 - 検索エンジンの結果ページ
 - コンテンツの内容を推測するのに役立つコメントだけを選択
- ブログ著者向けのコメント確認ツール
 - 反論や、炎上につながるコメントだけを提供

コメントのフィルタリング

- フィルタリングを実現するには？
→ コメントの分類が必要
- そもそもソーシャルブックマークのコメントとは？
 - 自分のためのメモ → 他人に対しての意見表明
(共有サービスから議論の場へと変容)



コメントの分類・調査



コメントの分類・調査(1)

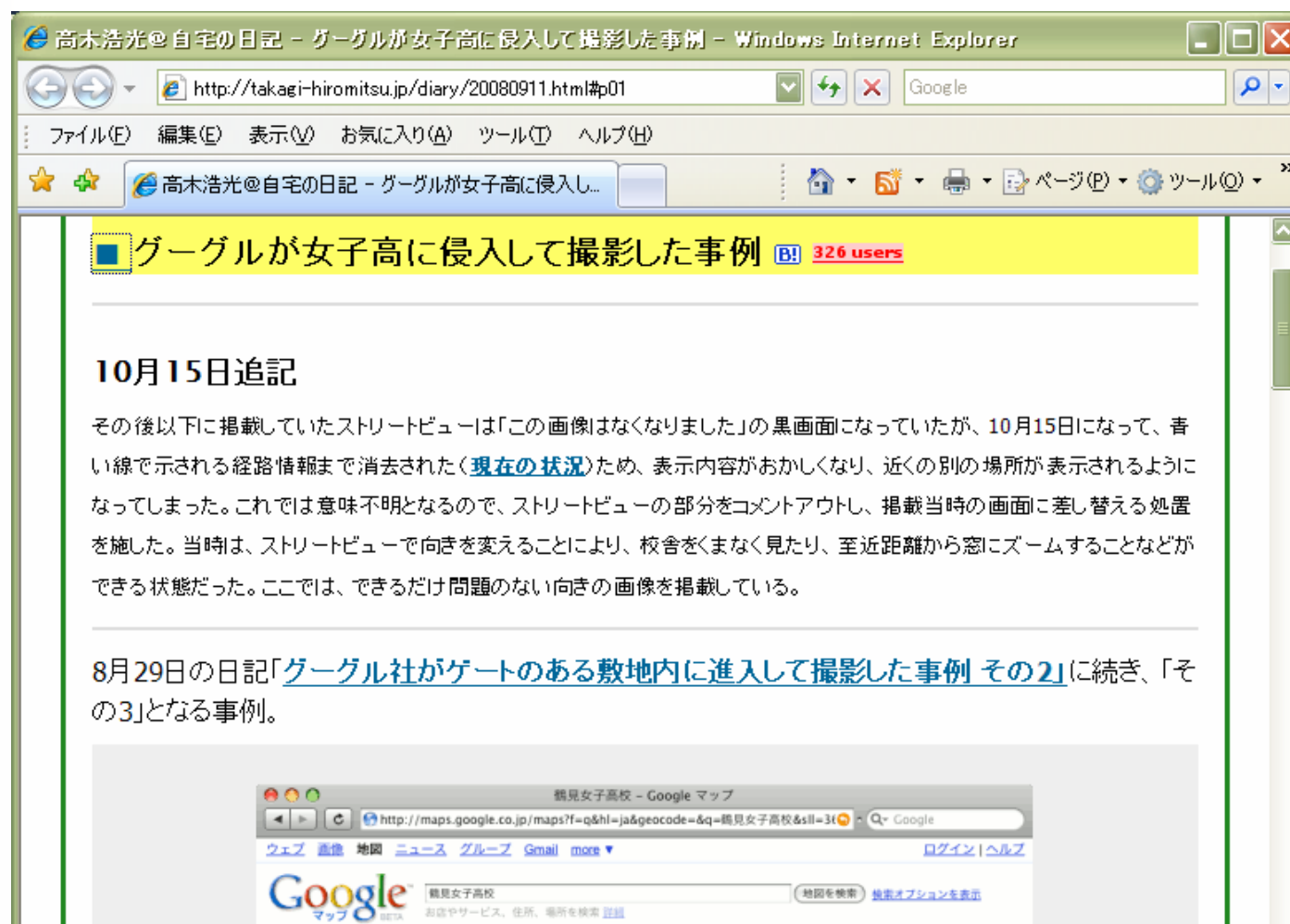
- ソーシャルブックマーク「本来の」コメント
 - タグでは表現できない情報を付与
メタ情報としての要約・引用が主体
- はてなブックマークのコメント
 - 他人の目を意識した意見・感想が主体？
 - 過去の del.icio.us の調査では意見・感想が15%程度
はてなブックマークは明らかに異なる性質



コメントの分類・調査(2)

- 要約・引用 (「本来の」使い方) VS. 意見・感想
- 意見・感想のうち、数の多い
 - ・評価、
 - ・疑問、
 - ・予想・希望を独立

例：対象コンテンツ





 [HiromitsuTakagi](#) ↓p-4 < 違うよ。ここが母校の人ならすぐ気付く。それが表に出てこないほどストリートビューは全然一告してください(グーグル)」では済まないってこと。2008/09/11  ★34★

 [ultraist](#) google, voyeur 一般人が女子高に侵入して写真を撮ると捕まりそうだけど。2008/09/11  ★19★

 [shout_poor](#) 爆弾とか..... 2008/09/11 

例：要約

と弄

 [shrkr](#) streetview 2008/09/11 

 [dhzk](#) google ちゃんとGoogleは説明すべきだね 2008/09/11  ★★★★★

 [tenkoma](#) google, map 2008/09/11 

 [tlfj](#) 計算機環境, 社会 2008/09/11 

 [pandaman385](#) google グーグルストーカーが、女子校に侵入。一步間違えるとストーカーになるという話。 2008/09/11 

 [taroleo](#) google, ストリートビュー これも「人のフィルターを通す前」の情報。この場合の「人」は、女子高(生)側も入る 200

 [zyugem](#) うーむ、ネガティブな事例を積み重ねてもグーグル問題はどうにもならない気がする。もっと本質を突いた問題

 [activecute](#) 反社会的企業 2008/09/11  ★

 [urahayate](#) 2008/09/11 

 [p-4](#) ネットサービス, 興味深い すごい！でもこれって前回事例発見からから1ヶ月以上女子高を探し続けてた人がいると考
上で自宅つくばで横浜市の事例を「通報した」って主体的に動く動機・・・ 2008/09/11  ★★★★★★

 [uxoru](#) 高木浩光, Google Maps 2008/09/11 

 [zu2](#) 高木先生すばらしすぎ / 一方米軍は <http://ryukyushimpo.jp/news/storyid-25588-storytopic-1.html> 2008/09/11

 [winny2b71](#) 校舎内に侵入といえば、米軍が養護学校内に装甲車で侵入して部活生のいる中グラウンドを走りUターン
<http://ryukyushimpo.jp/news/storyid-25588-storytopic-1.html> 2008/09/11  ★★★★★

 [retlet](#) googlestreetview 取り返しのつかない被害ってなんだろう 2008/09/11  ★★★★★★★★★★

 [anigaka](#) 裸の紐を持つ男のQRを思い出した 2008/09/11  ★★

例：引用

 [pochi-p](#) これはひどい, google, プライバシー 彼女達に謝まれ！ > Google 2008/09/12 

 [jbaucast](#) ... だんだん王様方向の運転手に乗らせたように思えてきた。 2008/09/12   

 [ocha39](#) google 2008/09/12 

 [m_yanagisawa](#) google, これはひどい 「再確認」の作業など行われているとは到底思えない」 2008/09/12 

 [Ubuntu](#) 2008/09/12 

 [cubed-l](#) セキュリティ, プライバシー 2008/09/12 

 [kurihara99](#) 2008/09/12 

 [jiwer5959](#) streetview, google 女子高というキーワードが関心を集める(かな)。「運が悪ければ女子生徒に取り返しのかにかかわらずに使えそう。(小さな出来事を過大に表現するのに便利) 2008/09/12 

 [yanok](#) Google, プライバシー 2008/09/12 

 [KANIBUCHI](#) google, これはひどい 秘密の花園画像満載を期待して開いた俺がバカだったよ・・・！！ 2008/09/12  

 [y_o_u_1](#) 2008/09/12 

 [nakag0711](#) 進んだらしい方向と1の状況を見る限り入ってしまったのは仕方ないが、2以降の部分は初めから削除して

 [t-murachi](#) Google, GoogleMap, プライバシー, コンプライアンス |ω|=3 2008/09/12 

 [chintaro3](#) いいぞもっとやれ 2008/09/12   

 [strairal](#) まともに対応してくれないからこそ、ネガティブな事例を積み重ねて対応をうながす、また(ネット)世論に影響を 2008/09/12    

 [hejihogu](#) ストリートビュー, セキュリティ, google 「運が悪ければ女子生徒に取り返しのかにかかわらずに使えそう。(小さな出来事を過大に表現するのに便利) 2008/09/12  

2008/09/12

treetview, google 女子高というキーワードが関心を集める(かな)。「運が悪ければ女子生徒に取り返しのない被害が出ていただろう」という
使えそう。(小さな出来事を過大に表現するのに便利) 2008/09/12

le, プライバシー 2008/09/12

google, これはひ

08/09/12

例:意見・感想

進んだらしい方向と1の状況を見る限り入ってしまったのは仕方ないが、2以降の部分は初めから削除しておくべきだったとは言える 2008/09/1

oogle, GoogleMap, プライバシー, コンプライアンス |ω°)=3 2008/09/12

いぞもっとやれ 2008/09/12 ★★★

もに対応してくれないからこそ、ネガティブな事例を積み重ねて対応をうながす、また(ネット)世論に影響を与えるようにするのもやむをえないの
★17★

レポートビュー, セキュリティ, google 「運が悪ければ女子生徒に取り返しのない被害が出ていただろう。」という言い回しには違和感を感じる。2

ogle 2008/09/12

oogleMaps 2008/09/12

見つけてくるなあ…。 2008/09/12 ★★

れだけ問題を指摘されてもやっぱりGoogleは無視し続けるんだよね、きっと。//撮り直すまで放置で、修正後にリリースを出して、それを褒める
見たような光景だけど。 2008/09/12 ★★★★★★

全にアンチになってしまわれた 2008/09/11 ★

gleはevil ↓↑この期に及んでまだこんなメタ議論or高木氏に非を求める行為を仕掛ける連中がいることが最大の問題。「たまたま問題が無かった
」と書いてある内容を全く読み取れていない。 17 2008/09/11 ★23★

ile 公開前に人によるチェックはしてないんだろなというのは問題。で、女子校に侵入可能なことを世界に知らしめているのは高木氏。 2008/0

ikani In-4 < 違うよ、ここが母校の人ならすぐ気付く、それが表に出てこないほどストリートビューは全然一般の人に知られてないってこと。「開

例：評価

 [A410](#) *あとで読む 2008/09/12 ☆+

 [dbfireball](#) google これは普通にやばい。 2008/09/12 ☆+

 [ocha39](#) google 2008/09/12 ☆+

 [m_yanagisawa](#) google, これはひどい「再確認」の作業など行われているとは到底思えない」 2008/09/12 ☆+

 [Ubuntu](#) 2008/09/12 ☆+

 [cubed-I](#) セキュリティ, プライバシー 2008/09/12 ☆+

 [kurihara99](#) 2008/09/12 ☆+

 [jiwer5959](#) streetview, google 女子高というキーワードが関心を集める(かな)。「運が悪ければ女子生徒に取り返しのかにかかわらずに使える。 (小さな出来事を過大に表現するのに便利) 2008/09/12 ☆+

 [yanok](#) Google, プライバシー 2008/09/12 ☆+

 [KANIBUCHI](#) google, これはひどい 秘密の花園画像満載を期待して開いた俺がバカだったよ・・・！！ 2008/09/12 ☆+ ★2

 [y_o_u_1](#) 2008/09/12 ☆+

 [nakag0711](#) 進んだらしい方向と1の状況を見る限り入ってしまったのは仕方ないが、2以降の部分は初めから削除して

 [t-murachi](#) Google, GoogleMap, プライバシー, コンプライアンス (w^)=3 2008/09/12 ☆+

 [chintaro3](#) いいぞもっとやれ 2008/09/12 ☆+ ★★

 [strairai](#) まともに対応してくれないからこそ、ネガティブな事例を積み重ねて対応をうながす、また(ネット)世論に影響を
2008/09/12 ☆+ ★17



調査対象

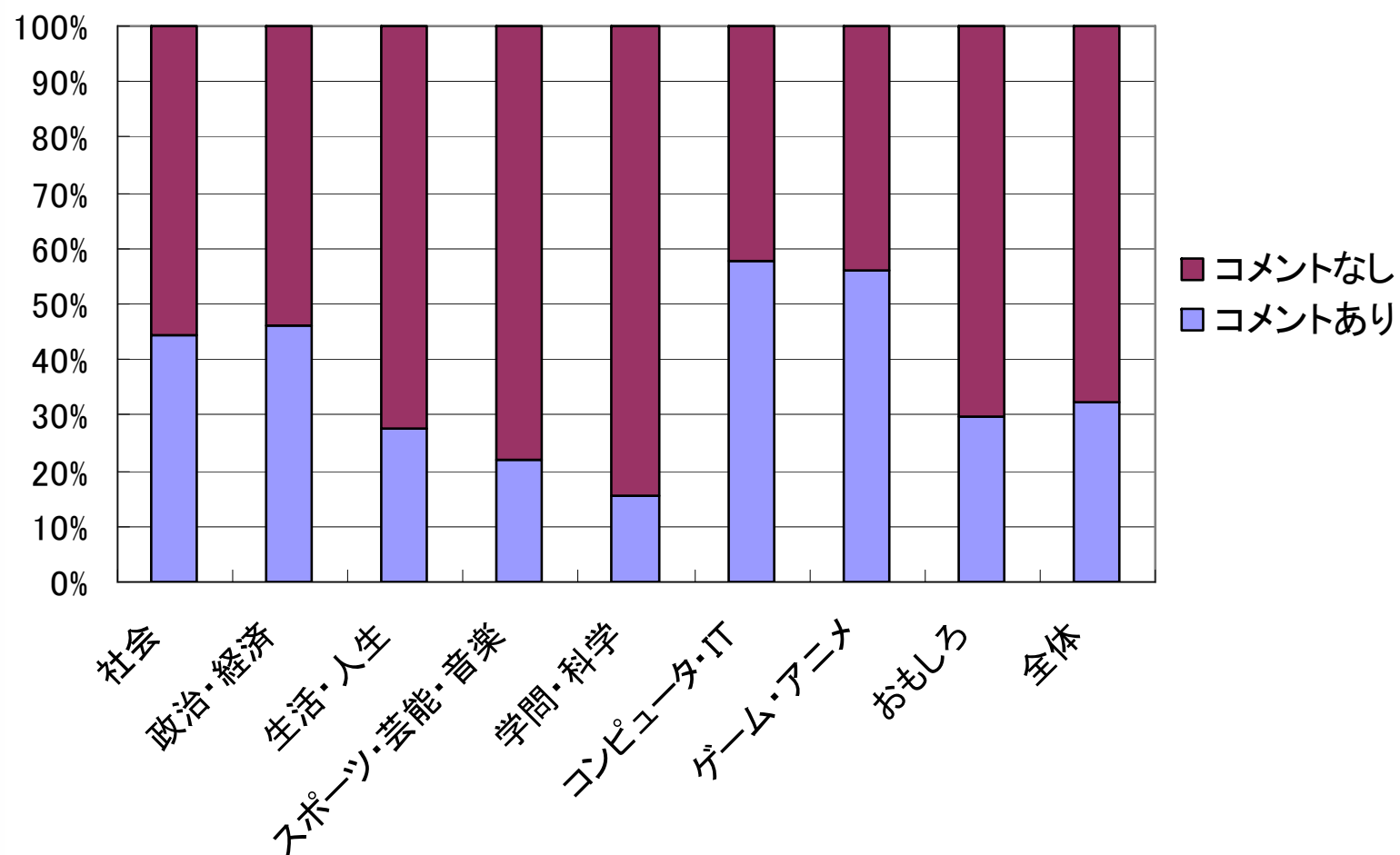
- 「はてなブックマーク」[†]の人気エントリーでブックマーク数が100以上のもの
- 調査期間 2009/1/9~2009/1/14
- 8つのカテゴリー[‡]より、24コンテンツを選択
- ブックマーク数 4285 、 コメント数 1396

[†]はてなブックマーク <http://b.hatena.ne.jp/>

[‡]8つのカテゴリー

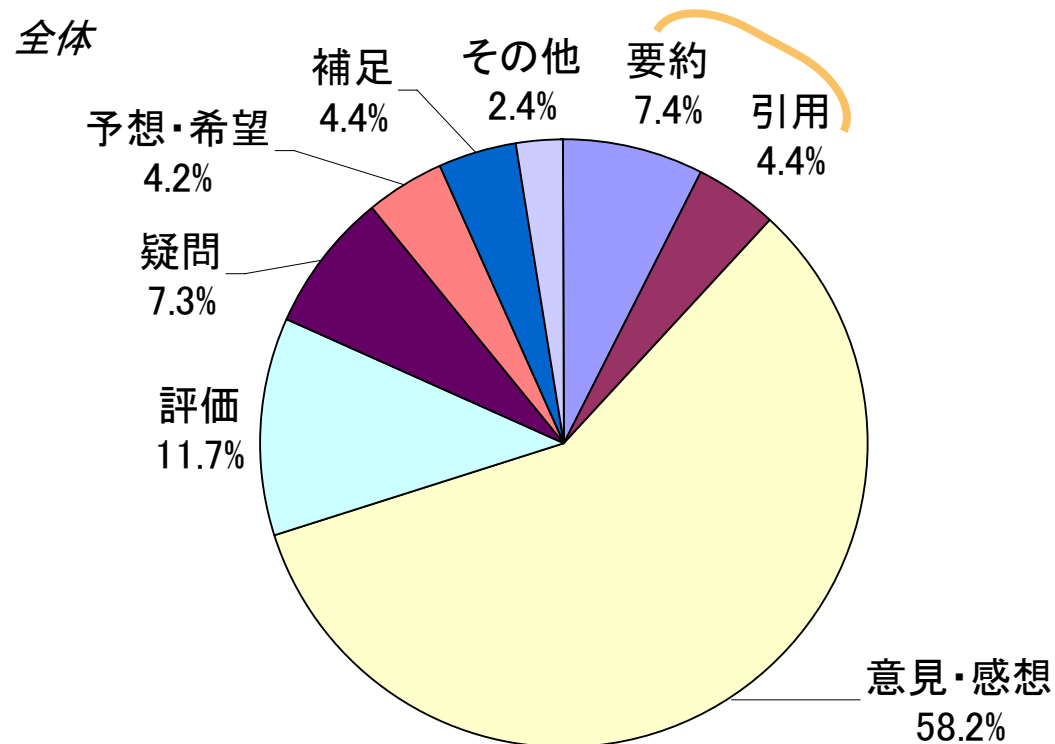
「社会」「政治・経済」「生活・人生」「スポーツ・芸能・音楽」
「科学・学問」「コンピュータ・IT」「ゲーム・アニメ」「おもしろ」

調査1: コメント記入率

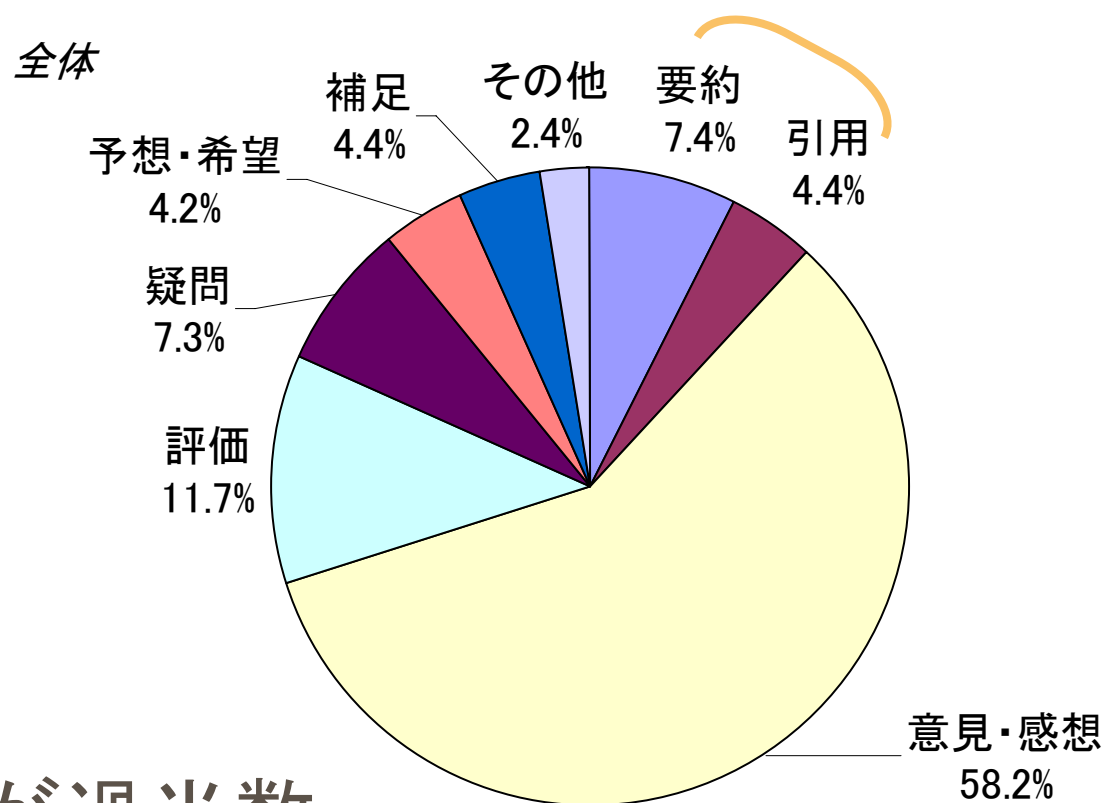


調査2: コメント分類（全体）

- 要約・引用（本来の使い方） vs. それ以外
- 意見・感想のうち、
 - ・評価、
 - ・疑問、
 - ・予想・希望を独立



調査2: コメント分類（全体）

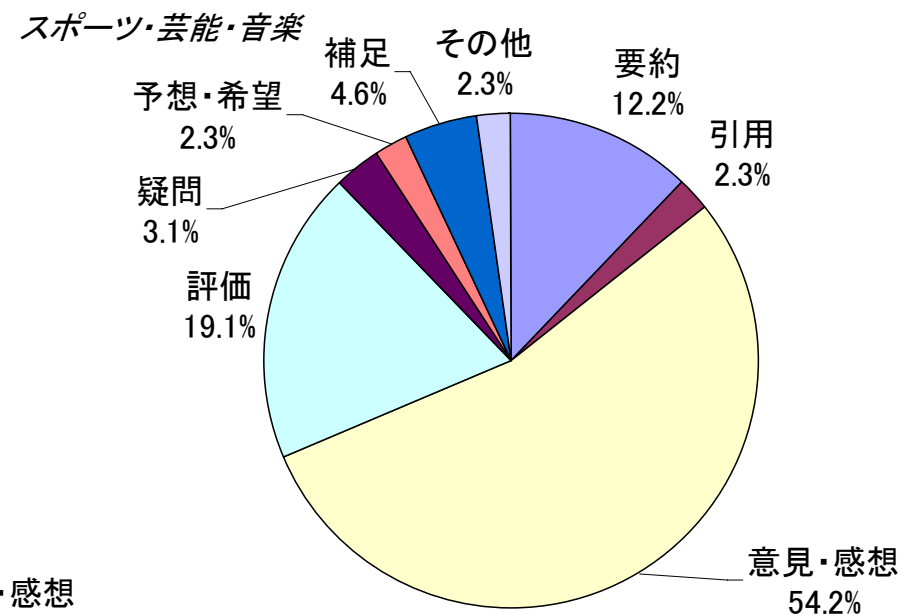
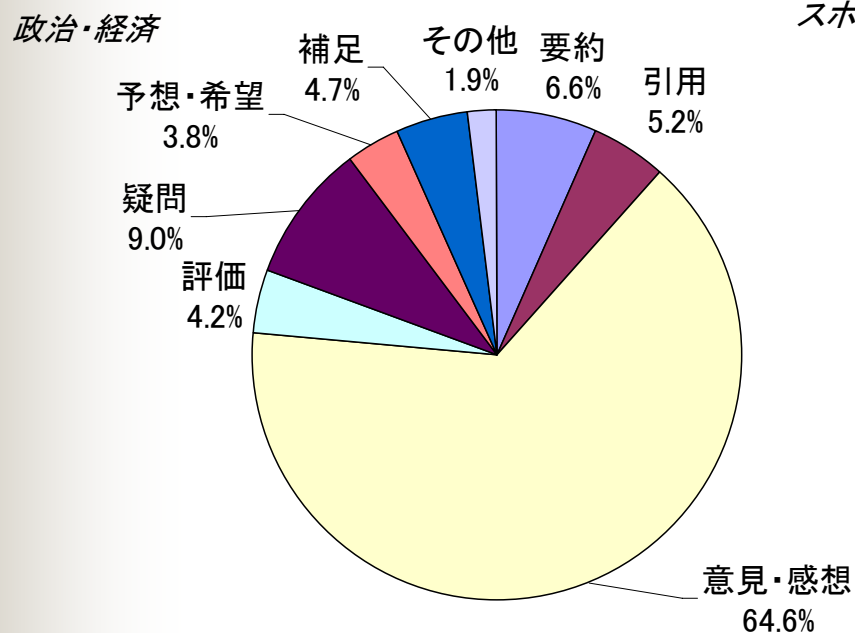


■ 意見・感想が過半数

← 意見表明の場として機能している

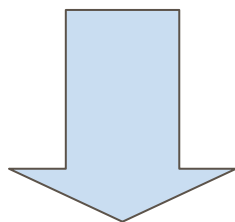
調査2: コメント分類（カテゴリー別）

「意見・感想」が最も多いのは「政治・経済」
最も少ないのは「スポーツ・芸能・音楽」



調査のまとめ

- 意見・感想が過半数
- 意見・感想が多いのは分野によらない



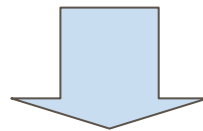
ブックマーク一覧

= コメントによる意見表明の場

フィルタリングに向けて

■ 問題点(再掲)

- コンテンツを見た上でないとわからないコメントと見る前でも参考になるコメントが混在



- コンテンツをまだ見ていない人
「要約」「引用」
- コンテンツをすでに見た人
「疑問」「予想・希望」「補足」
- コンテンツの既読/未読を問わない任意の人
「意見・感想」「評価」



言語表現の分析による分類

→ コンテンツをまだ見ていない人

「要約」「引用」

→ 引用表現、対象コンテンツの内容との比較

→ コンテンツをすでに見た人

「疑問」「予想・希望」

→ 主文述部の文末表現の特徴を利用

「補足」

→ URLやIDを検出

(これらを含まないものは意見・感想に含める)

→ コンテンツの既読/未読を問わない任意の人

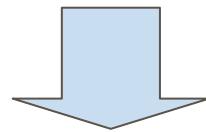
「意見・感想」「評価」

→ 上記以外

(評価を分離する場合には評価表現を抽出)

「意見・感想」vs.「要約」「引用」

- コメントの言語表現だけを見ても
「意見・感想」かは判断できない
 - 引用元・要約元の記事の内容が
「意見・感想」に分類される表現を持つかも

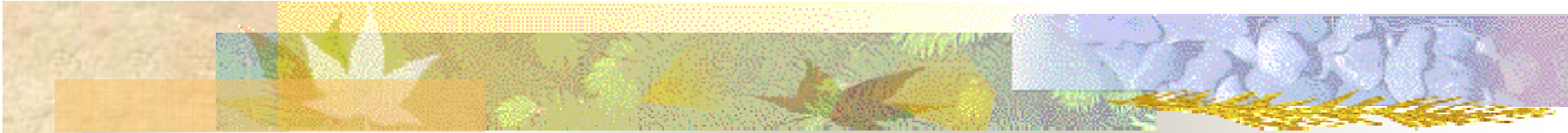


コンテンツを取得し、
「要約」「引用」かを先に判断



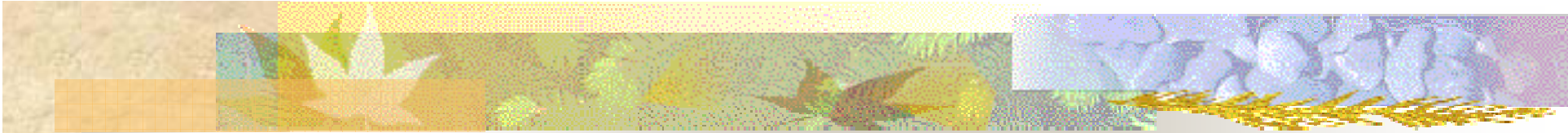
「要約」「引用」の判定

- コンテンツを取得しその内容と比較
- コメント・コンテンツの双方を形態素解析
 - 一語、二語のコメントは除外
- 「引用」: 形態素列(用言も含む)の完全一致でほぼok
- 「要約」: コメントの単語集合に含まれるコンテンツ内の単語の割合(7割で実験)



「意見・感想」の細分類

- 言語表現の特徴による分類
- 「疑問」「予想・希望」は主に文末表現
 - 形態素解析はしている
 - 構文解析をしているがいまのところ不要
 - 評価表現をきちんと扱う場合には
評価対象の特定に必要
 - ほぼとれている



まとめ

- ブックマーク一覧ページのコメントは意見表明のための新しいメディア
- コメントのフィルタリングがあるとよい
 - 対象コンテンツの既読/未読による分類
 - 言語表現による自動分類を検討
 - まずは「要約」「引用」vs.「意見・感想」を実験中

ご清聴ありがとうございました。