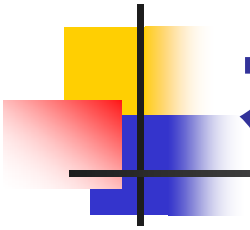
ブロードバンド時代における情報フィルタリングの動向

大阪大学工学研究科情報システム工学専攻

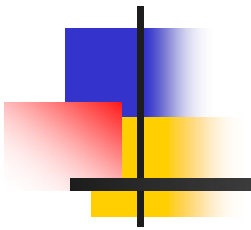
澤井 里枝



本日の内容

- これまでのブロードバンドを用いた情報フィルタリングサービス
- 情報フィルタリングとは
- 情報フィルタリングの研究動向
- これからのブロードバンド時代における情報フィルタリング

これまでのブロードバンドを用いた情報フィルタリングサービス



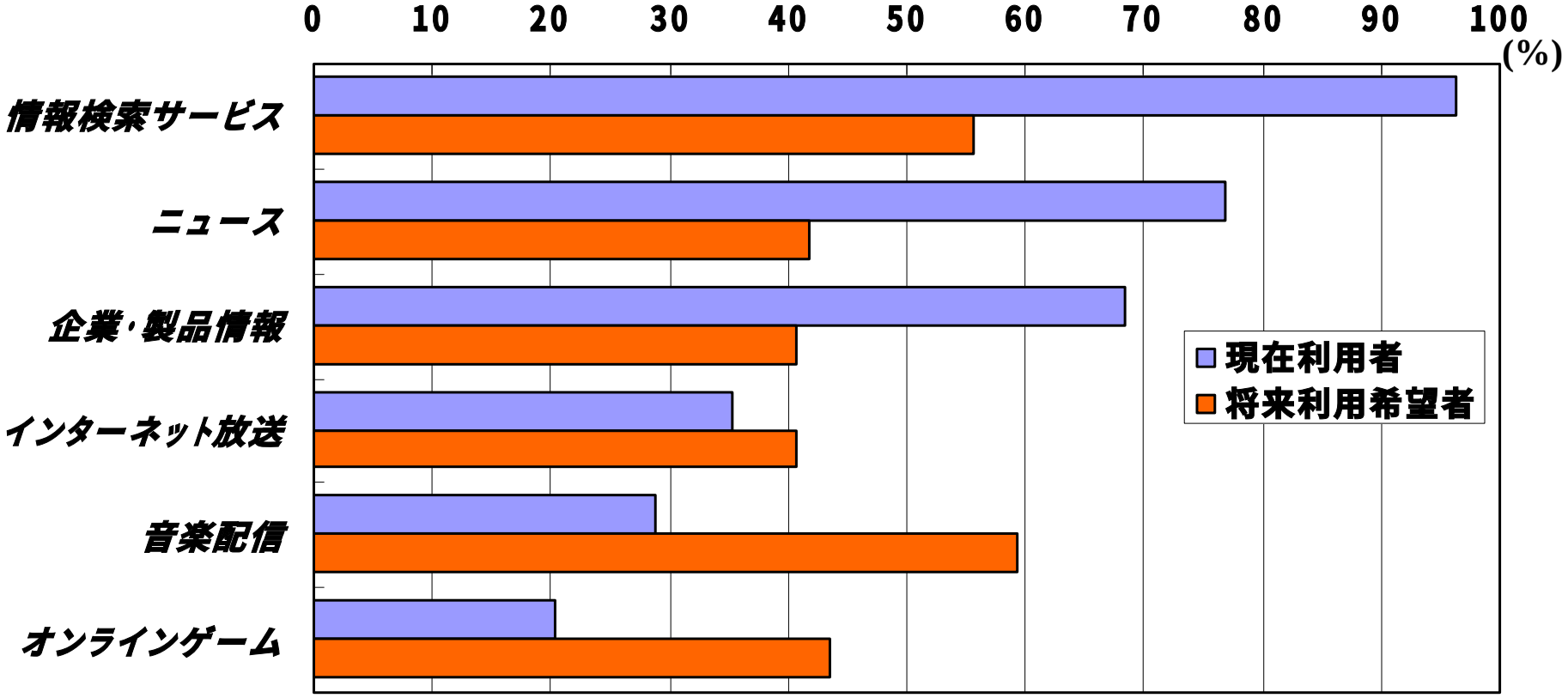
背景

- **ブロードバンドネットワークの普及**
  - ADSL
  - **衛星放送**
  - CATV
  - FTTH

さまざまなコンテンツが利用可能

コンテンツの変化

- 検索やニュースからエンターテイメントへ
  - ブロードバンド利用者が希望するコンテンツ



(出典：ブロードバンド利用動向調

ブロードバンドを用いたサービス

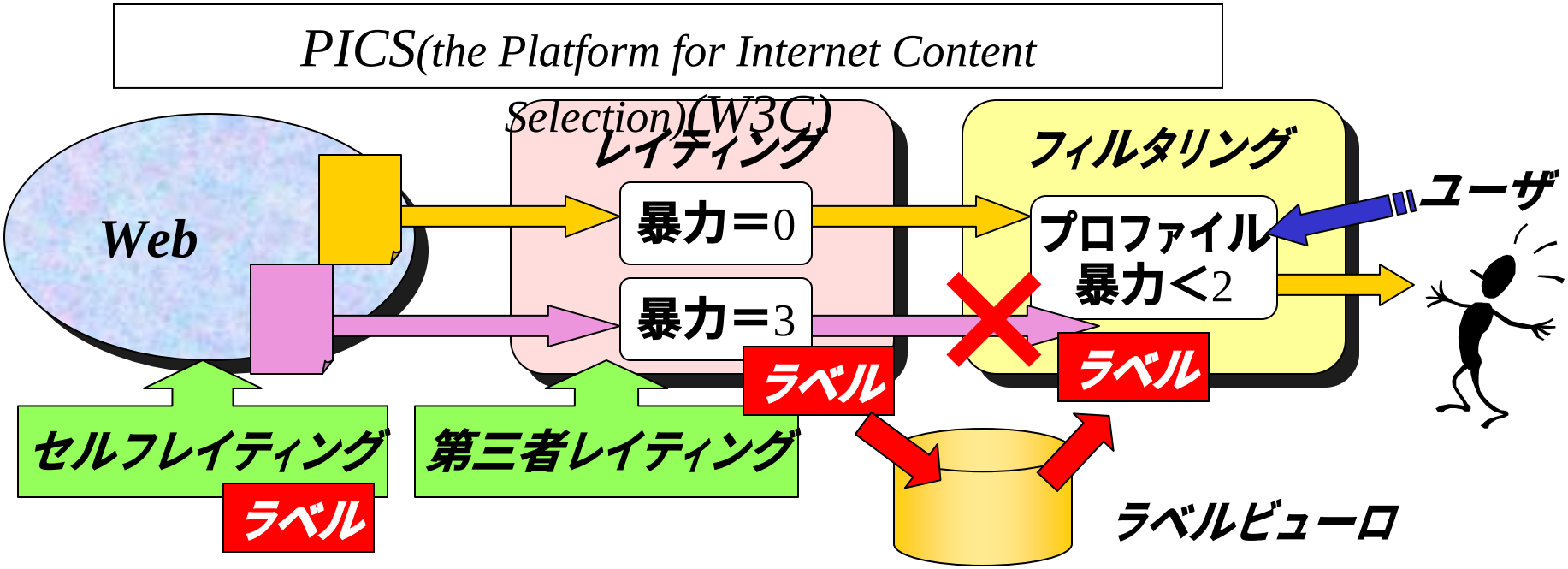
- 音楽配信
  - ソニーミュージックエンターテイメント
  - エイベックスネットワーク
  - ポニーキャニオン
  - キングレコード
- 映像配信
  - impress TV
  - Net-TV
  - itv24.com
  - Comin’ soon TV

現在運用中のフィルタリング

- フィルタリングを用いたサービス
  - 有害情報フィルタリング
    - AD-Guard(アイキュエス，電通テック)
    - Webフィルタリング(DreamNet)
  - Webデータのフィルタリング
    - SkyAgent(SkyCom)
  - 音楽データのフィルタリング
    - Napster(Napster)
    - Rhapsody(Listen.com)
  - 映像･番組のフィルタリング
    - アソシアガイド(NTT)

有害情報フィルタリング

- 学校や企業，官公庁などで有害情報をブロックする．
  - セルフレイティング・第三者レイティングによって，ページの有害度を格付けする．
  - プロファイルよりも有害度が小さい情報を提示する．



有害情報フィルタリング

- **レイティング基準の標準化**(1996**年以降**)
  - RSACi, ICRA, SafetyOnline, SafeSurf, NetShepherd…
- **フィルタリングソフト**
  - AdGuard(**アイキュエス, 電通テック**),
    Cyber Patrol(SurfControl,
    The Learning company),
    CYBERsitter(Solid Oak Software),
    Bess(N2H2)
- Web**フィルタリング**(DreamNet)
  - **アドレスブロック, セルフレーティング, 語句･単語フィルタの3重方式.**

RSACi レイティング基準

レベル	暴力
4	残虐な暴力
3	攻撃的な暴力, または人間の死
2	実在する物の破壊
1	人間の傷害
0	上記以外, または スポーツ関連

SkyAgent(SkyCom)

- One-to-One型インターネット情報配信システム
  - コンテンツとクライアントの個別属性データを利用.
  - コンテンツリストをクライアントに提示.
  - 音声合成による応答やサイバーキャラクタエージェントなどによるユーザインタフェース.

http://www.skyagent.ne.jp/

Rhapsody(Listen.com)

- 音楽配信サービス・インターネットラジオ
  - キャッシュサイズが制限.
  - アーティスト，アルバム，作曲家などで曲を検索.
  - ヒット曲リストを提供.
  - 登録したアーティストから，関連のある曲を推薦.

http://www.listen.com/

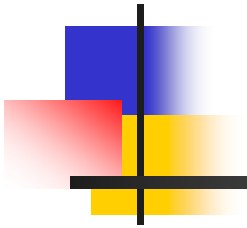
アソシアガイド(NTT)

- 番組ナビゲーションサービス
  - コンテンツに応じたマップ型インタフェース.
  - コンテンツのフィルタリング機能.
    - 新着モード
    - 視聴率に基づく
      ランキングモード
  - 番組表
  - ジャンル別番組表示
  - 広告表示

http://www.ntt.co.jp/news/news01/0112/011225.html

これまでのフィルタリングサービス

- まだ初歩的なことしかできない.
  - 嗜好の表現法が少ない.
    - ジャンルの選択
    - アーティストの選択
    - 年齢, 職業などのプロフィール登録
  - 情報の解析が不十分.
    - 単純なキーワードによるマッチング.
    - 手作業による情報の分類.



情報フィルタリングとは

情報フィルタリングとは

- **必要な情報をユーザに提示する手法**

  [Belkin92] *(Rutgers大, Comm. of the ACM)*.
- **広域な環境で情報を見つけるための強力な手法**

  [Yan00] *(Healtheon, TODS)*.
- **数多くの情報の中から必要な情報を選択するためのコストと，情報を得たことによるメリットの間のギャップを埋めるための技術**[森田93] *(北陸先端科技大)*.

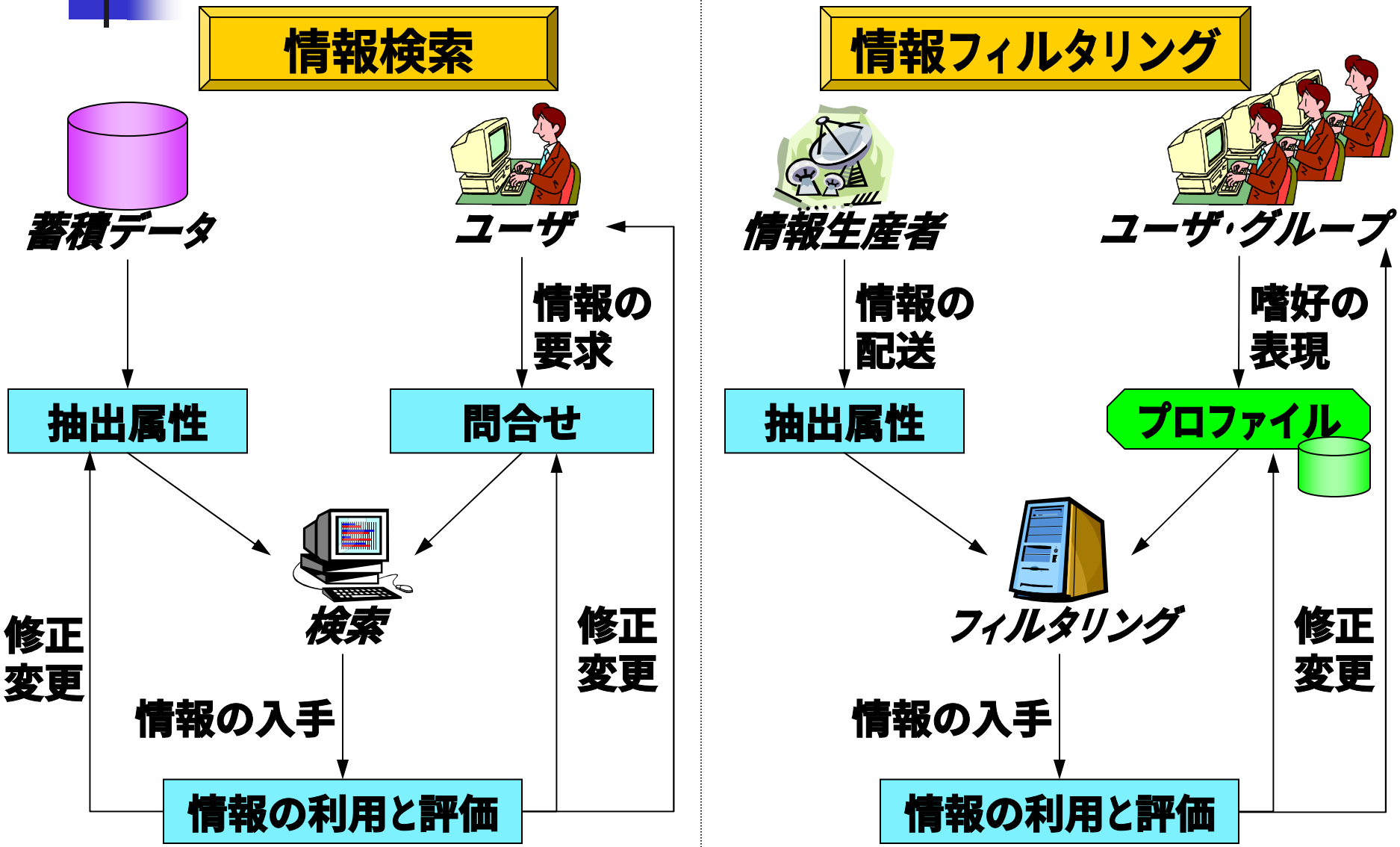
情報検索と同じ？

情報検索との違い

- **継続的な要求に対して情報を提供する**
  [Belkin92] *(Rutgers大, Comm. of the ACM)*,
  [Bell96] *(Melbourne大, SIGIR)*.
- **要求が動的に変化する**[Belkin92].
- **データソースが動的に変化する**[Belkin92].
- **複数ユーザの要求を満たす**[Belkin92].
- **プライバシー問題の解決が必要**[Belkin92].
- **情報提示のタイミングが重要**[Bell96].

時間軸が重要？
要望によって異なる？

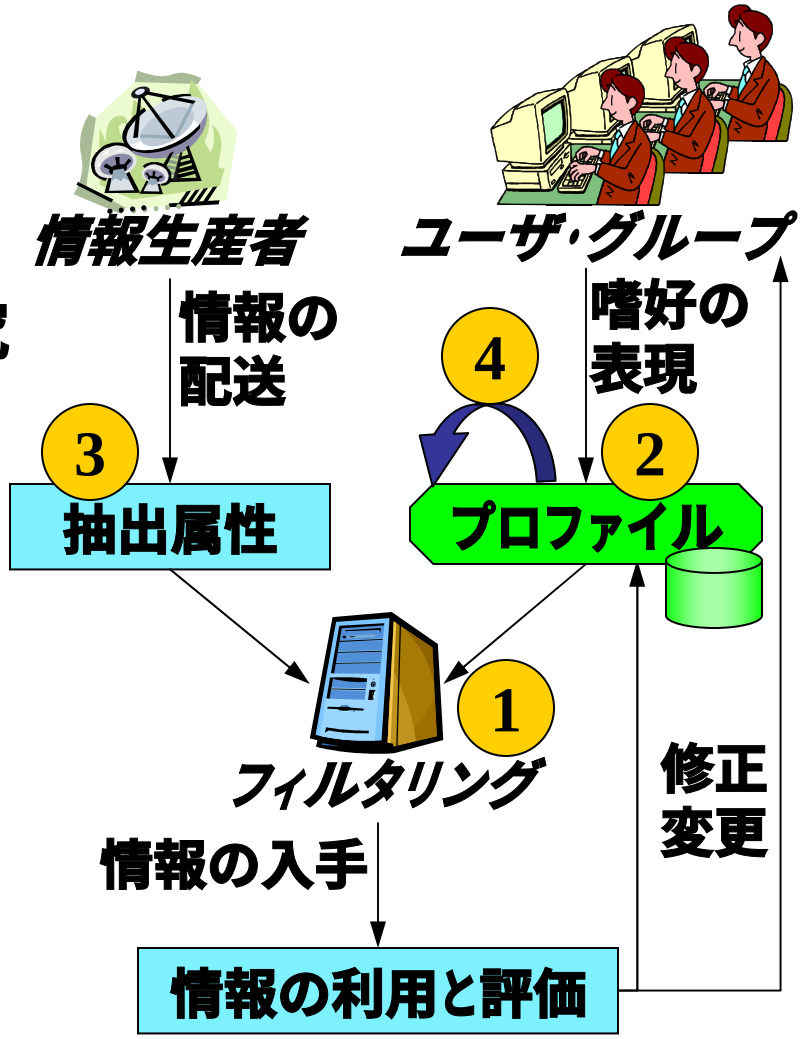
一般的なモデル[Belkin92](Rutgers大)



情報フィルタリングの研究動向

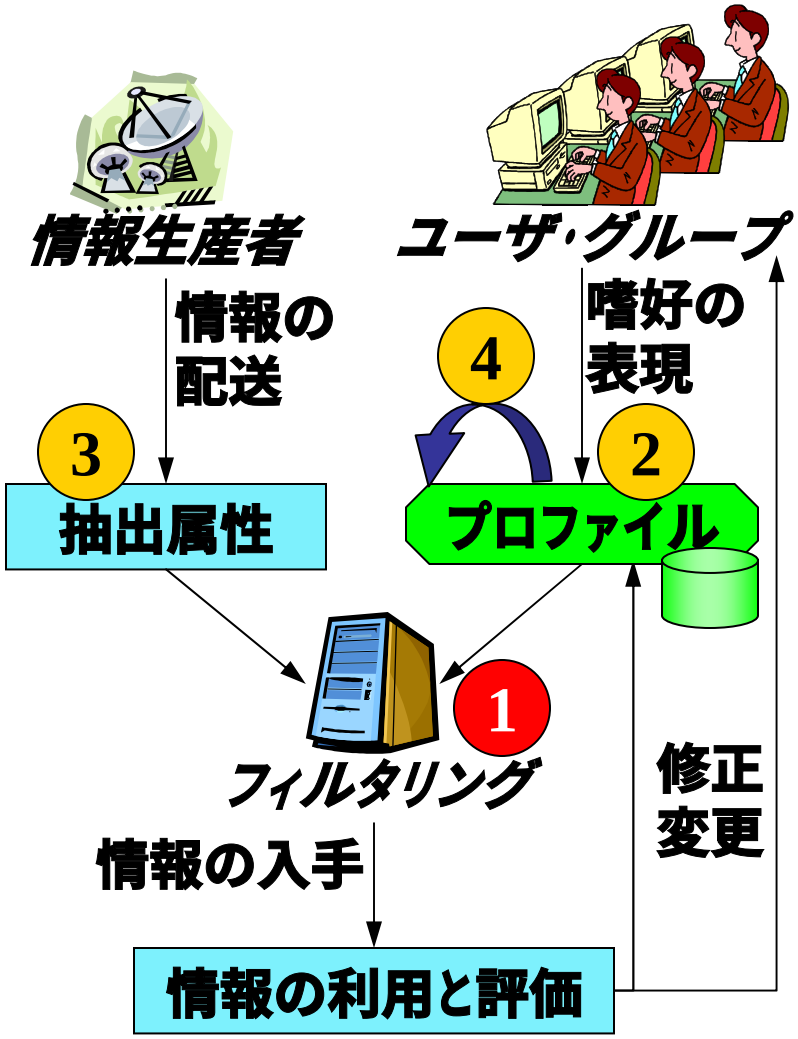
これまでのフィルタリング研究

1. データ抽出法に関する研究
   - セレクション
   - ランキング
2. プロファイル作成法に関する研究
   - 直接的方法
   - 半直接的方法
   - 間接的方法
3. データの分類法に関する研究
   - グルーピング
   - 意味ベクトル
4. 協調フィルタリングに関する研究



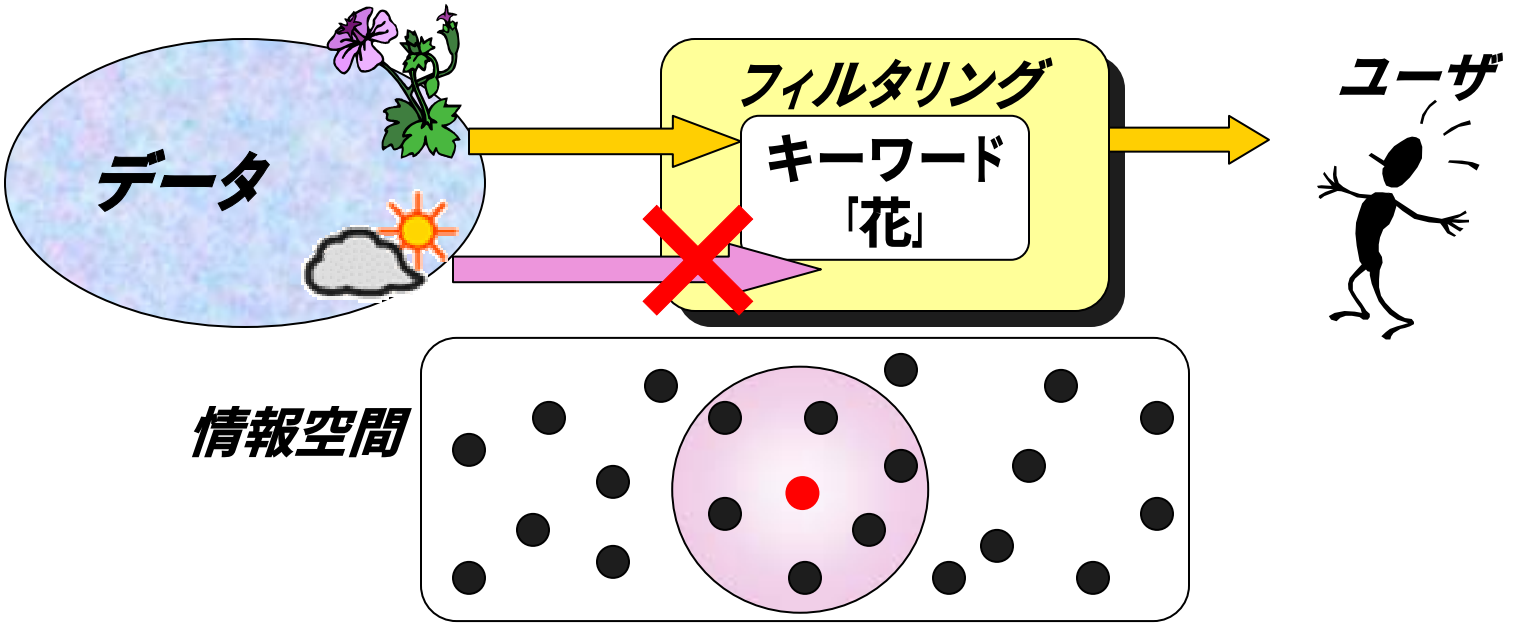
1. データ抽出法に関する研究

- セレクション
  - 条件を完全に満たすデータを抽出する手法.
- ランキング
  - 条件に合う度合いを評価値で表し, その上位から特定数のデータを抽出する手法.



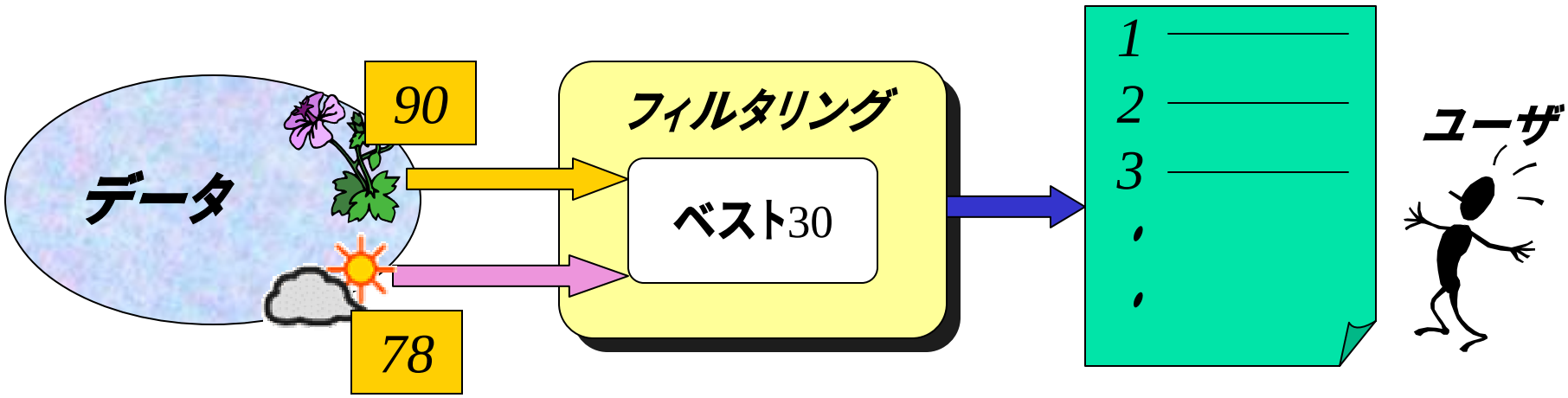
セレクション

- **条件を完全に満たすデータを選択する手法.**
  - INFOSCOPE*[Fischer91](Colorado大)*, Lyric-Time*[Loeb92](Bellcore, Comm. of the ACM)*, Tapestry*[Goldberg92](Xerox, Comm. of the ACM)*, GroupLens*[Konstan97](Minnesota大, Comm. of the ACM)*, XFilter*[Altinel00](Maryland大, VLDB)*, NiagaraCQ*[Chen00](Wisconsin-Madison大, SIGMOD)*, SIFT*[Yan00](Healtheon, TODS)*, FBDA*[Kan01](RICOH)*



ランキング

- **条件に合う度合いを評価値で表し，その上位から特定数のデータを提示する手法.**
  - Syskil&Webert *[Pazzani96](California大)*, ProfBuilder *[Wasfi99](Science大, Malaysia)*

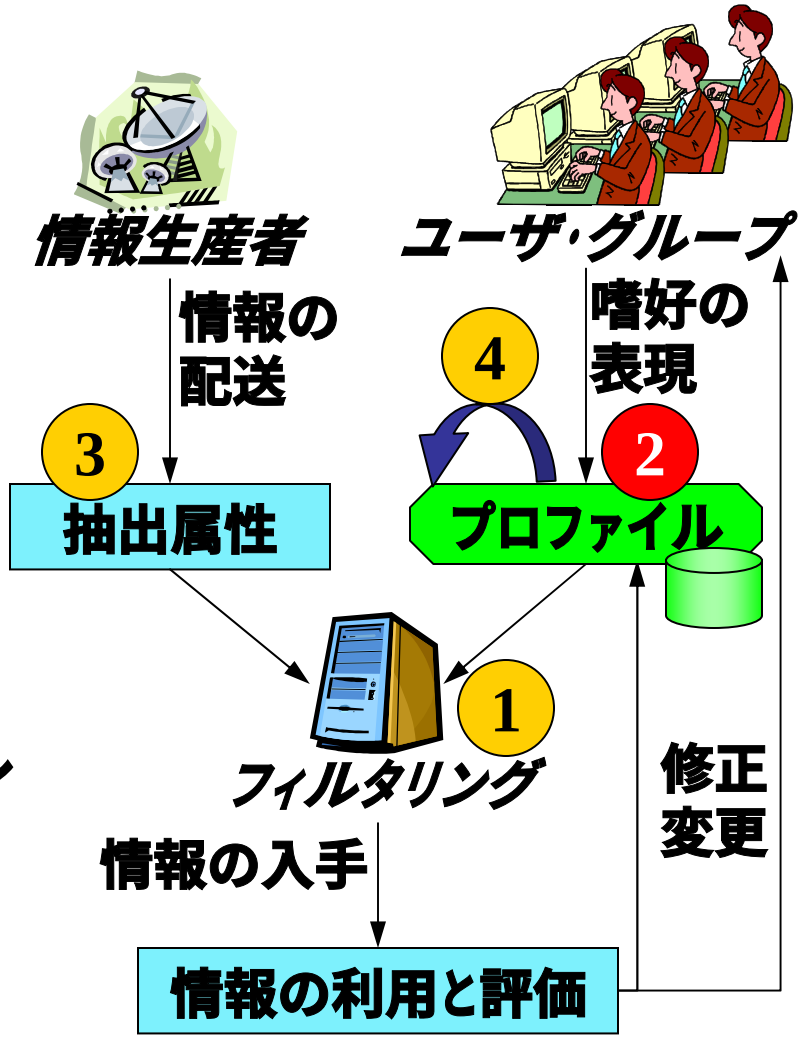


1. データ抽出法に関する研究(まとめ)

- セレクション
  - 計算コストの見積もりが容易.
  - フィルタリング対象によって, 結果のデータ量が大きく異なる.
- ランキング
  - ディスクにためるデータ数を制限できる.
  - 欲しい順番にデータを利用できる.
  - 評価値の計算方法でフィルタリングの精度が変わる.

2.プロファイル作成に関する研究

- 直接的方法
  - ユーザが自分でプロファイルを作成.
- 半直接的方法
  - ユーザとシステムが協調してプロファイルを作成.
- 間接的方法
  - システムだけでプロファイルを作成.



直接的方法

- **ユーザが明示的にプロファイルを記述.**
  - DATACYCLE*[Bowen92](Bellcore, Comm. of the ACM)*, Tapestry *[Goldberg92](Xerox, Comm. of the ACM)*, SIFT *[Yan00](Healtheon, TODS)*, NiagaraCQ *[Chen00](Wisconsin-Madison大, SIGMOD)*, [Fabret01] *(INRIA, SIGMOD)*

NiagaraCQのトリガ

CREATE *CQ_name*

XML-QL query

DO *action*

{**START** *start_time*} {**EVERY** *time_interval*}

{**EXPIRE** *expiration_time*}

time_interval ごとに問合せ *XML-QL query* を行う.

直接的方法

- **ユーザが明示的にプロファイルを記述.**
  - DATACYCLE*[Bowen92](Bellcore, Comm. of the ACM)*, Tapestry *[Goldberg92](Xerox, Comm. of the ACM)*, SIFT *[Yan00](Healtheon, TODS)*, NiagaraCQ *[Chen00](Wisconsin-Madison大, SIGMOD)*, [Fabret01] *(INRIA, SIGMOD)*

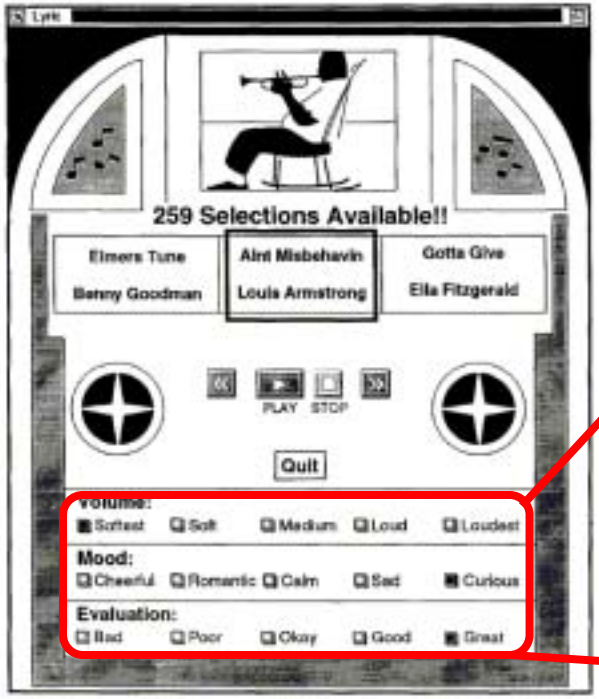
TapestryのTQL

```
m.to = 'BugReports'
AND  m.ts + [2 weeks] < now()
AND  NOT  EXSTS (mreply:
  mreply.in_reply_to = {m} )
```

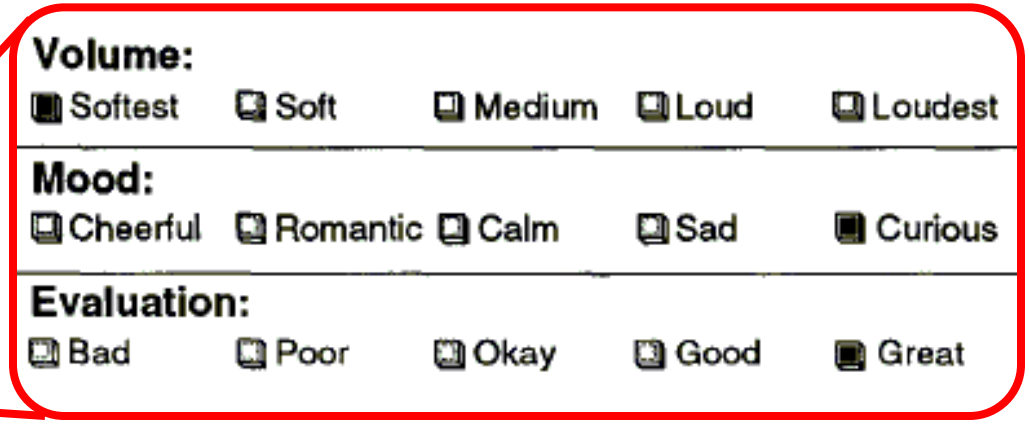
ここ2週間の間に受取り，まだ返事をしていない *BugReports* **宛のメールを抽出する.**

半直接的方法

- **ユーザがデータの評価値を与える(関連フィードバック).**
  - [Foltz92] *(Colorado大, Comm. of the ACM),* Syskill&Webert *[Pazzani96](California大)*, InfoFinder *[Krulwich97](AgentSoft)*, Ringo *[Shardanand95](MIT)*, Fab *[Balabanovic97](Stanford大)*



LyricTime *[Loeb92]*
(Bellcore, Comm. of the ACM)



間接的方法

- **システムがユーザの行動を監視することで，自動的に嗜好を検出する．**
  - Letizia*[Lieberman95](MIT)*, ProfBuilder*[Wasfi99](Science大, Malaysia)*, AIS*[Sanguantrakul99](Osaka大)*

AIS

- **ユーザがアクセスしたデータの評価値wを上げる．**

$$w \leftarrow (n \times w + 1) / (n + 1)$$

$$n \leftarrow n + 1$$

n: ***評価値が既に計算されているデータ数***

2.プロファイル作成に関する研究(まとめ)

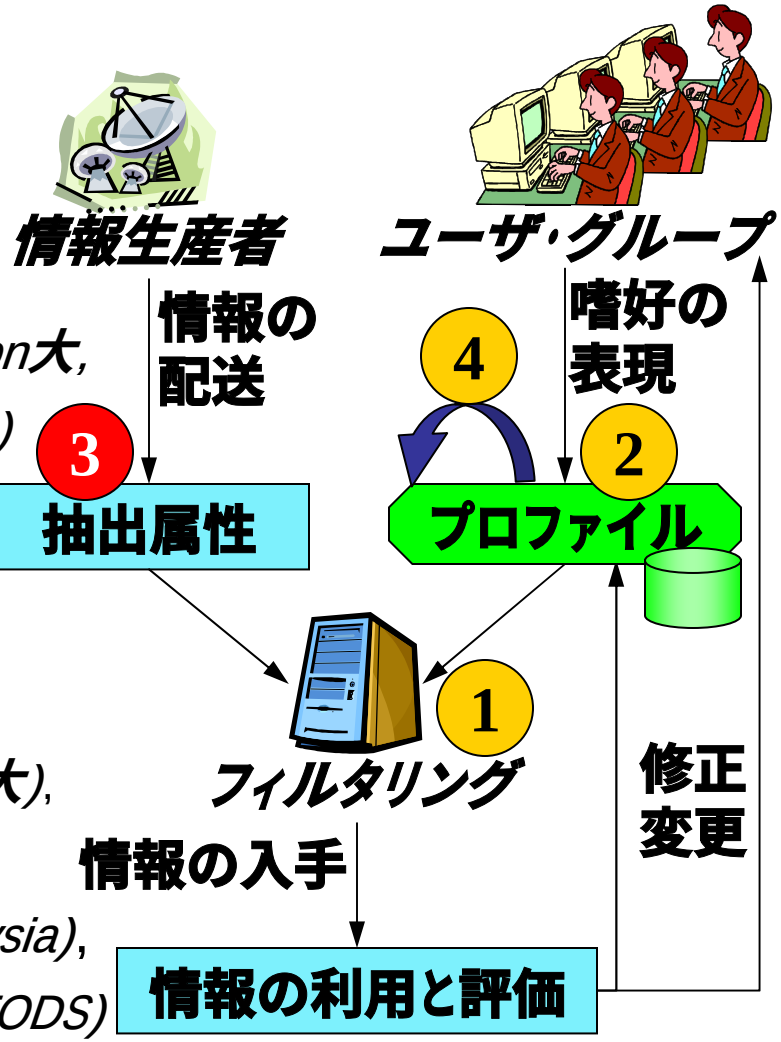
- 直接的方法
  - ユーザが直接記述するのはたいへん.
  - ユーザの主観によって記述方法が異なる.
- 半直接的方法
  - データを見るたびに評価を付けるのは面倒.
- 間接的方法
  - 関係のない情報が紛れ込む可能性が高い.
  - ユーザの嗜好を学習するまで時間がかかる.

その他のプロファイル作成法

- **直接･半直接的方法の組合せ**
  - **ユーザが直接キーワードを登録してから，関連フィードバックによってシステムがプロファイルを更新する．**
    - WebWatcher *[Armstrong95](Carnegie Mellon大)*, SiteHelper *[Ngu97](Monash大)*
  - **より的確なプロファイルが作成できる．**
- **半直接･間接的方法の組合せ**
  - **関連フィードバックとユーザの行動からプロファイルを更新する．**
    - Anatagonomy *[Kamba97](NEC)*
  - **ユーザの操作を軽減できる．**

3. データの分類法に関する研究

- グルーピング
  - ユーザの問合せを分類することで，データを抽出する．
    - NiagaraCQ *[Chen00](Wisconsin-Madison大, SIGMOD)*, [Fabret01] *(INRIA, SIGMOD)*
- 意味ベクトル
  - tf.idfを用いた手法
    - Lyric-Time *[Loeb92](Bellcore)*, Syskill&Webert *[Pazzani96](California大)*, SIFTER *[Mostafa97](Indiana大)*, ProfBuilder *[Wasfi99](Science大, Malaysia)*, SIFT *[Yan00](Healtheon Corporation, TODS)*



tf.idf

- TF(Term Frequency)
  - tf_{ij} : **データd_i中のキーワードk_jの出現頻度**
- IDF(Inverse Document Frequency)
  - idf_j : 1 / **全データ中k_jを含むデータの割合**
    データd_iでのキーワードk_jの重み

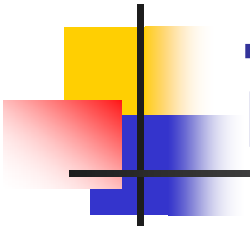
$$w_{ij} = tf_{ij} * idf_j$$

- **データd_iのベクトルD_iとユーザのプロファイルQ_jの類似度**

$$D_i = <(k_{i1}, w_{i1}),\ldots,(k_{in}, w_{in})>,\ Q_j = <(k_{j1}, z_{j1}),\ldots,(k_{jn}, z_{in})>$$

$$Similarity(D_i, Q_j) = \sum_k w_{ik} * z_{jk}$$

キーワードが増えるごとに計算量，メモリ使用量が増大する．



高度なフィルタリング手法

- 計算コストを軽減するための手法
  - ベクトル空間を小さくするための手法
    - 複数の手法を組合せた手法[Bell96] *(Melbourne大, SIGIR)*
  - インデックスによる効率化
    - SIFT *[Yan00](Healtheon Corporation, TODS)*
      - ドキュメントのみ扱う.
    - XFilter *[Altinel00](Maryland大, VLDB)*
      - XML文書のみ扱う.

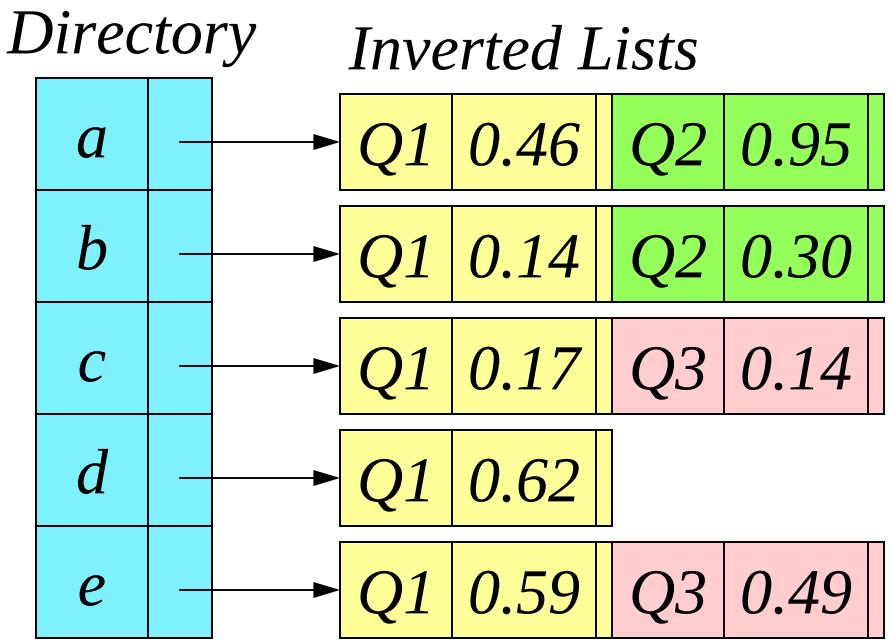
SIFT

- BF(Brute Force)手法
  - データごとに全ての問合せを調べるため，コストが高い．
- QI(Query Indexing)手法

D = <(a, 0.17), (b, 0.15), (d, 0.32)>

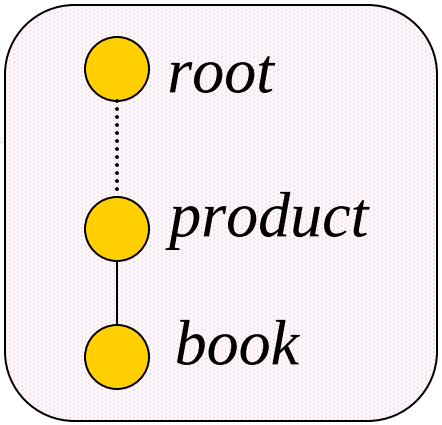
スコア
- $Q1$=0.29
- $Q2$=0.20
- $Q3$=0

閾値
- $Q1$=0.25
- $Q2$=0.20
- $Q3$=0.25

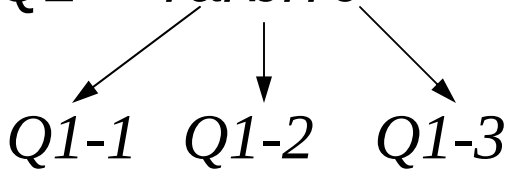


XFilter

- XMLの構造を利用したフィルタリング.
- XPath　　//product [price/value<300]/book



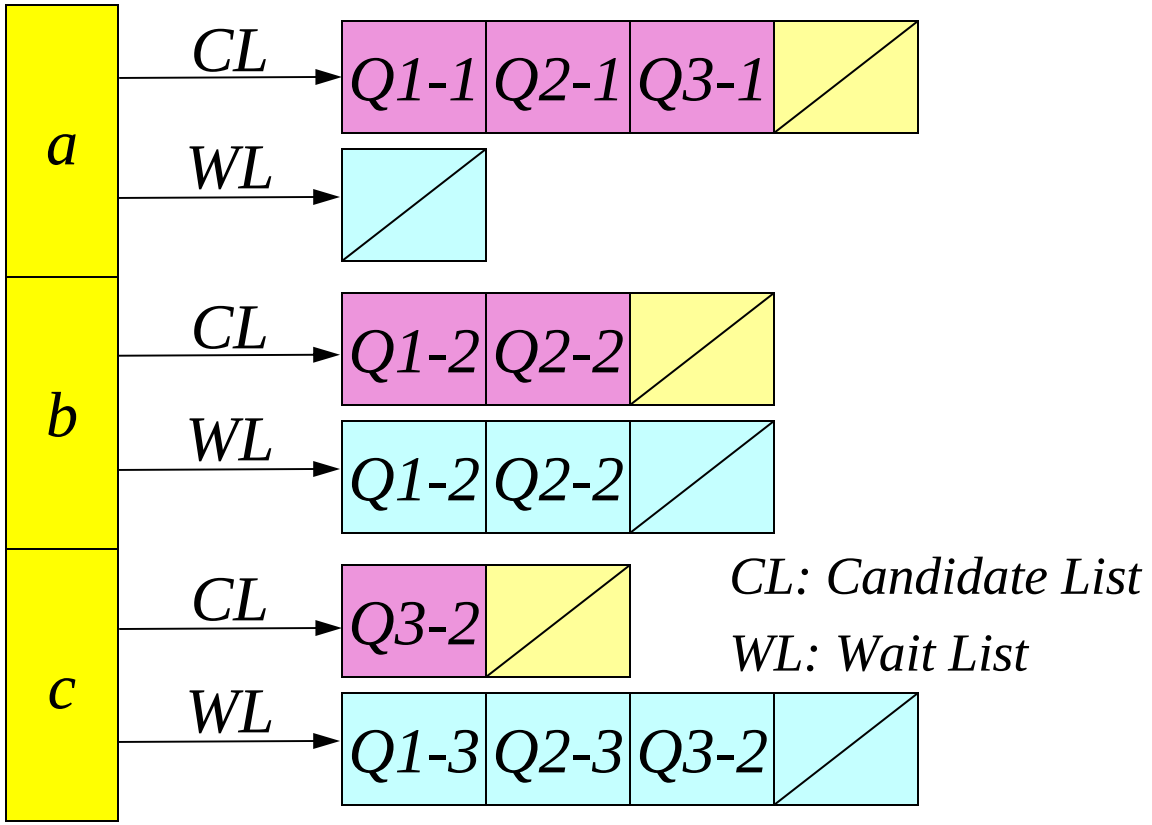
Q1 = /a/b//c



Q2 = //a/*/b/c

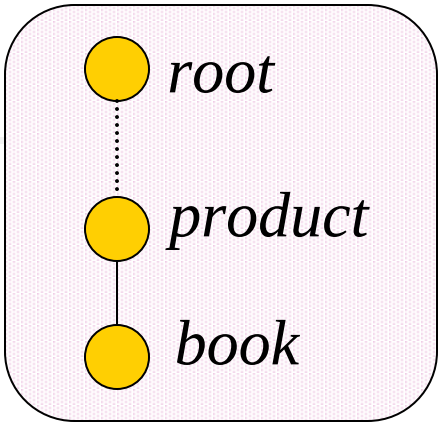
Q3 = /a/c

到着文書=/a/b



XFilter

- XMLの構造を利用したフィルタリング.
- XPath　　//product [price/value<300]/book



Q1 = /a/b//c

Q1-1　Q1-2　Q1-3

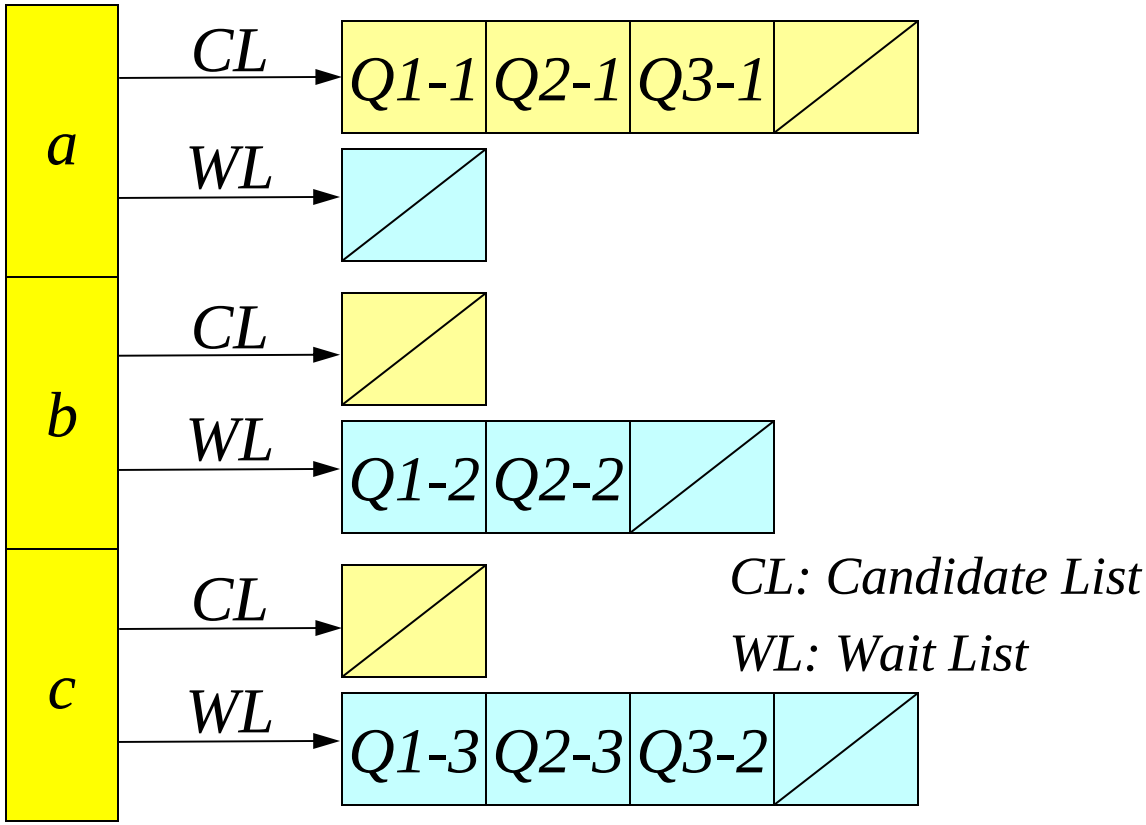
Q2 = /a/*/b/c

Q3 = /a/c

pivot

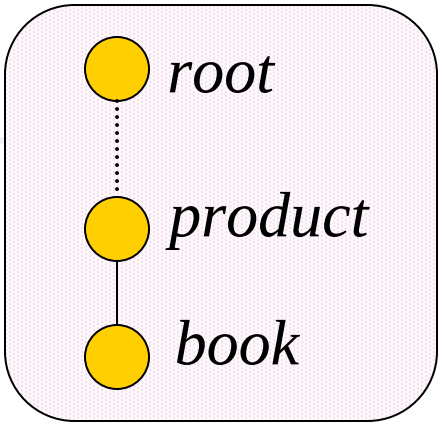
prefix

到着文書=/a/b



XFilter

- XMLの構造を利用したフィルタリング.
- XPath　　//product [price/value<300]/book



Q1 = /a/b//c

Q1-1　Q1-2　Q1-3

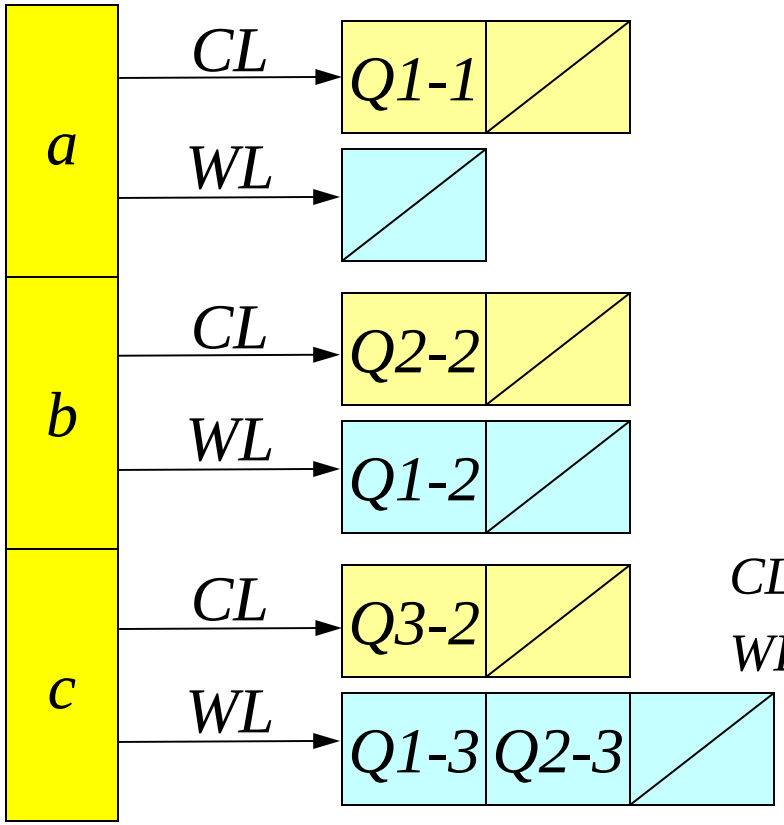
Q2 = /a/*/b/c

Q3 = /a/c

pivot

prefix

到着文書=/a/b



CL: Candidate List

WL: Wait List

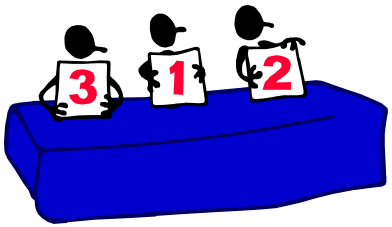
3. データの分類法に関する研究(まとめ)

- 計算コストを軽減するための手法
  - 近似を用いた手法
    - 精度 vs. 効率
  - tf.idfを用いた手法
    - ドキュメントしか扱えない.

あらゆる要素を含むマルチメディアデータに対しては不向き.

4.協調フィルタリングに関する研究

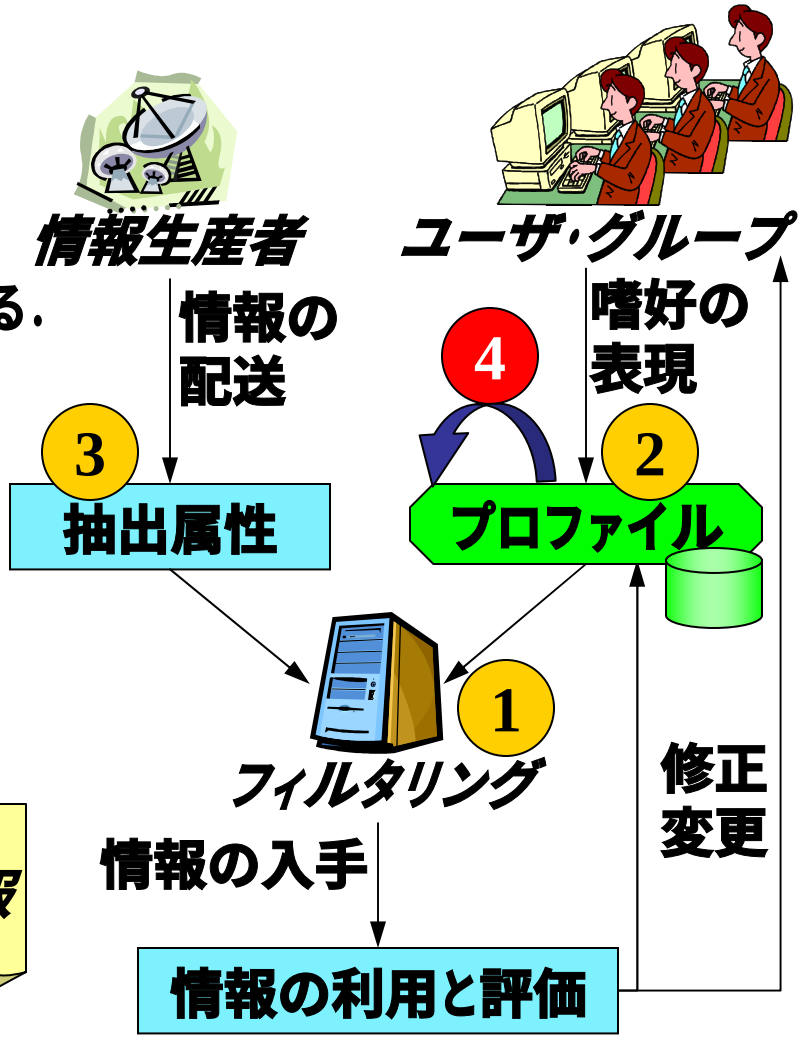
- 他人のプロファイルを利用する.
  - 嗜好が類似したユーザの
    プロファイルを組合せる.
  - 他人が評価したコンテンツを提示する.
  - 自分でプロファイルを作成
    しなくてもよい.
- プロファイルのテンプレートを
  利用する.
  - 話題のデータを取得できる.



今週1週間の
ニュース

週間天気予報

ヒット曲
ベスト10



協調フィルタリング

- **協調フィルタリングのみによるもの**
  - Tapestry *[Goldberg92](Xerox, Comm. of the ACM)*
  - Ringo *[Shardanand95](MIT, CHI)*
  - NewsWeeder *[Lang95](Carnegie Mellon大)*
  - PHOAKS *[Terveen97](AT&T, Comm. of the ACM)*
  - GroupLens *[Konstan97](Minnesota大, Comm. of the ACM)*
  - EachMovie *[Yu01](Munich大)*
- **コンテンツに基づくフィルタリングと組合せたもの**
  - Fab *[Balabanovic97](Stanford大, Comm. of the ACM)*
  - Active WebMuseum *[Kohrs99](EURECOM)*
  - ProfBuilder *[Wasfi99](Science大, Malaysia)*

協調フィルタリング

- GroupLens
  - 6段階評価
  - 嗜好の類似度 r

$$r_{KF} = \frac{\sum_i (K_i - \overline{K})(F_i - \overline{F})}{\sqrt{\sum_i (K_i - \overline{K})^2} \sqrt{\sum_i (F_i - \overline{F})^2}}$$

評価表

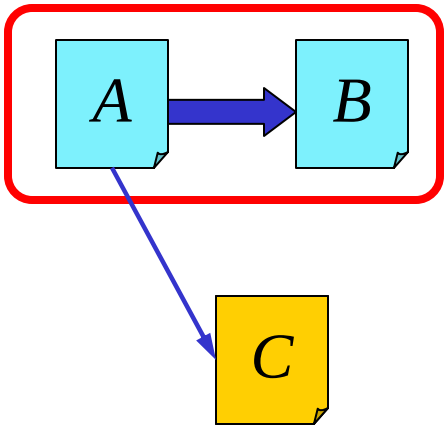
データ	Ken	Fred	Tom
a	1	2	5
b	3	2	4
c	5	6	2
d	4	3	1

  - r が大きい人のプロファイルを利用する.
  - 高い評価が付けられたデータを提示する.

類似したユーザを見つけるのがたいへん.

協調フィルタリング

- ProfBuilder
  - A → B となる確率が高い場合.
    - A と B との関連性が高いため, ユーザの嗜好を反映するものではない.
  - A → C となる確率が低い場合.
    - A と C の内容に関係はないので, ユーザは C にたいへん興味がある.



多くの人の推移率を調べるまで精度が上がらない.

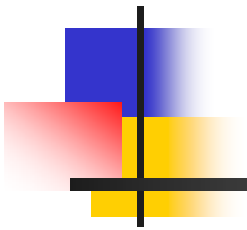
コンテンツに基づくフィルタリングとの組合せ

- ProfBuilder
  - content-based filtering
    - **ページの内容とユーザのプリファレンスの相関性を考慮.**
    - **まだ他の人が見ていないページを扱える.**
  - collaborative filtering
    - **ユーザのアクセス系列と過去のユーザのアクセスパターンを比較.**
    - **あらゆる内容, 未知の領域のページを探せる.**

手動による切替が必要.
協調フィルタリングによらない手法の問題点は解決されない.

4. 協調フィルタリングに関する研究(まとめ)

- 類似度が高いユーザのプロファイルを利用する手法
  - 組織内やグループ内では有効.
  - 不特定多数のユーザが利用するブロードバンド環境では不向き.
  - 新着データに対する評価は存在しない.



これからのブロードバンド時代における情報フィルタリング

従来のフィルタリング研究は

1. データ抽出法
   - セレクションとランキングの利点を兼ね備えた手法がない.
2. プロファイル作成法
   - ユーザの労力なしで，すぐに正確なプロファイルは作成できない.
3. データの分類法
   - あらゆる形態のデータを扱えない.
4. 協調フィルタリング
   - 多数のユーザとの類似度を調べきれない.

・・・・・

ブロードバンド時代の情報フィルタリング(チャレンジ1)

- 防犯フィルタリング
  - 街中に設置された防犯カメラを用いて，事件や不審者をフィルタリング．
- ビデオフィルタリング
  - 有害ビデオの意味的フィルタリング．

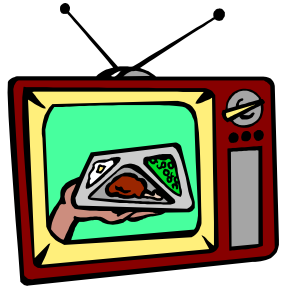
動画の意味解析
プライバシーの保護

街頭の防犯カメラ
(新宿歌舞伎町)

ブロードバンド時代の情報フィルタリング(チャレンジ2)

- 大きなデータは分割してさらにフィルタリング
  - 従来手法では，各データがまるごと得られるか捨てられるかのどちらかしかない．
  - 長時間番組は，抽出されても見るのがたいへん．
    - 好きな俳優が出ているシーンだけをフィルタリング．
    - 好きなコーナー順にランキングを作成．
    - 番組の途中で分断されたコーナーを接続．
    - CMの除去：したい vs. させたくない．

プロファイル記述方法の多様化



ブロードバンド時代の情報フィルタリング(チャレンジ3)

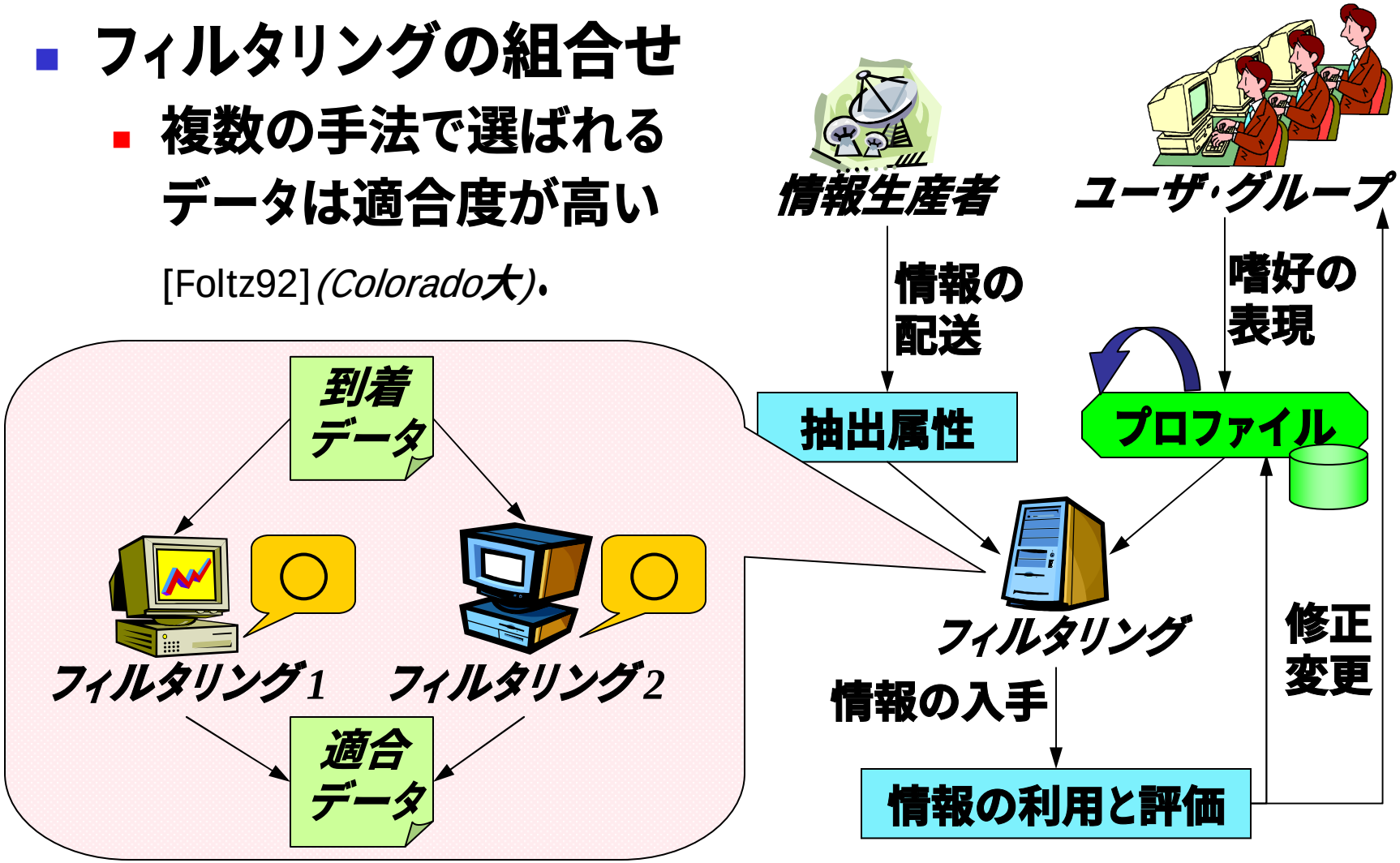
- フィルタリングのフィルタリング
  - 環境に特化したフィルタリング手法を動的に選択.
    - ディスク容量 → セレクション or ランキング？
    - ユーザによるプロファイル更新頻度 → 直接的 or 間接的？
    - 嗜好の類似したユーザ数 → コンテンツ or 協調？
    - 提供サイトによる情報料の相違 → 一番安いサービスは？
    - セットトップボックスの処理能力 → 直列 or 並列処理？
    - ユーザ数の増加によるネットワークの過負荷
      → 現在空いているサービスサイトは？

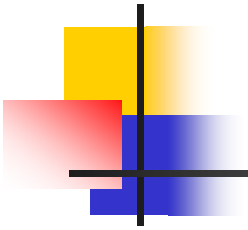
フィルタリングのコストを見積もる枠組みが必要

ブロードバンド時代の情報フィルタリング(チャレンジ4)

- フィルタリングの組合せ
  - 複数の手法で選ばれるデータは適合度が高い

[Foltz92] *(Colorado大)*.





まとめ

- ブロードバンド環境におけるフィルタリングの動向.
- 情報フィルタリングの研究動向.
- 今後のフィルタリングに必要なこと.

まだまだ課題がたくさん