SBMデータを用いたwebコンテンツ推薦

東京工業大学
宮田　高道 / 佐々木　祥

Agenda

- Aパート：SBMと研究（宮田）
  - SBM研究の着眼点と目的
  - SBMデータを用いた推薦
- Bパート：Anti-FolksonomyアプローチにもとづくSBMデータからのWeb推薦（佐々木）

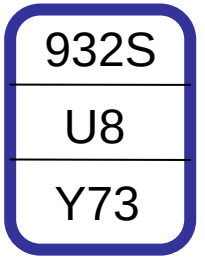
SBMと研究

- SBMとは何か？
  - URLに対してタグやコメントをつけられるもの
  - 日本の「はてなブックマーク」では…
    - コメントを受け付けないブログにもコメント可能
    - コメント欄での議論など，コミュニケーションツールの側面
  - 一方，多くのSBM関係の研究では
    - タグとコンテンツ，ユーザの関係に着目
    - コメントを分析/利用する研究は少ない

なぜタグが重視されるのか？

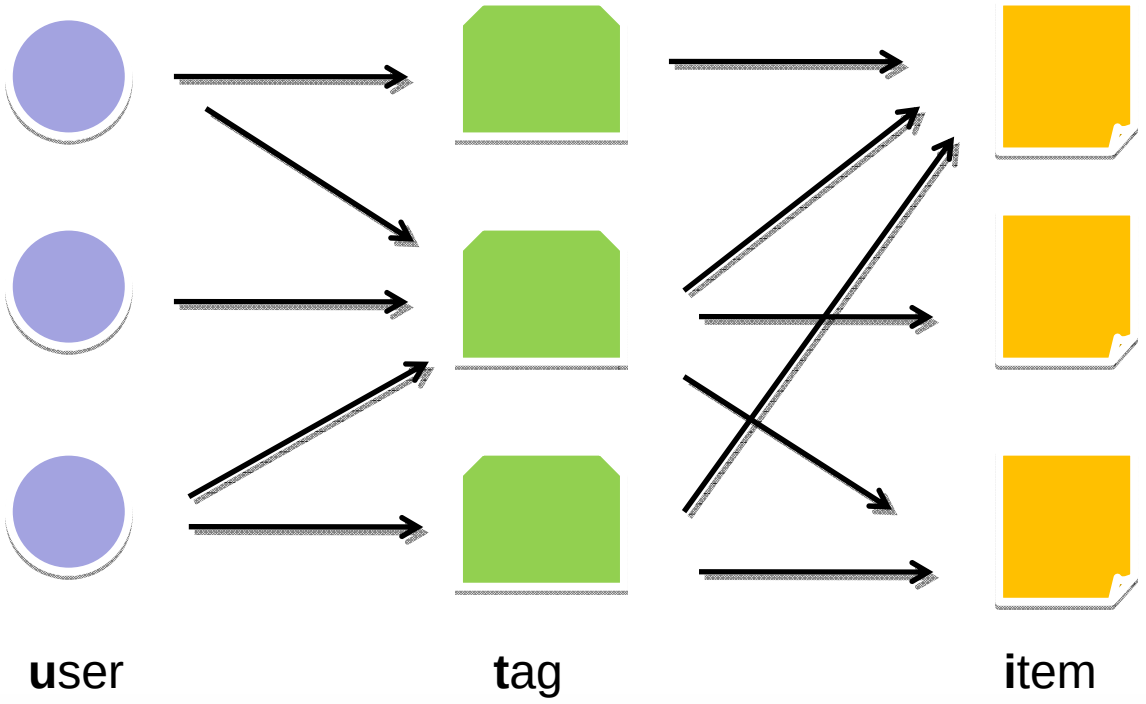
なぜタグが重要なのか

- 分類のパラダイムシフト:
  Shirky, 2005 “Ontology is Overrated”
  - Taxonomy
    - 権威による統制語彙をもちいた分類
      例:図書分類，リンネの生物分類，Yahoo!
    - 階層化されている (ネコ目-イヌ科-イヌ属-イヌ)
  - Folksonomy
    - みんな(folks)が自由な語彙でtag付け(権威はいない)
    - すべてのtagはフラット(階層なし)
    - 分類に関するコンセンサスの変化に対応可能



- SBMはFolksonomy＝「ボトムアップでの対象物の分類」を実現できるか？→実際にはいろいろな課題がある

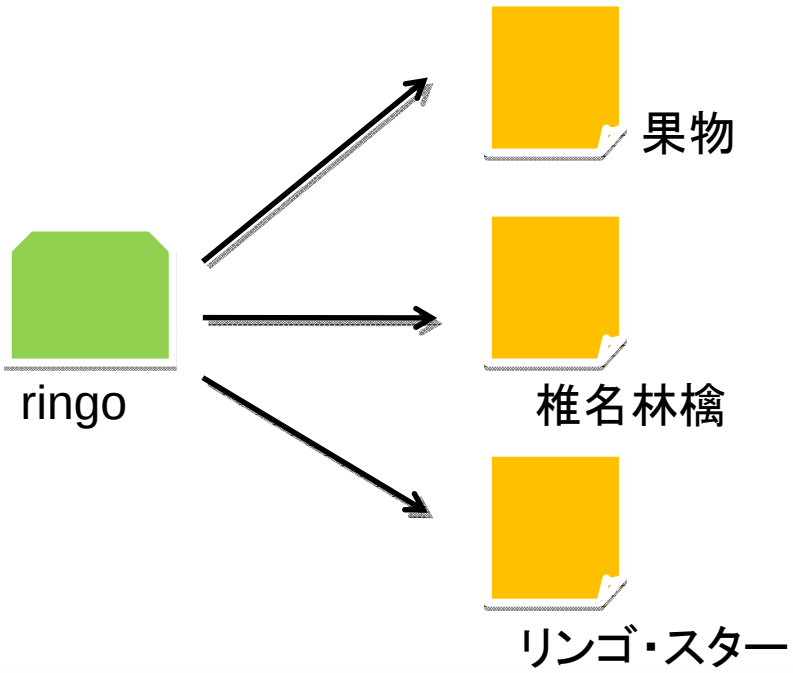
研究におけるSBMデータ



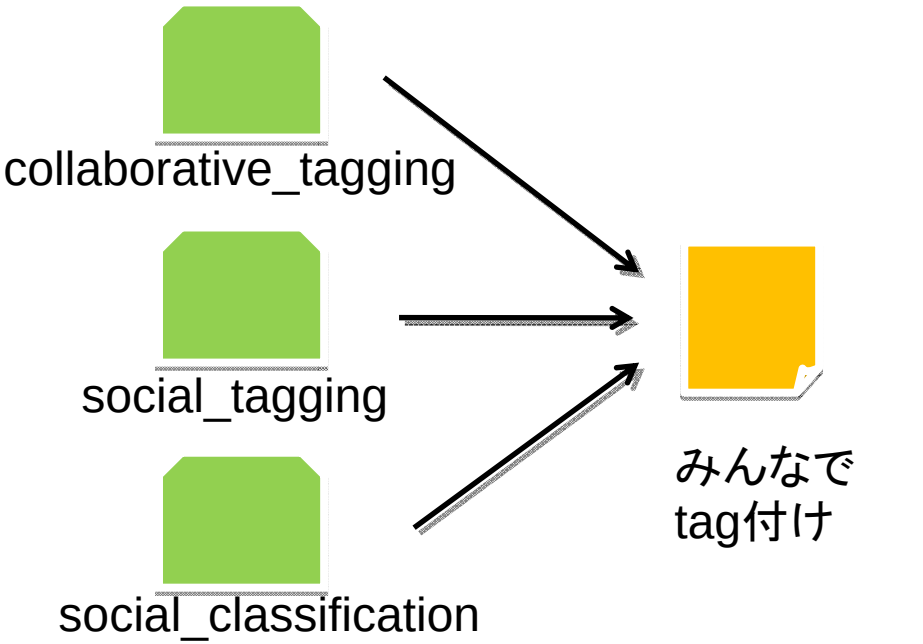
- **U**ser, **T**ag, **I**temの3つ組（3-tuple）
- Bookmarkした時刻(**D**ate)を追加することも

tag付けにおける問題点

・多義性(Polysemy)



・同義性(Synonymy)



→”生の”tagでは，itemの分類は困難

SBM研究の分類

これまでの背景をふまえ，SBM研究を分類：

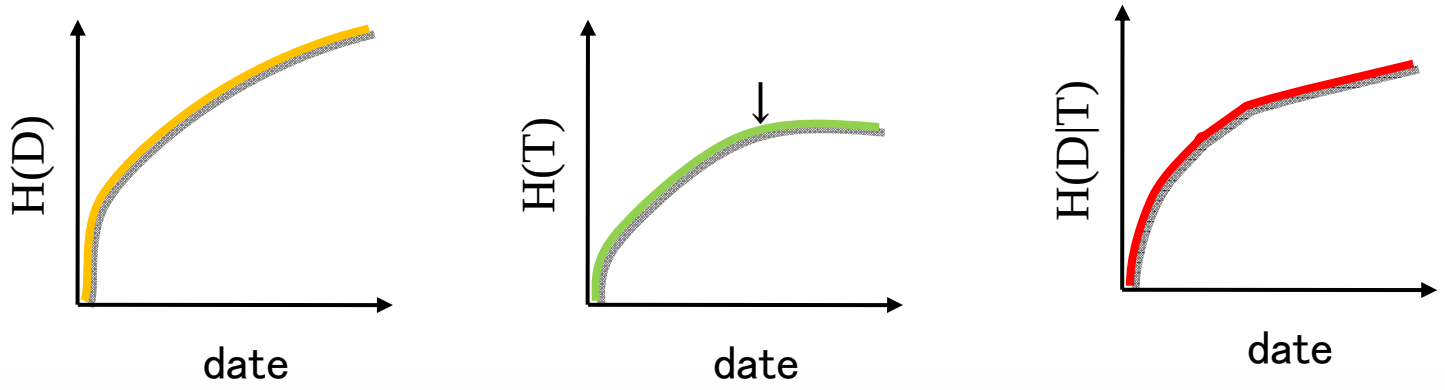
- SBMの分析
  - userのtag付け傾向などを分析
- SBMデータの補正
  - tagデータを利用したitemの分類を実現
    - Tagを階層化する
    - 類義語tagをひとまとめにする（tagのcluster化)
  - tagをクエリとした検索を可能とする
- SBM推薦：tag, user, dateのデータを活用したitemの推薦

SBM分析—tag付け傾向の分析

- Golder, 2005 “The Structure of Collaborative Tagging System”
  - del.icio.usのtagを機能別に7種類に分類
    - アイテムが何に関するものか ex) cat, Microsoft, Steve Jobs
    - アイテム自身が何であるか ex) article, book, blog
    - アイテムの品質や性質を示したもの ex)funny, stupid, これはひどい
    - タスク管理　ex) toread, jobsearch,あとで
    - などなど…
  - Itemにつけられた各tagの出現頻度は長期的に一定値に収束
    - 「tagによるitemの分類」にとって明るい材料
- Marlow 2006 “HT06, Tagging Paper, Taxonomy, Flickr, Academic Article, ToRead”

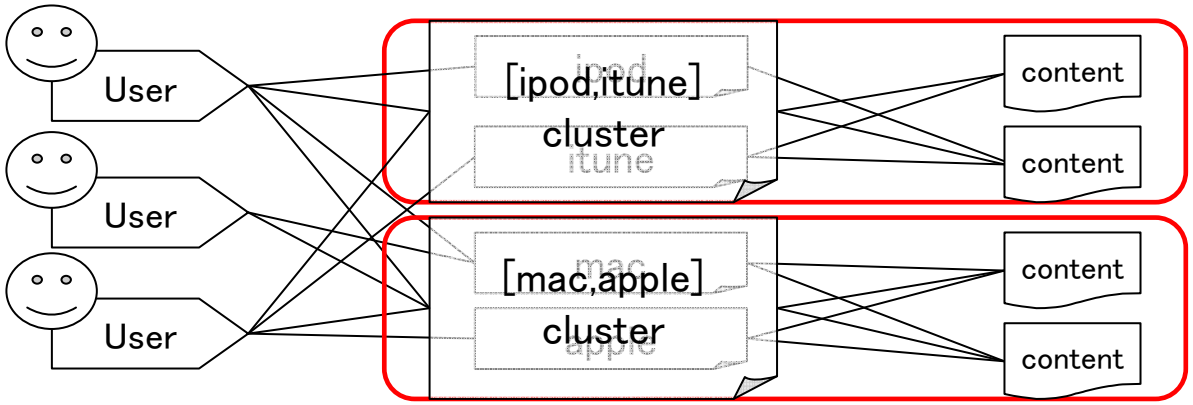
SBM分析—tagの持つ検索能力とその推移

- Chi, 2008 “Understanding the Efficiency of Social Tagging Systems using Information Theory”
  - タグTと文章Dの条件付きエントロピー $H(D|T)$は時間がたつにつれて上昇($H(D|T)=H(D,T)-H(T)$



  - (単一の)タグが(単一の)コンテンツを特定する能力は時間が経つにつれて低下している

SBM補正に関する研究



- 複数のtagをひとまとめに（clustering）
  →tagをクエリとしたitem検索を容易化
  - Begelman, 2006 “Automated Tag Clustering: Improving search and exploration in the tag space”
  - Rui, 2007 “Towards Effective Browsing of Large Scale Social Annotations”
    - tag-itemの関係から類似するtagをまとめる→tag cluster形成
    - Intersection rateなどに基づき類似tagを階層化

SBMを用いた情報推薦

SBMデータの分析

↓

SBMデータの補正

↓

SBMデータを使った情報推薦

・多くの検討が広義の協調フィルタリングを使用

協調フィルタリング(CF)

GroupLens(1994)

負の相関性　類似度= -0.8

	Alice	Bob	Clair	David
Happening	1	5	2	2
Matrix	5	2	5	3
Gattaca	–	–	3	–
Sixth sense	2	4	2	2
Star wars	4	1	–	3
Speed racer	?	2	5	2

正の相関　類似度= 0.9

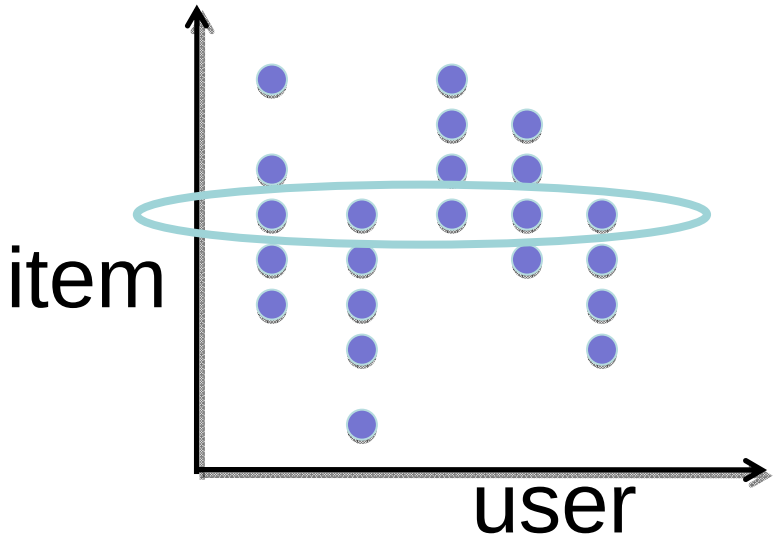
- AliceはSpeed racerに何点をつけるだろうか？
- Aliceとその他のユーザのratingの様子を比較→類似度算出
  - ピアソン相関係数
  - Cosine係数
- 他ユーザのSpeed racerについてのratingを類似度に基づいて足し合わせる→Aliceのratingを予測

協調フィルタリングの課題

- 協調フィルタリングの限界
  - ユーザの趣味があらゆる側面ですべて一致することは少ない
    - SBMのtagを利用すれば，ユーザ–ユーザ推薦以外の推薦が可能？
  - データが十分に密(dense)でないといけない
    - データが大量にあつまる環境が望ましい

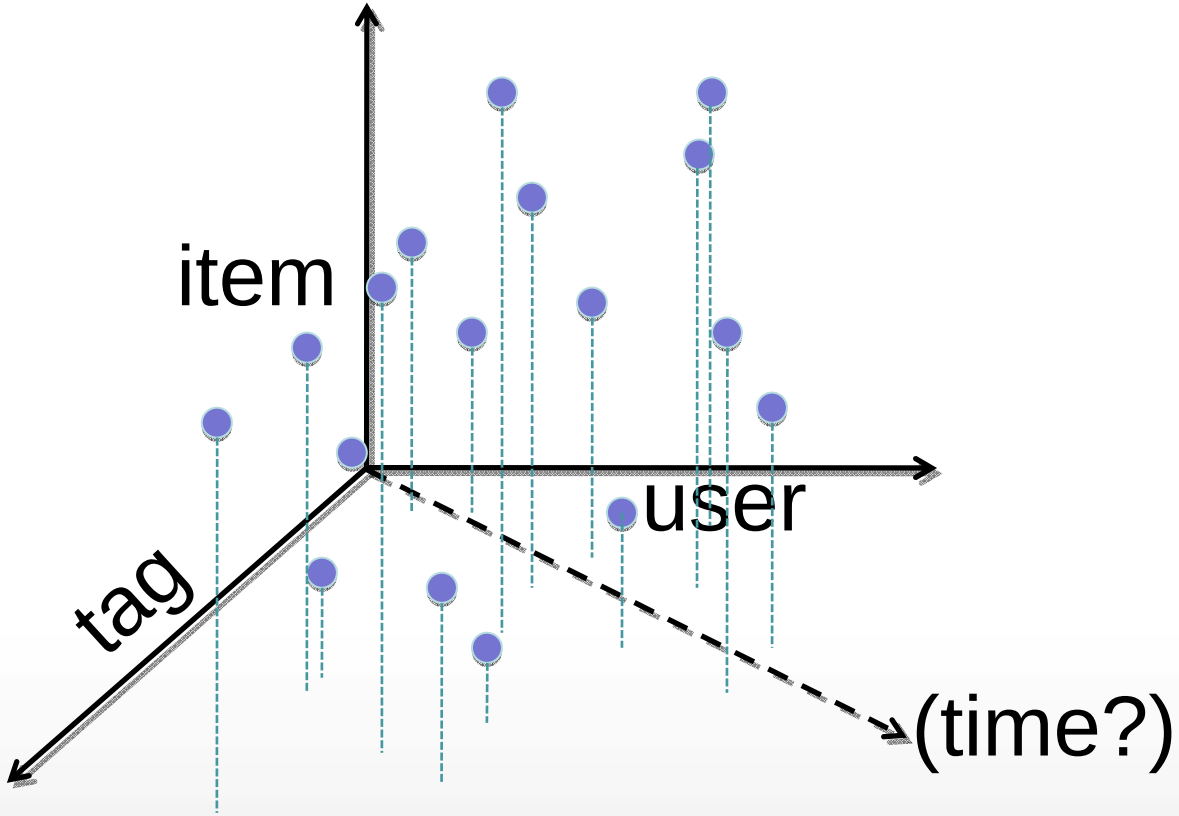
協調フィルタリング×SBMデータ

- SBM data is too sparse to apply CF!



協調フィルタリング×SBMデータ

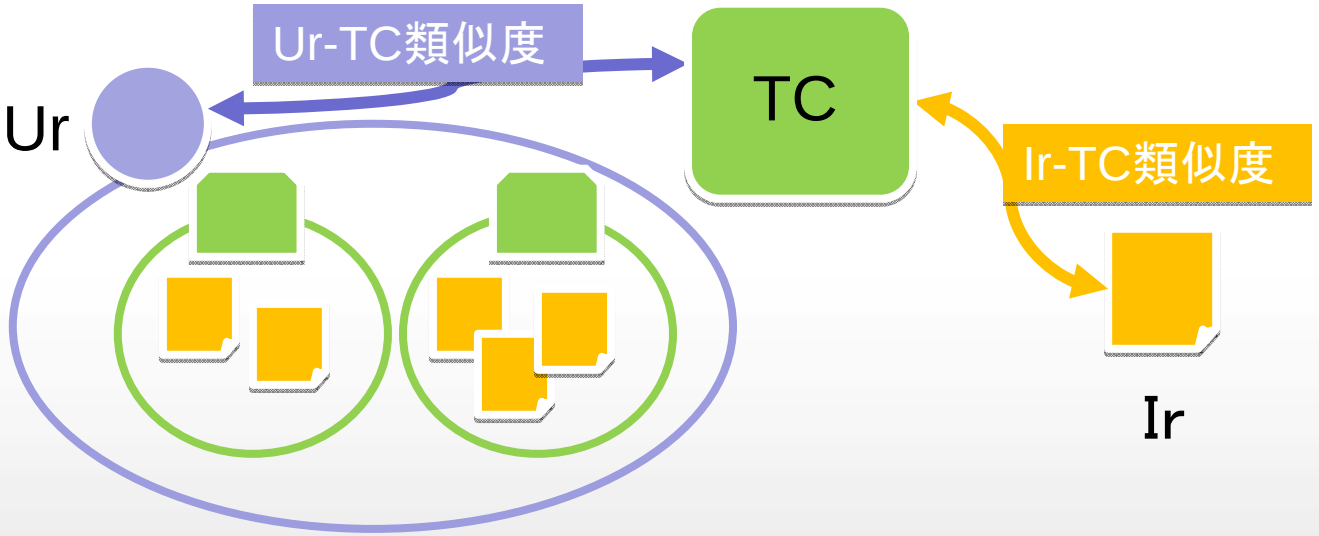
- SBM data is too sparse to apply CF!



- いかにして圧縮を行うか？

アプローチ(1) タグの縮約と推薦

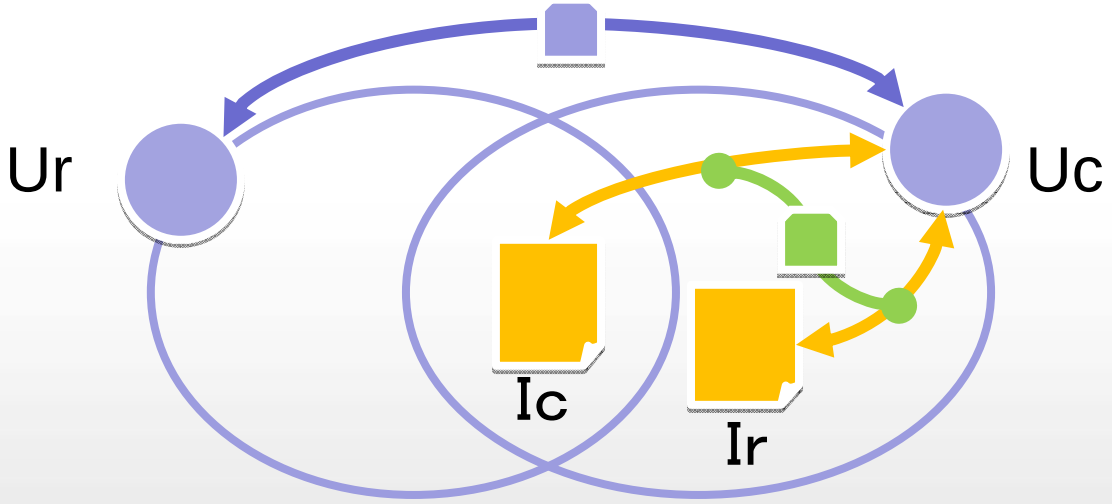
- Niwa, 2006 “Web Page Recommender System based on Folksonomy Mining”
  - tagをクラスタ化(tag cluster, TC)して”表記のゆらぎ”を吸収
  - 全userと全itemについて，それぞれ全TCとの類似度を算出
  - 被推薦user:Urと推薦対象Item:Irとすると，
    Irの推薦度=Ur-TC類似度 * TC-Ir類似度



アプローチ(1) タグの縮約と推薦

- 中本, 2008 “多様に変化するユーザの興味を考慮したウェブサイト推薦システム”
  - tag-item共起から100個のTCを構成
  - user間の類似度＝両userの共通itemに関する100次元TCベクトルのcosine係数の平均値(…とブックマーク共起数との重み付き和)
  - いま，Urと共通のitem:Icと，Irの両方をもつuser:Ucとする

    Irの推薦度＝UcとUrの類似度 × IcとIrに関するUcのタグ付け傾向の一致度(…とUcとUrの類似度の重み付き和)



アプローチ(2) 時系列データの活用

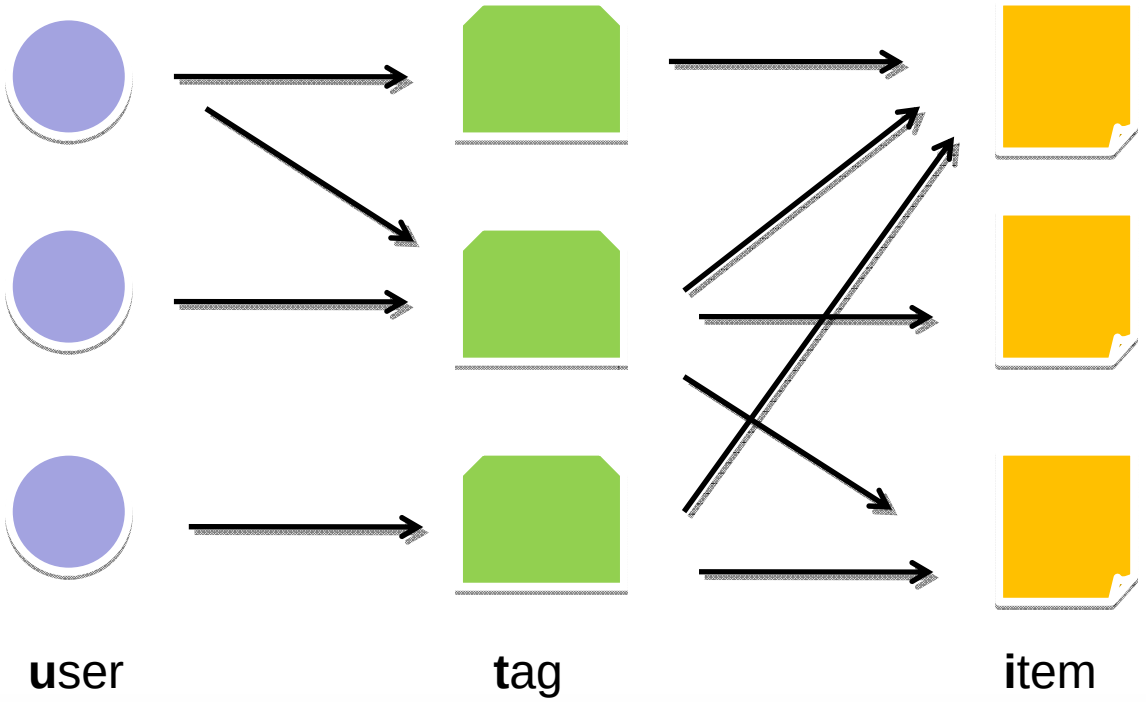
国内では”αブックマーカ発見”などの時系列データを用いた推薦が流行

- 大力, 2008 “ソーシャルブックマークにおけるイノベータに着目した情報推薦手法の提案”
  - あるURLを最初にSBMに登録した人=1ゲッター
  - 全体URLの9割くらいが僅か83人(全体の2%)のユーザに1ゲットされてる
  - 上位83人のユーザ(イノベータ=αブックマーカ)と被推薦ユーザの間のcosine類似度で推薦
  - 推薦に必要なデータが少ない
- 百田, 2008 “ソーシャルブックマークに基づく情報発見”
  - Intersection rateを使ったタグの階層化→Folksonomyの問題を解決
  - αブックマーカを発見
  - ランダムに選んだユーザから推薦されるより，αブックマーカから推薦させたほうが精度がよい

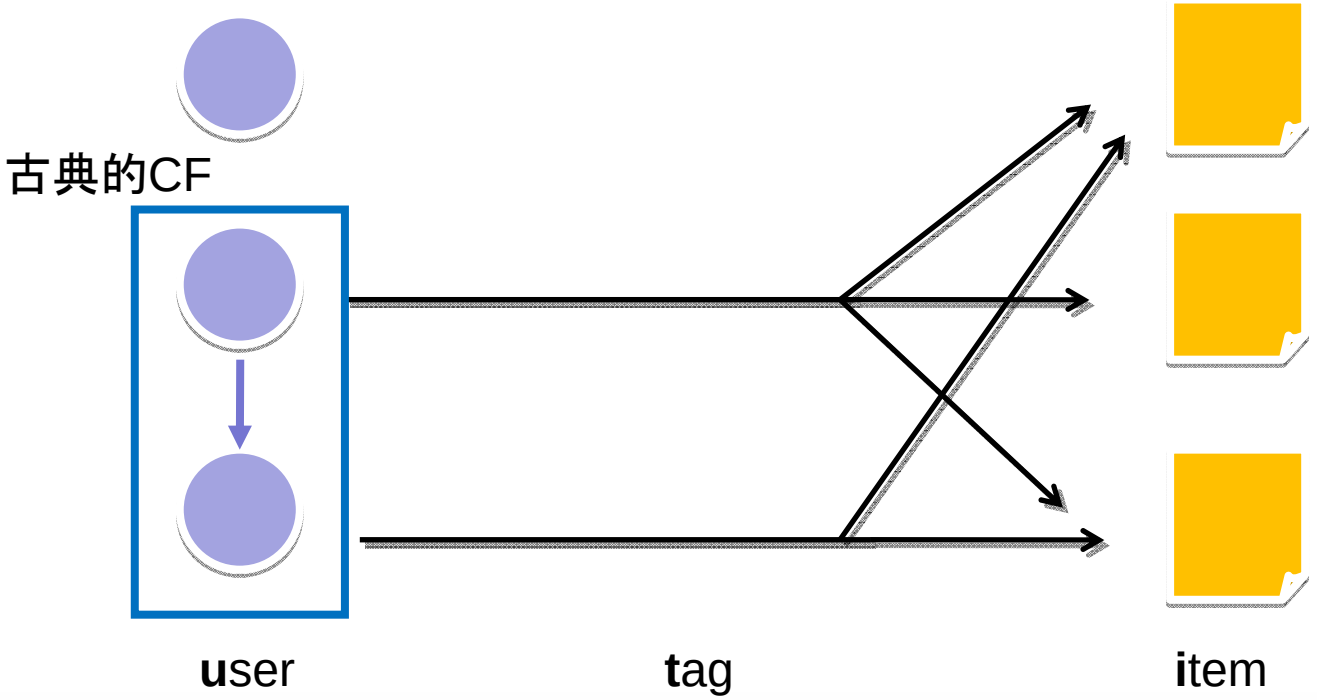
アプローチ(3) その他

- 正統的なLSAアプローチ
  - Xu, 2005 “Cubic Analysis of Social Bookmarking for Personalized Recommendation”
- Webコンテンツの内容（文章）を利用
  - 矢島，2008 “ソーシャルブックマークにおける文書解析を利用した類似文書および類似ユーザの推薦方法の提案”

既存手法のまとめと課題

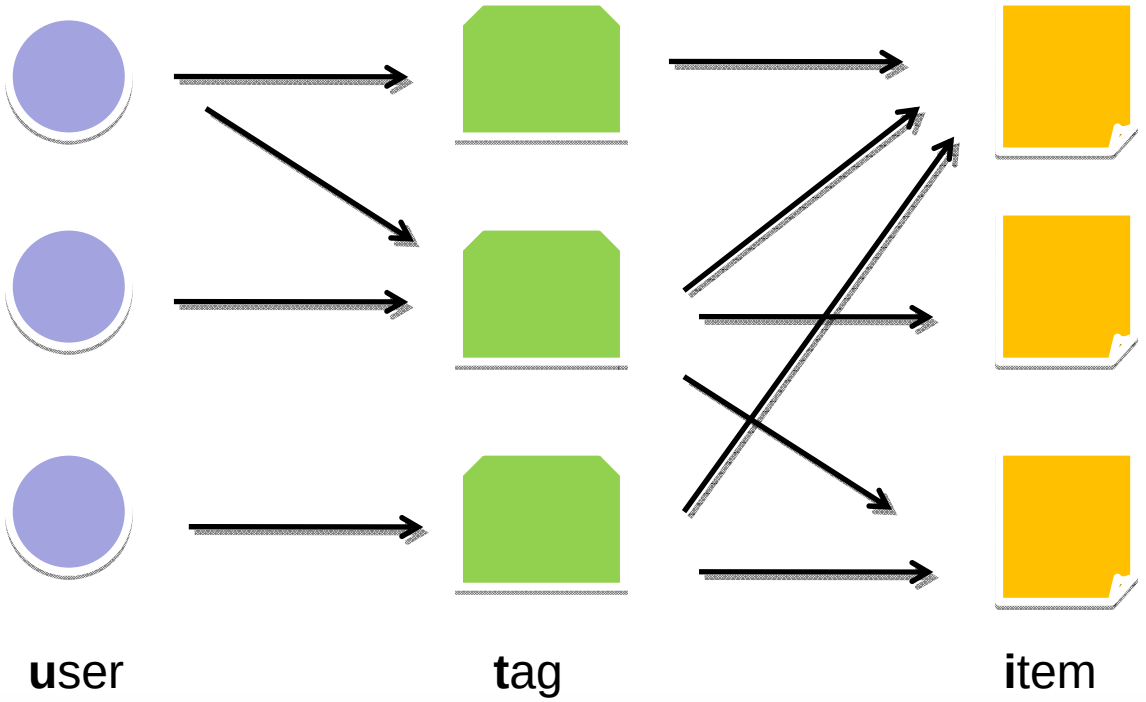


既存手法の課題:古典的CF



- すべての趣味が一致するとは限らない
- 時系列データからαブックマーカを抽出する研究も多くはこちらに属する

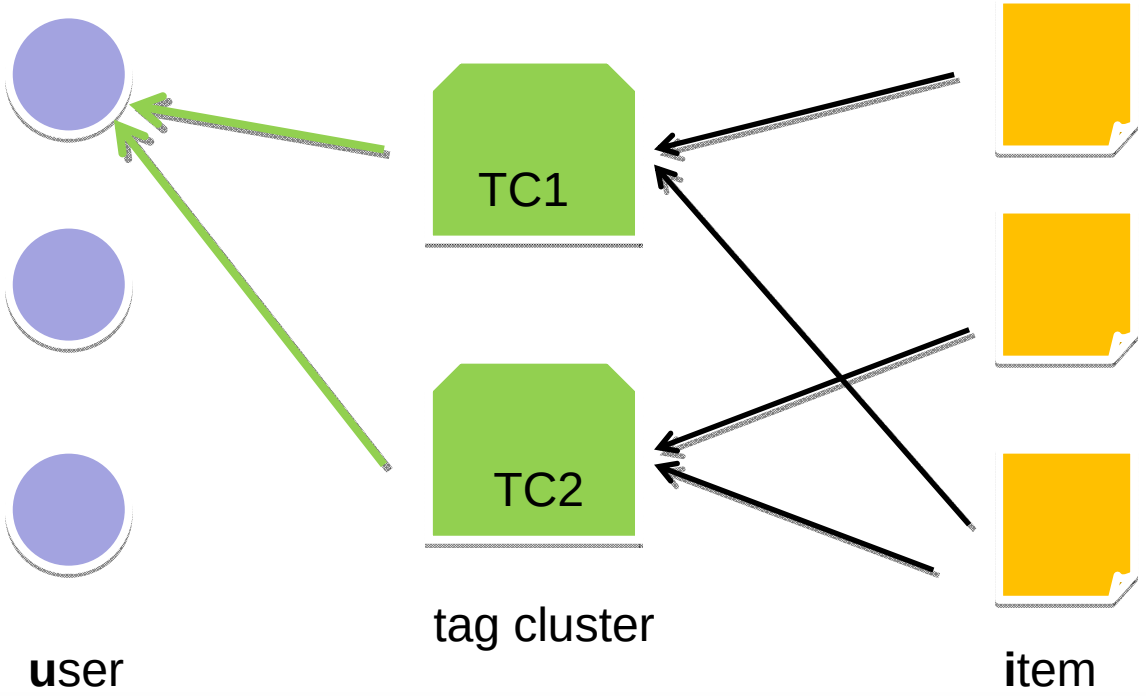
既存手法の課題:SBM利用の問題点



- 趣味の多様性を反映:SBMデータの利用
  →tagのvocabularyのばらつきが問題に

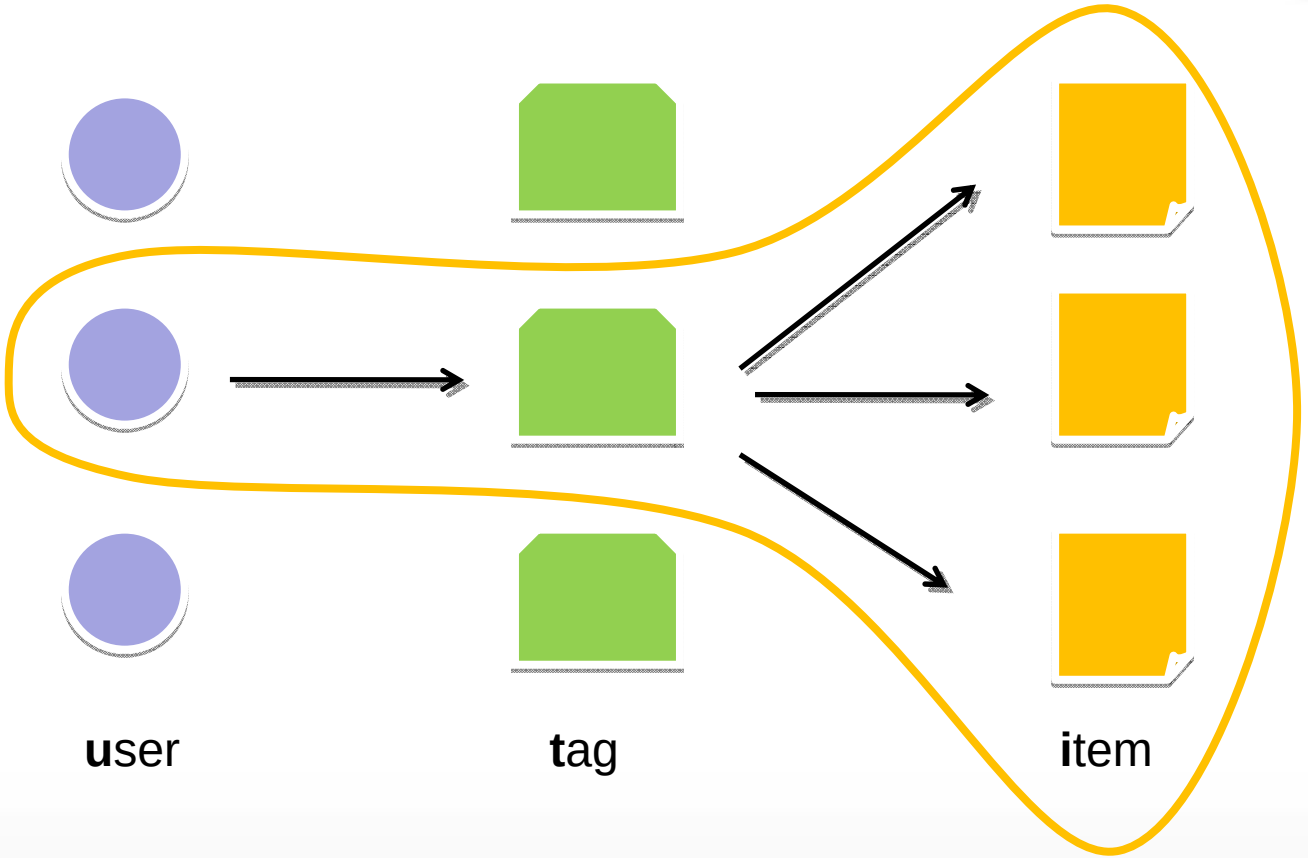
既存手法の課題:タグ縮約の利用

tag clusterを使う手法



- Itemがいくつかのtag clusterに所属:sparistyの解消
  →Itemは正しいtagで分類されるべき，という考え方
- **推薦**にとって，本当に”itemの正しい**分類**とtag付け”は必要なのか？

Vocabulary is NOT important, only group is.



- タグに使われているwordなんて飾り
  →Tagの役割はitemのグルーピングだけ
- グループ間の類似度をどう定義するか　→Bパートへ続く…

さまざまなSBM研究

- 時系列データの活用，
  - 毛受, 2008 “ブックマークの時系列情報を利用したソーシャルブックマークにおける注目度予測”
- 被ブックマーク数による検索結果の再ランキング
  - 山家, 2007 “ソーシャルブックマークの特性を利用したWeb検索のランキング精度の向上”
    - あるitemがどれだけbookmarkされたか＝SBrank
    - SBrankとPagerankを両方利用して検索結果を再ランキング

Bパート:
Anti-Folksonomyアプローチにもとづく SBMデータからのWeb推薦

東京工業大学大学院理工学研究科
集積システム専攻　酒井研究室
佐々木　祥

本研究のねらい

- SBMを利用したwebコンテンツ推薦
  - SBM履歴からの自動的な嗜好予測
  - 情報共有ツールとしてのSBMの延長線上にあるサービス

- タグ単位の推薦
  - タグ別の嗜好単位推薦
  - 人間の多様な側面には多様な推薦が必要

本研究の新規性

- タグ名称を用いない情報推薦
  - 従来のSBMに基づく推薦の研究
    - タグの名称に着目，Folksonomyに基づく・・・

⇒ Anti-folksonomyな推薦！

- 統計的信頼性を考慮した協調フィルタリング
  - 従来システムは試行錯誤的
    - Jaccard係数，コサイン係数など単純な比率
    - 発見的に閾値を設定して利用範囲を決定

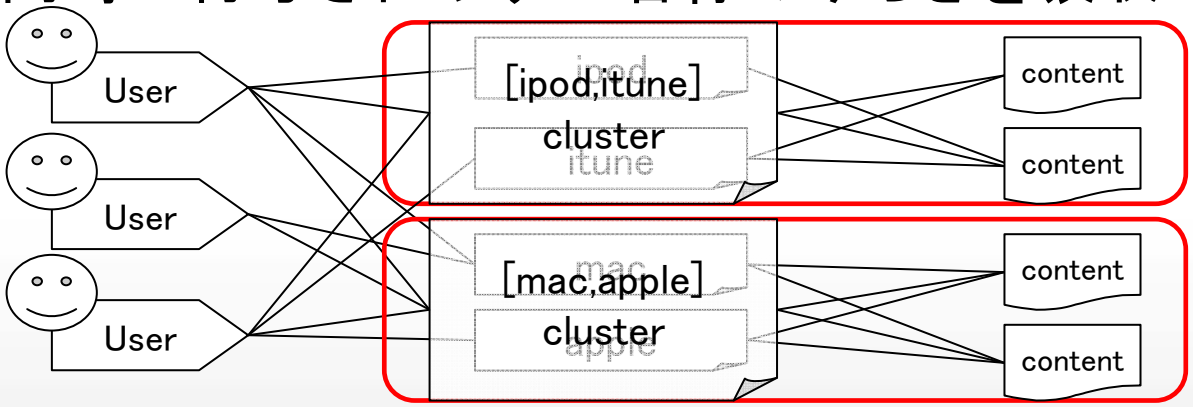
⇒ 尤度比検定を導入！

従来研究・タグの名称を利用したwebコンテンツ推薦

既存研究：Folksonomyを利用した推薦システム

Web Page Recommender System based on Folksonomy Mining (ITNG'06)

- コンテンツ-タグの関係が類似するタグをクラスタ化
- 同時に付与されやすい名称のゆらぎを吸収



タグの名称の類似性を判定・・・表記ゆらぎの解決

Folksonomyへのアンチテーゼ

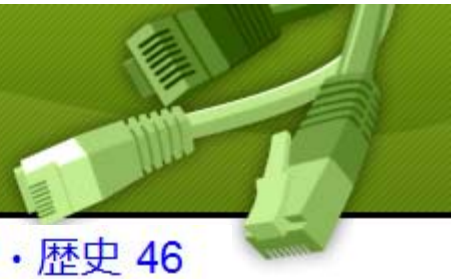
- タグのゆらぎは単なる記述結果の違いか
  - Folksonomyの根底に「協調的タギング」
    - 他のユーザの利益のため・・・

  SBMerはコンテンツを解説する義務はない

  - 事実，みんな好き勝手につけている
    - 顔文字
    - 勝手に階層構造化
    - アルファベット検索・入力のための頭文字

  自己の利便性を重視している

SBMの実データの検証
ユーザが使っている極端なタグ例



tags

10 (#ﾟДﾟ)ｺﾞﾙｧ!!
121(`･ω･´)ｼｬｷｰﾝ
22 (´Д`)σ)Д`)ﾌﾟﾆｮﾌﾟﾆｮ
17 (´･ω･`)ｼｮﾎﾞｰﾝ
36 (σ･∀･)σｹﾞｯﾂ!!
12 (　´д)ﾋｿ(´д｀)ﾋｿ(д｀　)
72 (　ﾟДﾟ
11 (　ﾟДﾟ
52 (　ﾟДﾟ)ﾎﾟｶｰﾝ
8 (・A・)ｲｸﾅｲ!
124・∀・)ｲｲ!!
14 (ﾟДﾟ)ﾏｽﾞｰ
36 (ﾟдﾟ)ｳﾏｰ
1 *.do

21 う_宇宙
2 う_歌
8 え_エッシャー
17 え_エミュレータ
31 え_映画
37 え_絵
51 え_英語
6 お_おもちゃ

本当にこのデータで推薦できるの？

172・ω・)っあとで
4 (・ω・) あれ？
6 (・ω・) うん？
2 (・ω・`)あらら
39 (・ω・`)おぉ
13 (・ω・`)なーる
80 (・ω・`)ふむふむ
5 (・ω・`)禿げしく同意

4 study/english
5 study/P2P
2 study/pdf
10 study/SBM
10 study/scale-free
8 study/scale-out
11 study/search

45地理・歴史 46
食 47ラーメン 48
酒 50文化 51漫
画、アニメ 51美・
デザイン 52音楽
53映画 54作家・本

かん3世 おぞ
ましい おたく おっ お
なら おはきな おめでと
うございます おわび お
フランスざんす おロシ
ア お仕事るんるん お勉
強 お受験 お年頃 お店
お役人クオリティ お役

推薦にFolksonomyは必要か

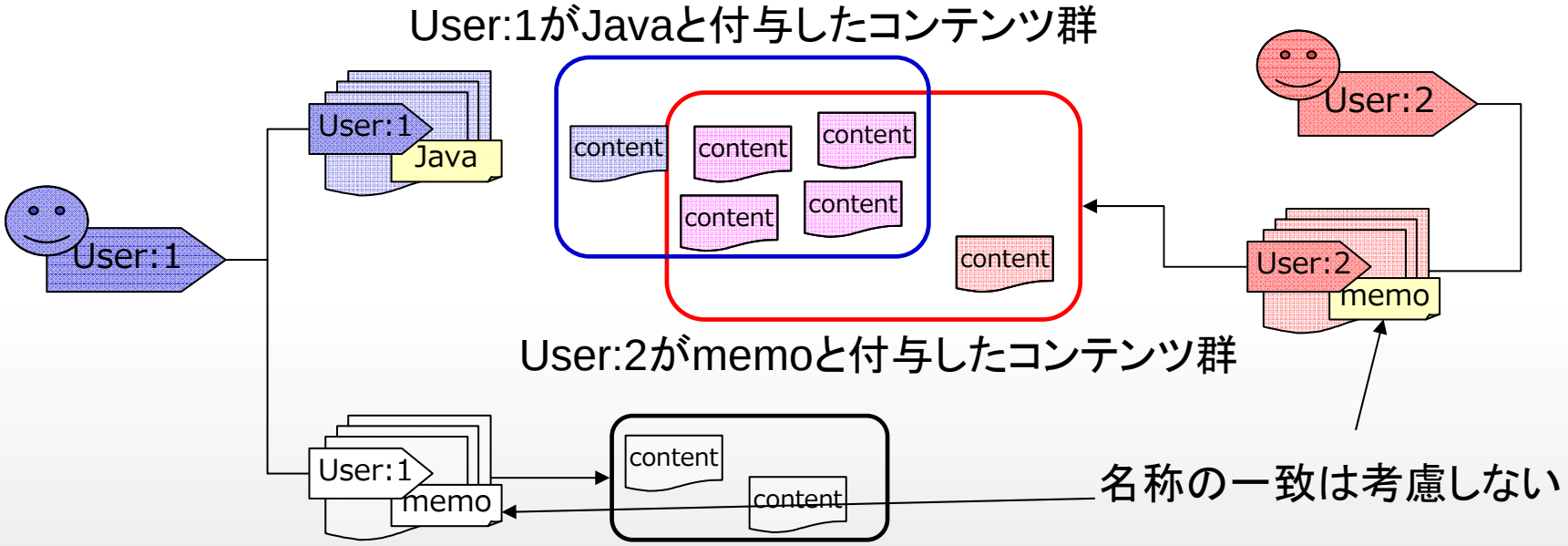
- 多くのSBMを用いた推薦研究
  - タグの名称に着目してしまう
    - どうやってセマンティックを見出すか・・・
    - タグの表記ゆらぎを解決するには・・・
  - そこにタグ情報があるから・・・？

- 偏見を捨てよ　さすれば道は開けん
  - タグの表面的な字面を破棄
  - タグによる個人のグルーピングに着目
    - 興味を持って整理したコンテンツ群にこそユーザの嗜好を見出せるのではないか

研究アプローチ・タグ名を見ない

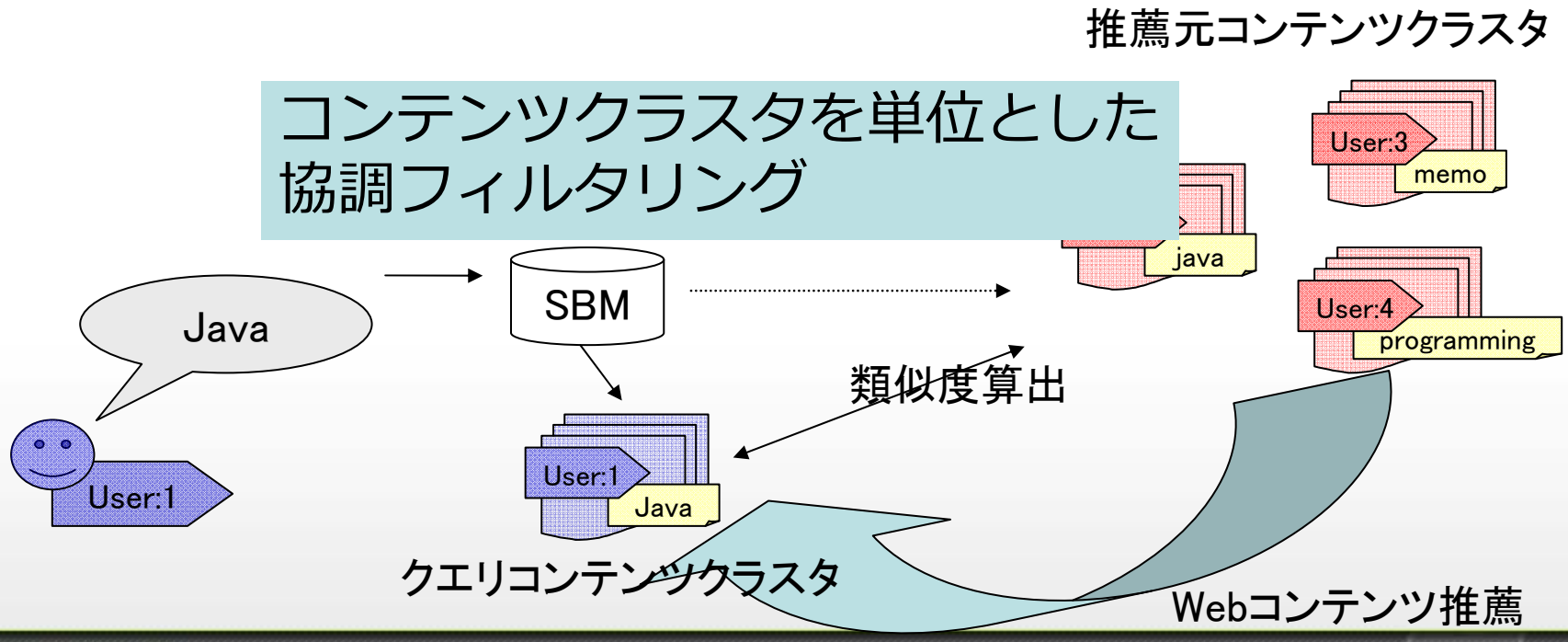
タグの名称に依らない推薦

- あるユーザがタグでまとめたコンテンツ群に着目（以下，コンテンツクラスタ）
- コンテンツクラスタ同士の共起性に基づき推薦



提案システムの概要

1. クエリコンテンツクラスタをシステムに入力
   （ユーザID，タグ名が入力となる）
2. SBM中の複数の推薦元コンテンツクラスタとの類似度算出
3. 推薦元コンテンツクラスタからwebコンテンツを推薦



本研究の新規性

- タグ名称を用いない情報推薦
  - 従来のSBMに基づく推薦の研究
    - タグの名称に着目，Folksonomyに基づく・・・

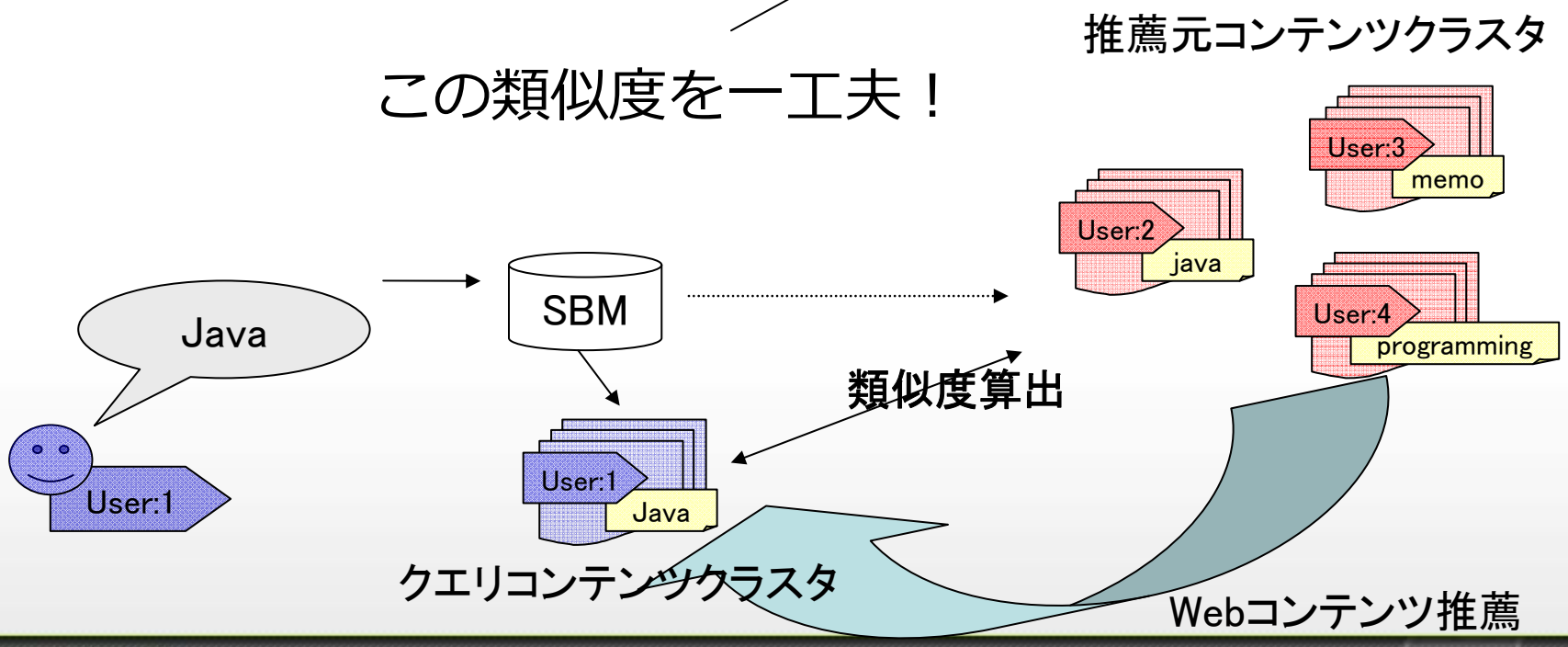
→ Anti-folksonomyな推薦！

- 統計的信頼性を考慮した協調フィルタリング
  - 従来システムは試行錯誤的
    - Jaccard係数，コサイン係数など単純な比率
    - 発見的に閾値を設定して利用範囲を決定

→ 尤度比検定を導入！

システム概要

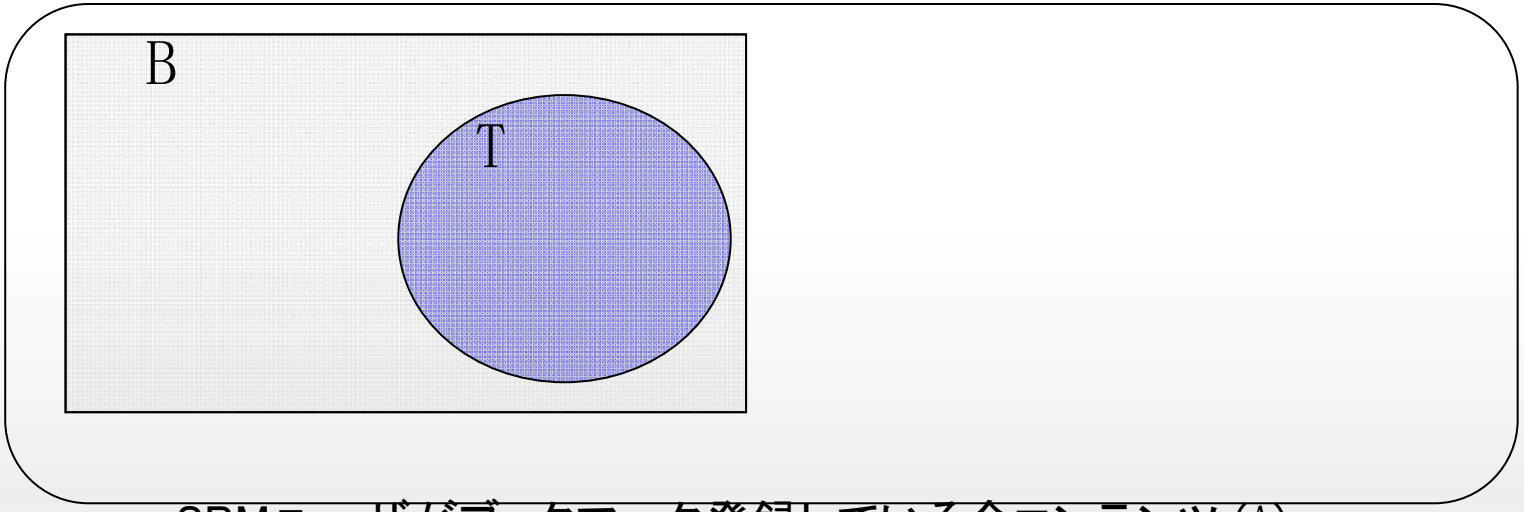
1. クエリコンテンツクラスタをシステムに入力
   （ユーザID，タグ名が入力となる）
2. **SBM中の複数の推薦元コンテンツクラスタとの類似度算出**
3. 推薦元コンテンツクラスタからwebコンテンツを推薦



コンテンツクラスタ

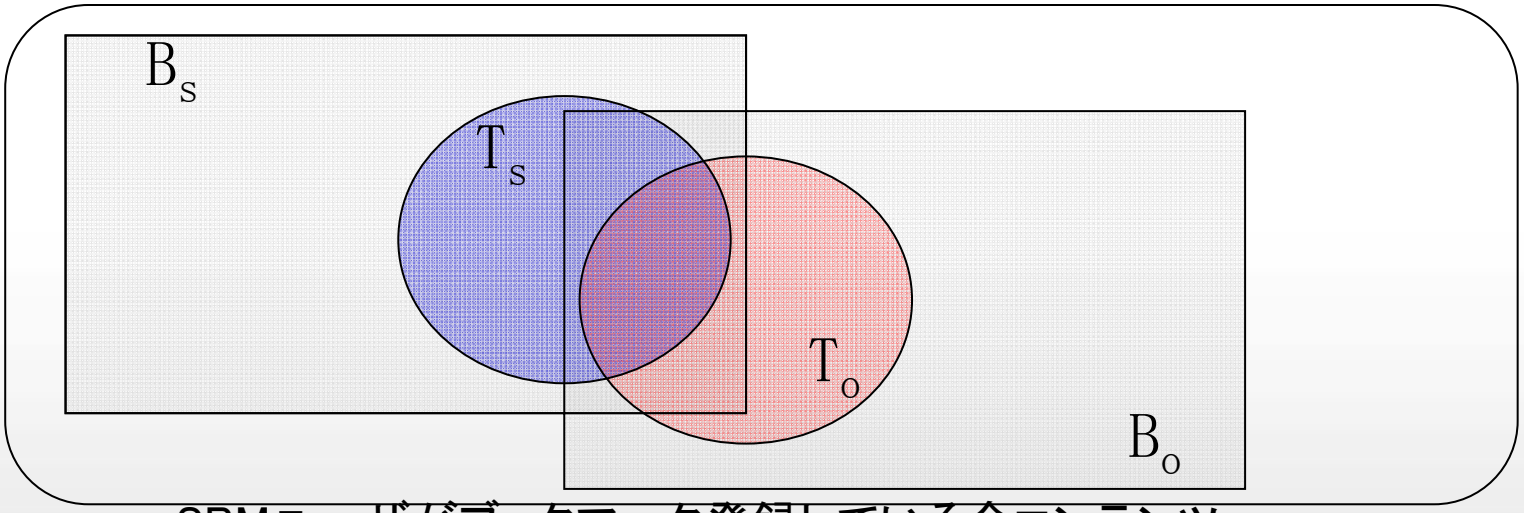
User:1 の Tag:a によるコンテンツクラスタ

- 全ユーザのブックマーク集合 (A)
  - あるユーザUser:1のブックマーク集合 (B)
    - User:1にタグTag:aを付与されたブックマーク集合 (T)



二つのコンテンツクラスタ

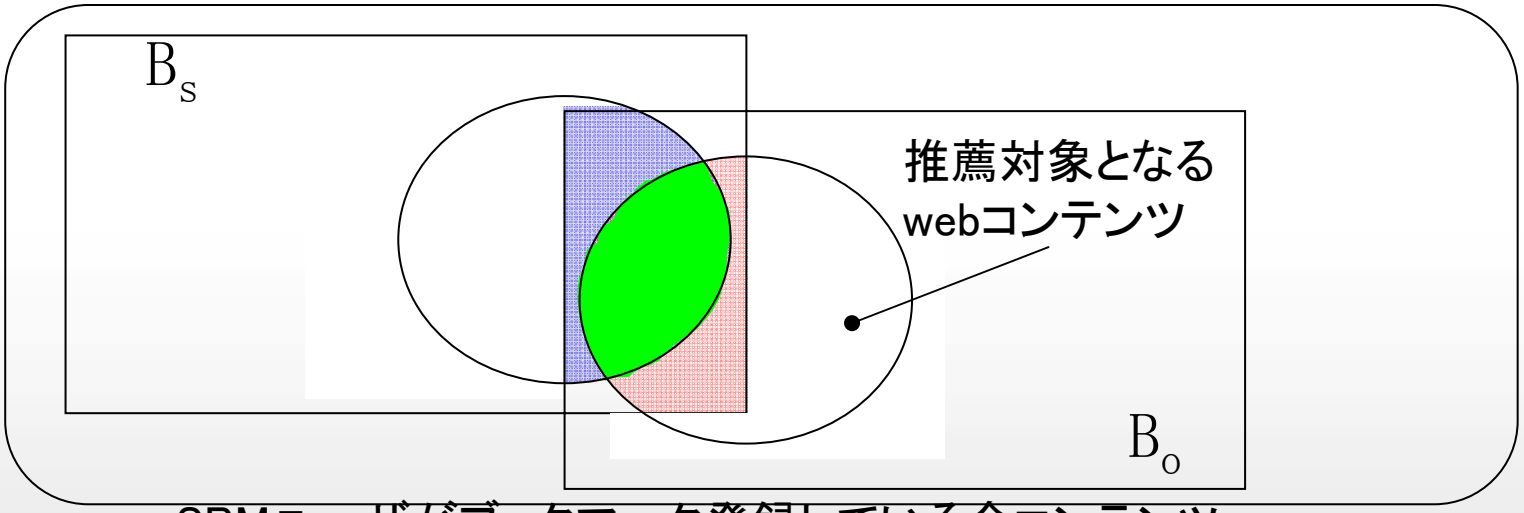
- クエリコンテンツクラスタ s (User:1,Tag:a)
  - B_s User:1にブックマークされているwebコンテンツ集合
  - T_s User:1にTag:aを付与されているwebコンテンツ集合
- 推薦元コンテンツクラスタ o (User:2,Tag:b)
  - B_o User:2にブックマークされているwebコンテンツ集合
  - T_o User:2にTag:bを付与されているwebコンテンツ集合



SBMユーザがブックマーク登録している全コンテンツ

コンテンツクラスタ間の共起関係

- コンテンツクラスタ s,o 間の共起性に関する数
  - 和集合 $n(s, o) : (Bs \wedge Bo) \wedge (Ts \vee To)$
  - 積集合 $k(s, o) : Ts \wedge To$
  - $n(s, o)$, $k(s, o)$に従ってwebコンテンツを推薦



SBMユーザがブックマーク登録している全コンテンツ

従来の協調フィルタリングと問題点

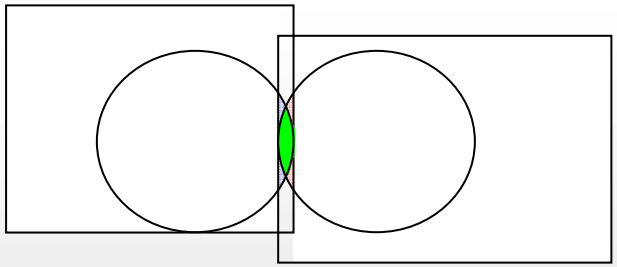
- 集合の類似性尺度
  - Jaccard，cosine係数など

$$\mathrm{Jaccard} = k(s,o) \ / \ n(s,o)$$

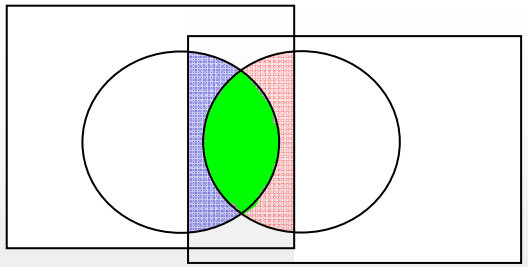
以降，(s,o)略

  - 単純な割合
  - 分母nが小さいときkによって値が大きく変化
    - 偶発的な一致が推薦に悪影響を与える

2 個中 2 個hit ＞ 40 個中 35 個hit ←?



全体が小さく，Jaccard大



全体は大きいが，Jaccard中程度

提案アプローチ・確率論を導入

- 統計的な信頼性を考慮した類似度の算出
  - タギングのベルヌーイ試行モデル化
    - 2者のコンテンツクラスタに
      コンテンツが同時に帰属する確率（一致度）p
    - 試行回数: n , 成功回数: k となる尤度は？
  - 仮説検定を利用した一致度の判定
    - 二つの一致度設定：p0（不一致） , p1（一致）
    - どちらの一致度であるのが尤もらしいか
      （尤度比検定） → 類似度とする

仮説検定の例:コイン判定 1/2

- コインCはどちらかである
  - Ct:正確に表，裏が1/2で出るコイン　p1=1/2
  - Cf:表が1/5で出る偏ったコイン　　p0=1/5
- コインCがどちらか，振って判定してみる
  - p1,p0である尤もらしさ(尤度)を比較
    - コインをn回振って，k回表が出る確率(二項分布)
    - もし{Ct|Cf}なら，どれくらいの確率で発生しうるか

$$ {}_nC_k\, p^k (1-p)^{n-k} $$

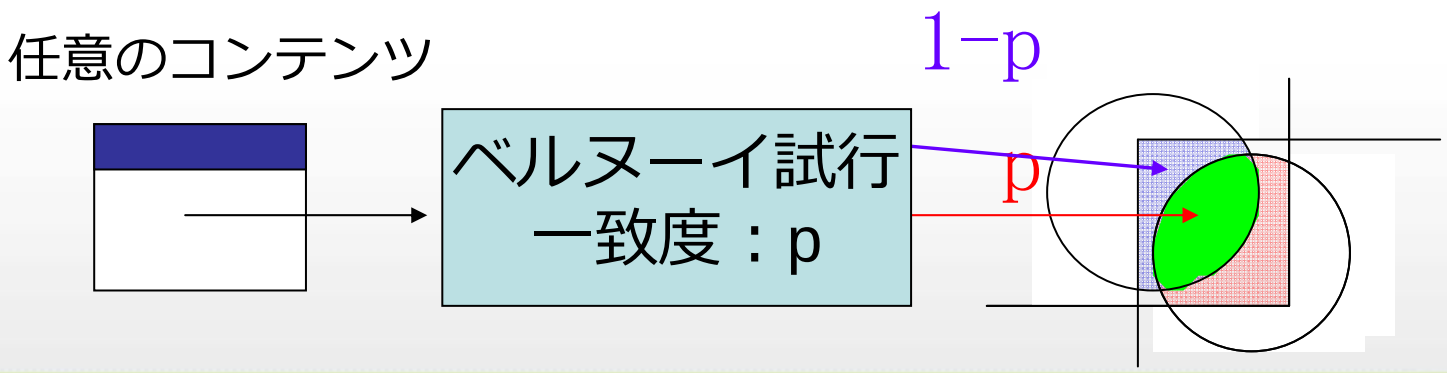
仮説検定の例:コイン判定 2/2

- コインCはどちらかである
  - Ct : p1=1/2 , Cf : p0=1/5
- コインCを振ってみた結果
  - 100回振ったら51回表
    - 尤度(p1)=0.0780，尤度(p0)=$3.97*10^{-12}$
      - あきらかに差があるので，強い確信度でCtと判定できる
  - 2回振ったら1回表
    - 尤度(p1)=0.500，　尤度(p0)=0.32
      - 非常に近い値なので判定しかねる
  - 尤度比較により信頼性を考慮できる

ベルヌーイ試行モデル

- 2コンテンツクラスタへの共起をモデル化
  - 一致度 p で一致，1-pで不一致
    - p ：両方のCCに含まれる.（緑部分）
    - 1-p：片方のCCにのみ含まれる（赤，青部分）
  - 上記試行のn回反復の結果は二項分布に従う
    - (n,p)が与えられたとき
      k回一致する尤度

$$ {}_nC_k\, p^k (1-p)^{n-k} $$



仮説検定

- 類似度計算を，以下の問題として捉える
  - 試行 n 回中 k 個一致から，pを測定すること

- 仮説検定：二つの一致度を設定
  - p0（意見不一致），p1（意見一致）（p0 < p1）
  - 各一致度のときの尤度を見比べて
    どちらに近いかを判定
  - 尤度　：　$_nC_k\, p^k (1-p)^{n-k}$

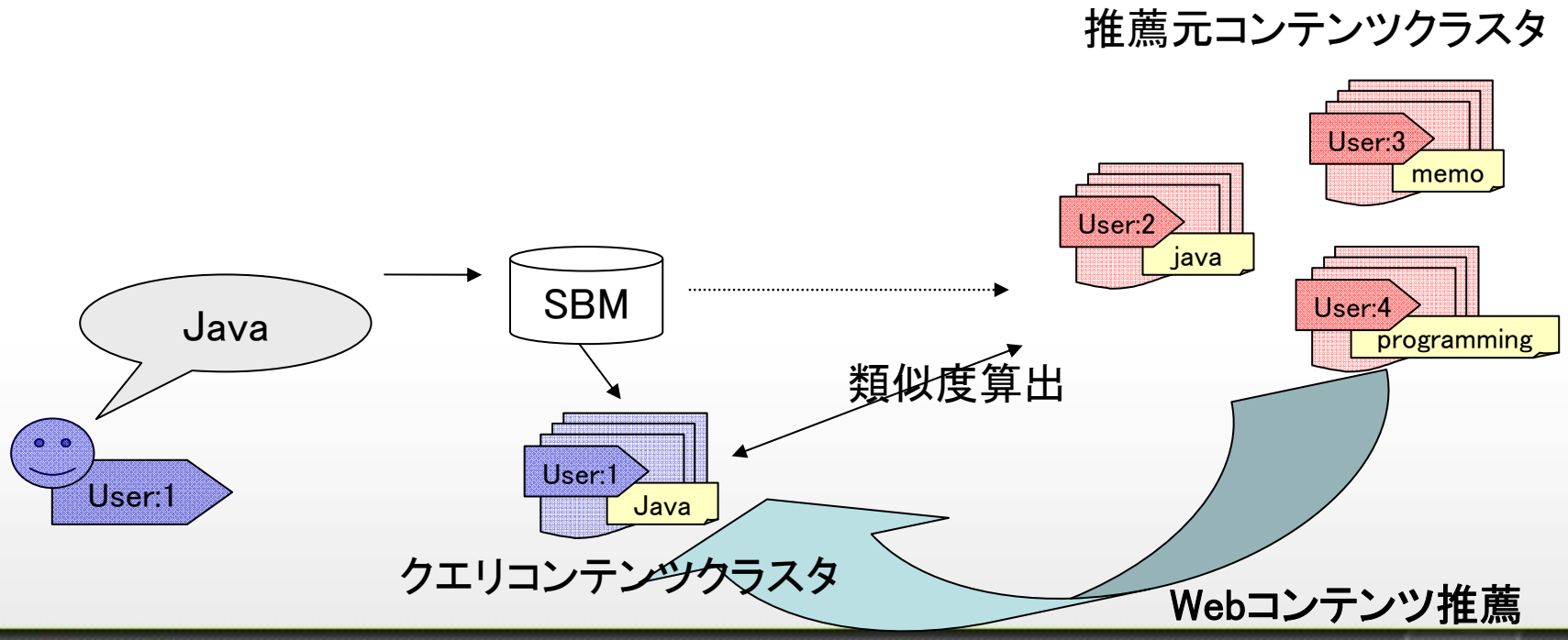
ここをp0,p1に入れ替えたとき値を比較

対数尤度比による類似度算出

- 2尤度の比較方法 → 対数尤度比
  - 式 ： $$\log\left(\frac{{}_nC_k\, p_1{}^k (1-p_1)^{n-k}}{{}_nC_k\, p_0{}^k (1-p_0)^{n-k}}\right)$$ =sim(s,o)
  - p1側か否か(判定不可を含む)の判定手法
  - 数値的な性質
    - (n, k)が大きいときのみ大きい値を取る
      - (n, k)=(2, 2)　Jaccard=1.00高 ， proposed=3.58低
      - (n, k)=(40, 5)　Jaccard=0.13低 ， proposed=−19.4低
      - (n, k)=(40, 35)　Jaccard=0.88高 ， proposed=58.7高
  - この値をそのまま類似度として利用

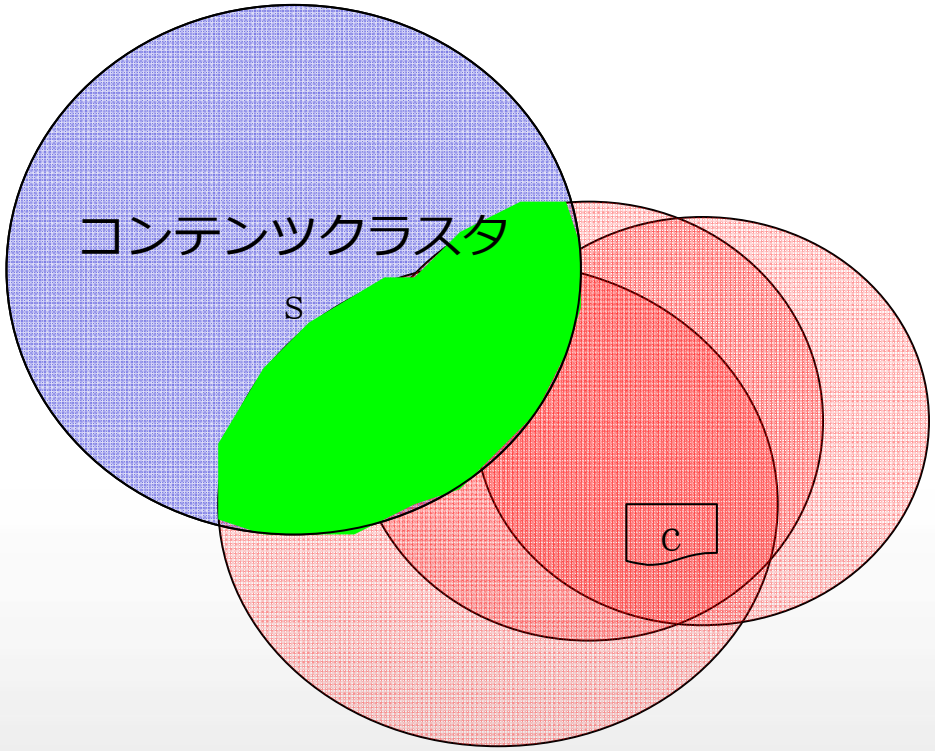
システム概要

1. クエリコンテンツクラスタをシステムに入力
   （ユーザID，タグ名が入力となる）
2. SBM中の複数の推薦元コンテンツクラスタとの類似度算出
3. **推薦元コンテンツクラスタからwebコンテンツを推薦**



webコンテンツの推薦度算出

- webコンテンツcが属する全てのコンテンツクラスタ間T_oの類似度の和で推薦度 R(s , c) を定義



$$R(s,c) = \sum_{o\{c \in T_o\}} sim(s,o)$$

実験

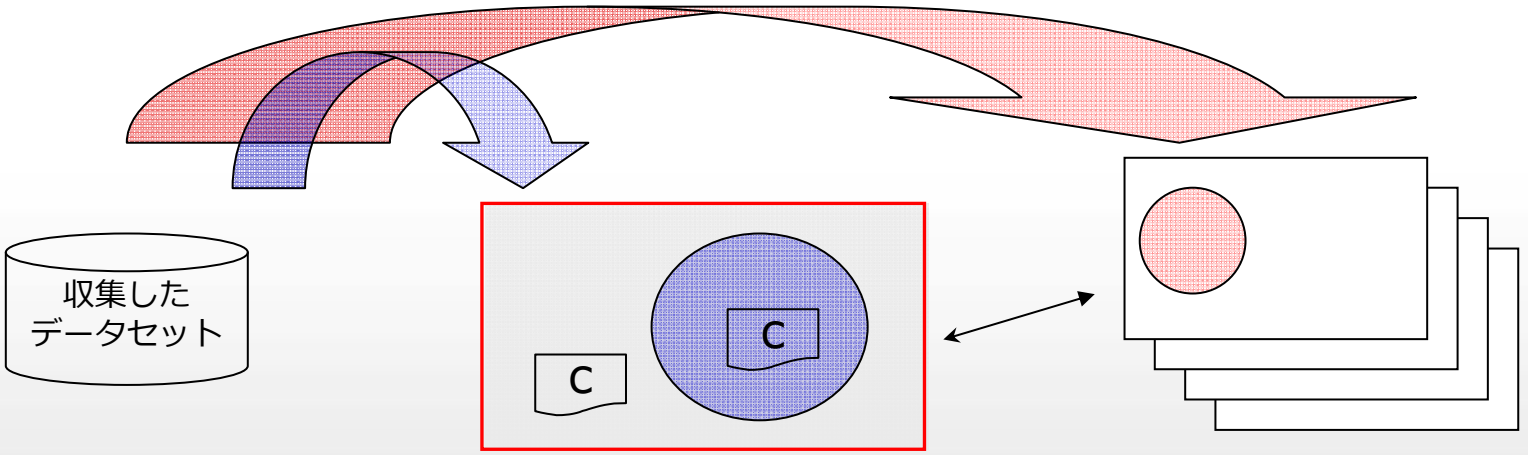
実験環境

- del.icio.us (http://del.icio.us/)
- ランダムサンプリングした1000人のデータを利用
  - コンテンツクラスタ総数：260,000
  - webコンテンツ総数　　：310,000

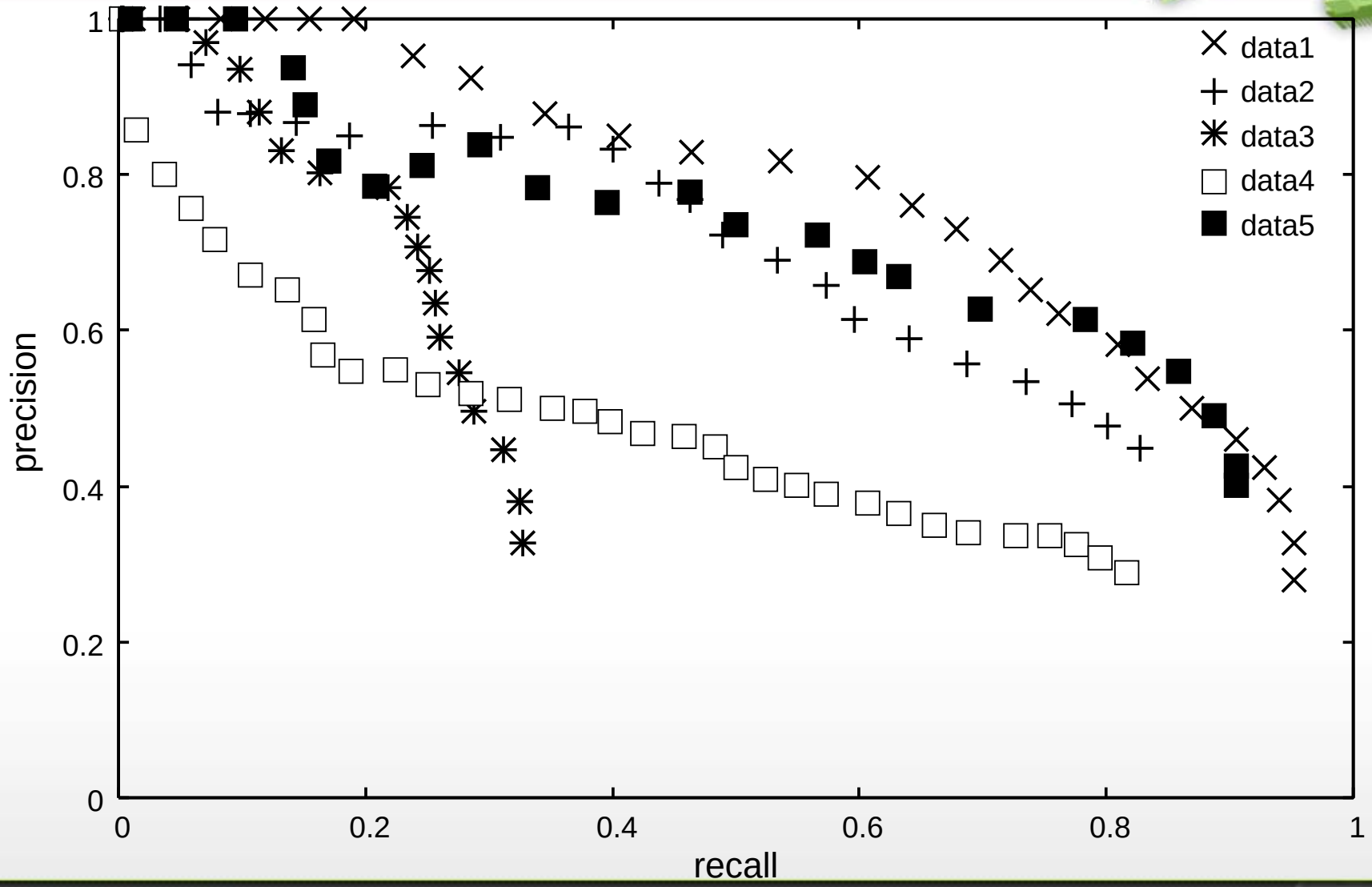
実験

実験方法

- 検証するコンテンツクラスタ(VC)を選択
- 類似度を計算後，VCのタグ・非タグを隠す
- VCのブックマーク内のコンテンツを数件推薦
- 実際のVC{タグ・非タグ}と比較
- Recall=漏れの少なさ,Precision=正確さ で評価



実験結果
（推薦不可コンテンツ除去後）



比較検証

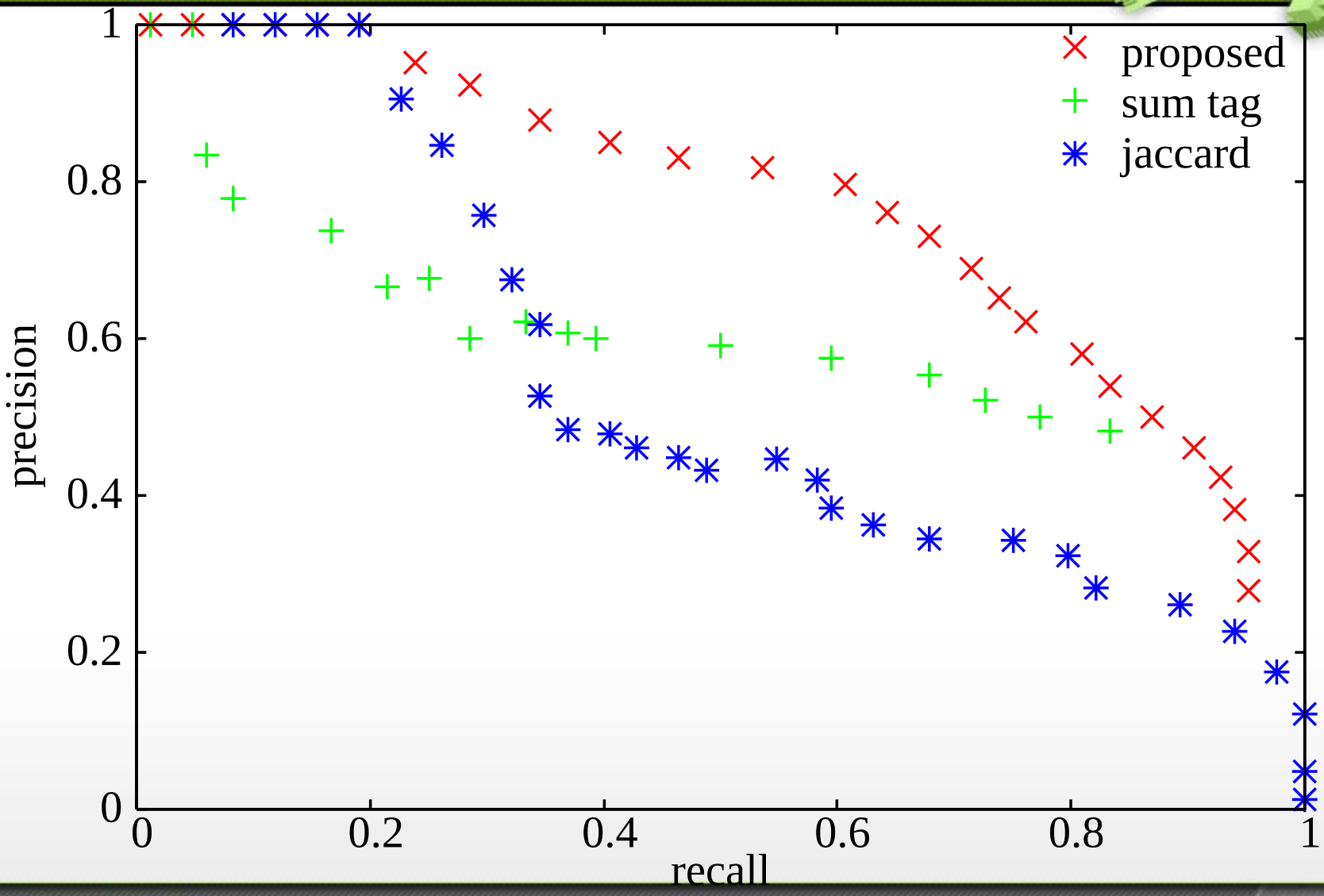
- 比較1　タグの名称のランキング
  - 同一のタグ名が多く付与されたwebコンテンツを推薦
  - 多くのユーザが同一のタグ名を付与したものを上位

タグの名称を利用しない推薦の有効性を検証

- 比較2　Jaccard係数による類似度算出
  - コンテンツクラスタ間の類似度をJaccard係数に置換
  - 他は提案手法と同様に推薦度を算出

仮説検定の有効性を検証

比較結果



まとめ

- 目的
  - SBMを利用した情報（webコンテンツ）推薦
  - タグ単位の推薦
- 新規性
  - タグ名称を用いない情報推薦
  - 統計的信頼性を考慮した協調フィルタリング
- 提案法の有効性検証
  - 高いRecall，Precision
  - 他の手法より優れた推薦精度

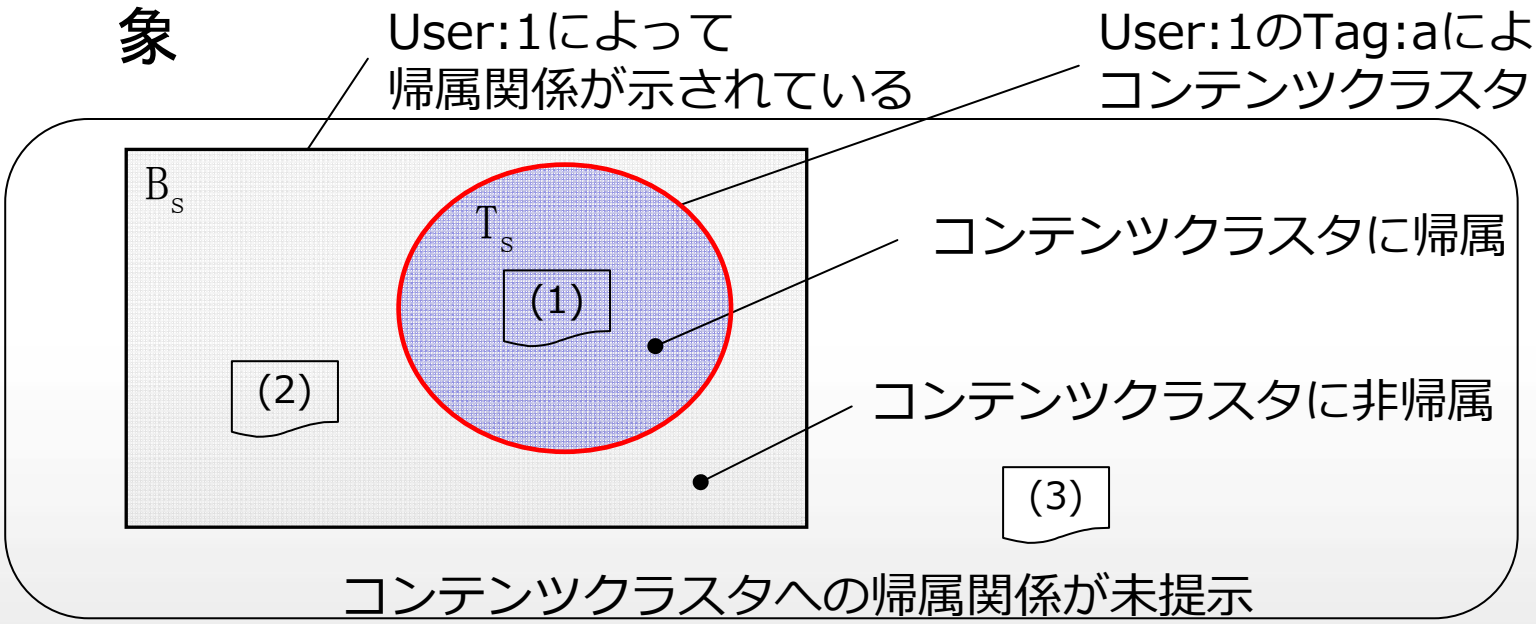
補助:コンテンツクラスタの概念

コンテンツクラスタとコンテンツ

(1)明示的にコンテンツクラスタに帰属

(2)明示的にコンテンツクラスタに非帰属

(3)コンテンツクラスタへの帰属関係が未提示　→　推薦対象



補助：Recall，Precision

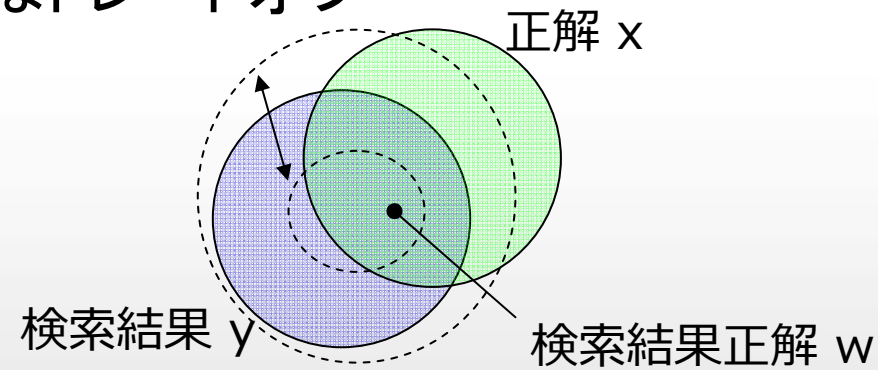
Recall（再現率）

- 推薦漏れの少なさの評価基準
- 検索結果正解数 w / 正解数 x

Precision（精度）

- 推薦間違いの少なさの評価基準
- 検索結果正解数 w / 検索結果数 y

一般的にRecall，Precisionはトレードオフ

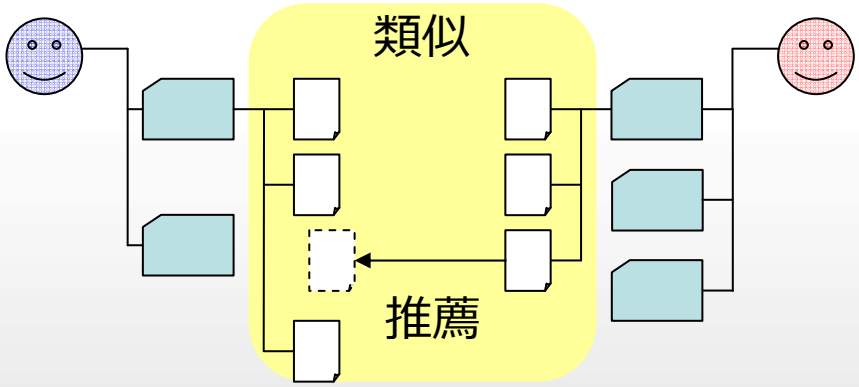


補助：Siteseer

webコンテンツの共起性のみを用いた推薦手法

- （ローカル）ブックマークのフォルダ構造に着目

- 具体的な計算方法が未提示
- web コンテンツの推薦順位を決定できない
  （本研究ではwebコンテンツ単位で推薦度算出）



補助：仮説検定について

二つのコンテンツクラスタ間の関係

- 類似関係にある　→　p = 0.6 (高確率)
- 類似関係でない　→　p = 0.1 (低確率)

二点で尤度比検定を行うメリット

- 計算しやすい
- 類似度の和が和集合の類似度と一致する (次ページ参照)
  - 推薦度の算出方法の理論的裏付け

補助：類似度の和について

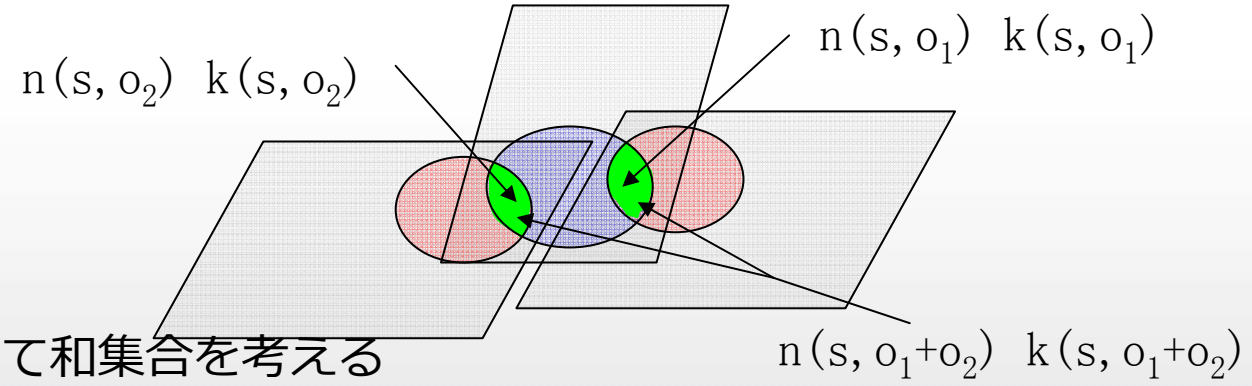
推薦元コンテンツクラスタの和集合を考える

- $n(s, o_1+o_2) = n(s, o_1) + n(s, o_2)$
- $k(s, o_1+o_2) = k(s, o_1) + k(s, o_2)$

類似度の和

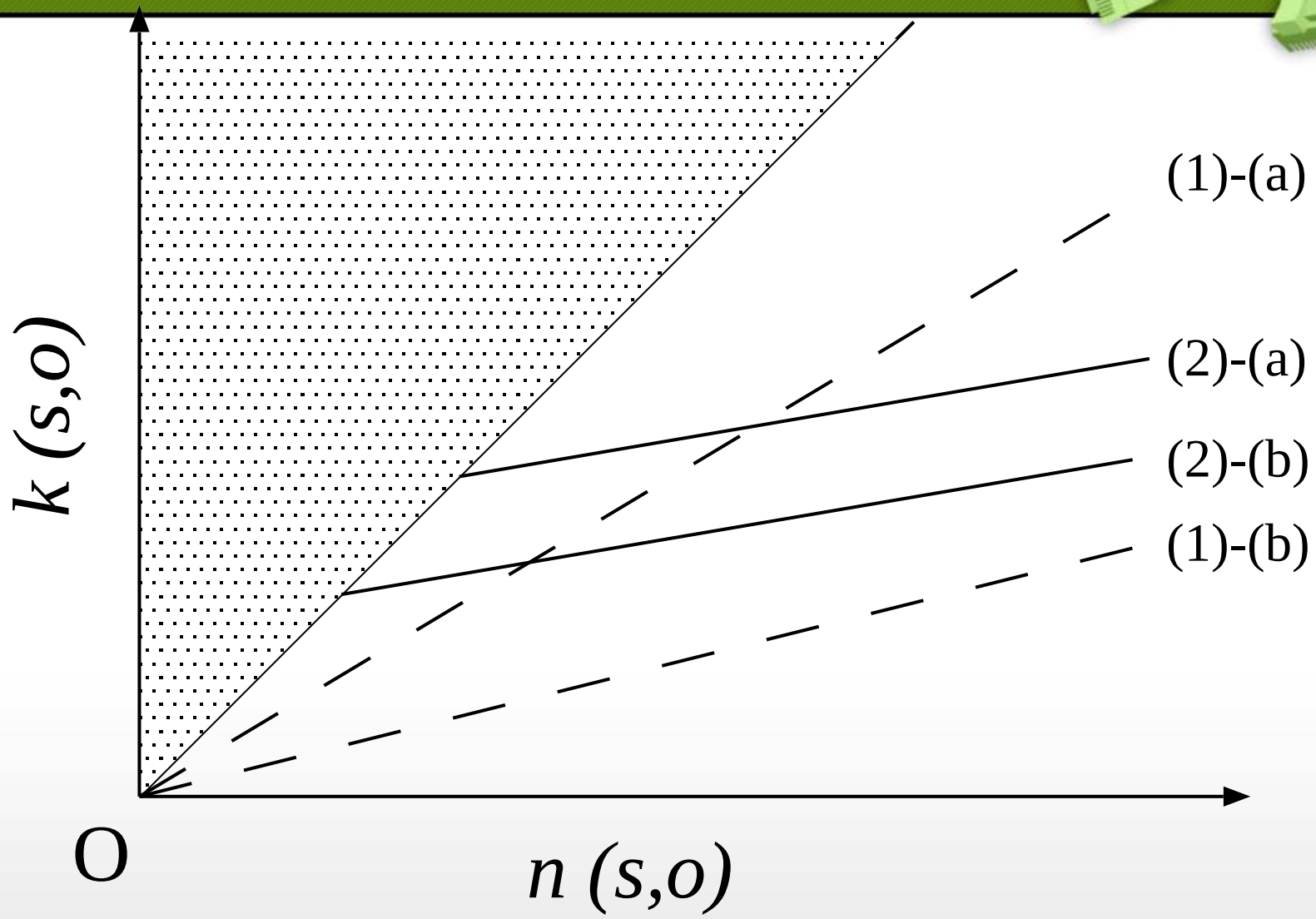
$$sim(s, o_1+o_2) = sim(s, o_1) + sim(s, o_2)$$

→ webコンテンツの推薦度を定義できる



重複を許して和集合を考える

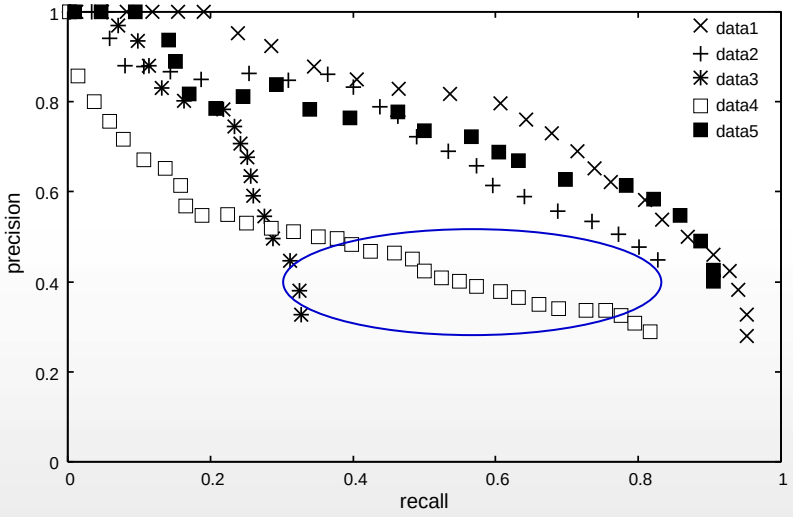
補助：仮説検定とJaccard

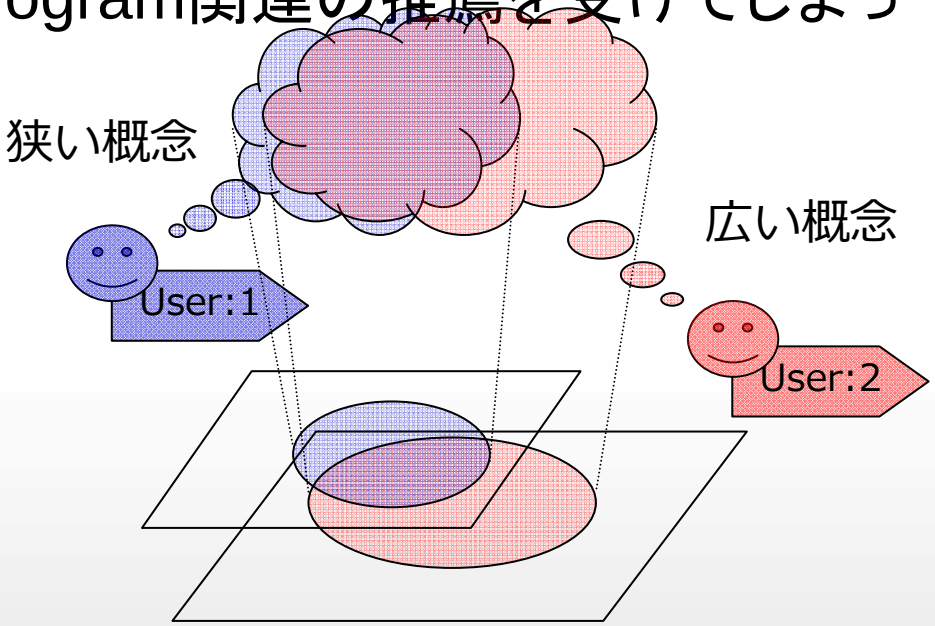


補助：precisionが低くなる理由

- 推薦先コンテンツクラスタのタグ名は「javascript」
- 推薦元コンテンツクラスタのタグ名は「programming」等

→　javascript以外のprogram関連の推薦を受けてしまった

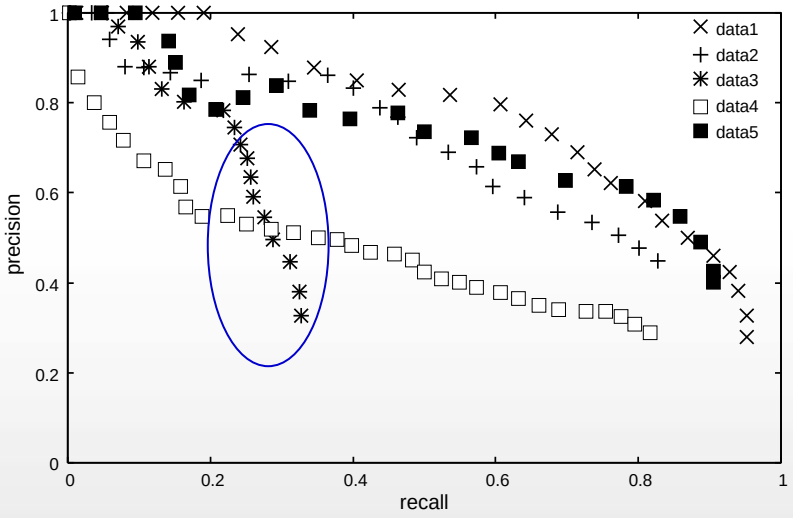


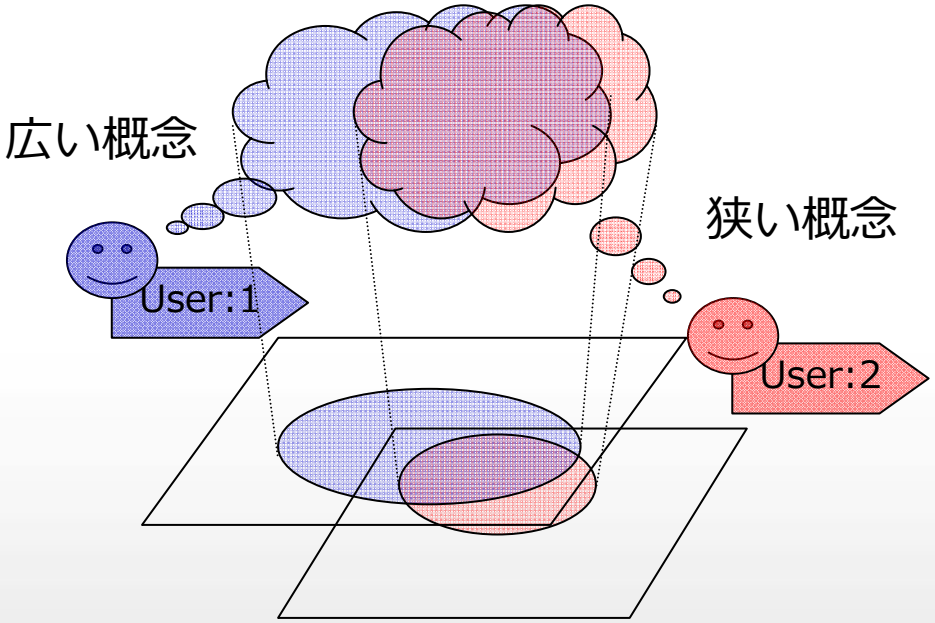


補助：recallが低くなる理由

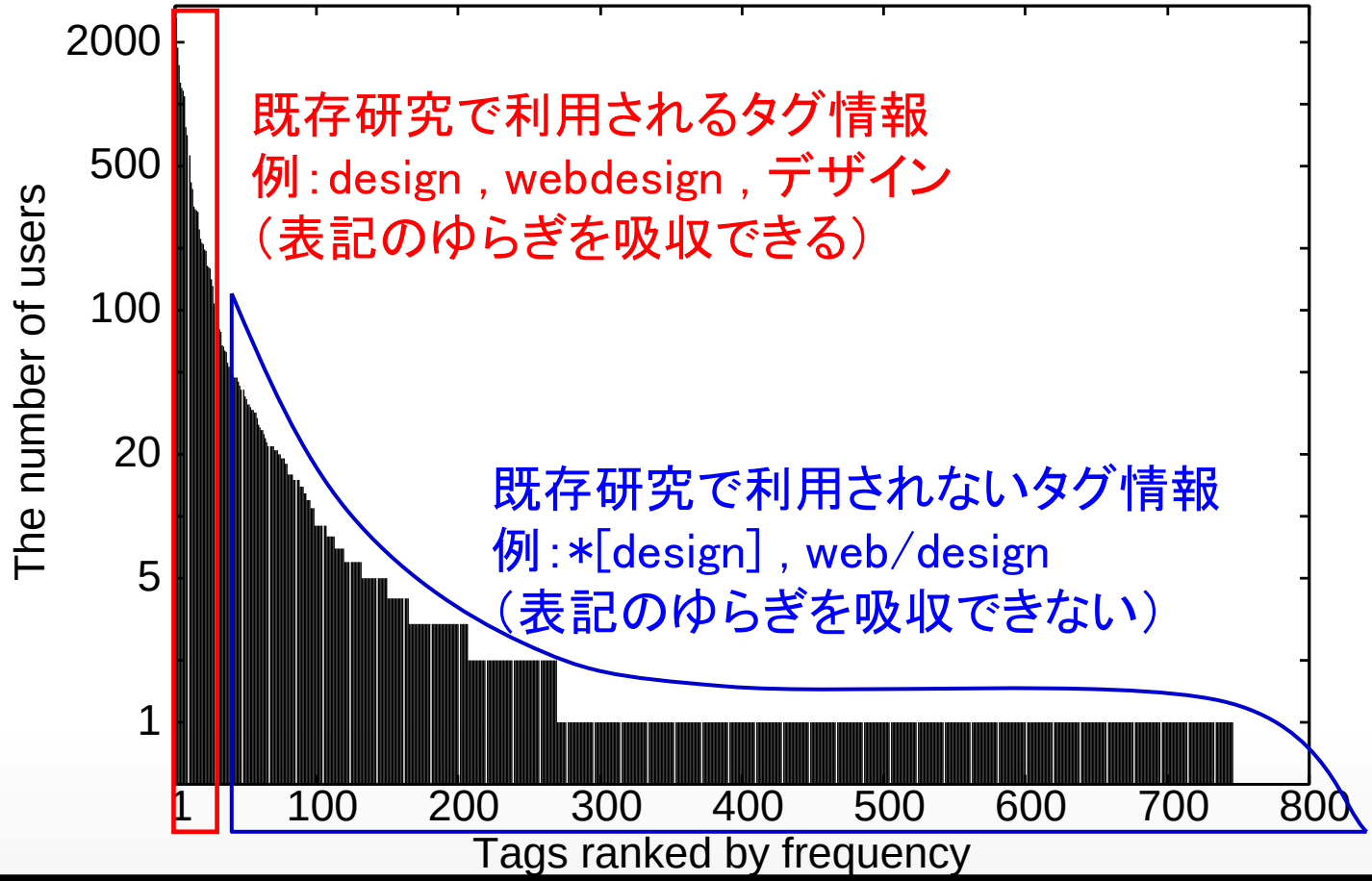
- 推薦先コンテンツクラスタのタグ名は「art」
- 推薦元コンテンツクラスタのタグ名は「webdesign」等

→　デザイン系以外のart関連の推薦を受けられなかった





補助:タグ名の頻度分布もロングテール



少数意見のタグが大部分を占めている

類似度いろいろ

- Pearson correlation coefficient
  - 共分散をそれぞれの標準偏差で割っただけ

$$S_P = \frac{\sum (x_{ik} - \bar{x}_i)(x_{jk} - \bar{x}_j)}{\sqrt{\sum (x_{ik} - \bar{x}_i)^2 \sum (x_{jk} - \bar{x}_j)^2}} \quad (6)$$

- Cosine coefficient
  - ベクトルの角度と考えるとわかりやすい

$$S_C = \frac{\sum (x_{ik} x_{jk})}{\sqrt{\sum (x_{ik})^2 \sum (x_{jk})^2}} \quad (5)$$

余談:これも問題(!)

- 「みんなでタグ付け」という概念に対してつけられたキーワードがいっぱい
  - Social tagging
  - Social classification
  - Collaborative filtering
- 「ソーシャルブックマーク」に限ってみても2つ
  - Social bookmark
  - Social bookmarking

類似研究を検索するときの罠
→tagの統合ってやっぱり難しい

何が問題か？

- SBMデータの分析：itemに”正しいtag”をつける分類が主流→推薦もそちらに引きずられる
- 推薦という目的を考えたときは，コンテンツの分類は本当に必要なのか？
- なぜcosine係数なのか
  - データが少ないと精度が下がる
    - ちょっとした補正項を入れている例も多い
  - あと，そもそもt,u,iは直交じゃないのでは？
    - BiasのキャンセルSVDやTF-IDF