May 22, 2019

Folksonomy マイニングに基づく Web ページ推薦システム

沼田　賢一

富山県立大学　情報基盤工学講座

1. はじめに
2. Folksonom とは y
3. ソーシャルブックマークとは
4. システム概要
5. 実験 A,B
6. 考察

はじめに

目的

従来の Web ページ推薦システムの多くは協調性フィルタリンを用いているが, 共通のページにアクセスしていないと類似度が 0 とみなされたり, 同一ページを評価しているユーザが少ないため Web ページの嗜好をどのように抽象化し表現するかという問題があった.

方法

ソーシャルブックマーク上のデータをユーザの Web 嗜好データとして利用し,Web ページを Folksonomy によって与えられるタグ情報でマイニングすることで, ユーザの Web ページ嗜好の抽象化表現と Folksonomy のタグ表記のゆれを解決する.

TOYAMA Prefectural University

Folksonomy

Folksonomy

多数のエンドユーザが各々の Web ページに対しタグ（キーワード）を付与することで, 様々な分類処理を行うことができる. 従来のあらかじめ分類木に基づいて Web ページを分類していく Taxonomy と呼ばれる方法に対して, ユーザの実際の興味や利便性をリアルタイムに反映した分類情報を構築できる.

Folksonomy の問題点

ユーザのタグ付け方法に制限を行わないため, タグ表記のゆれ問題が起こる. これは, 同じ意味なのに表記の異なるタグが氾濫してしまうこと.

TOYAMA Prefectural University

ソーシャルブックマーク

ソーシャルブックマーク

一つの Web サイト上で複数ユーザのブックマーク情報を共有するサービス. ブックマークを登録するときに,Folksonomy に基づいてページに自由にタグ付けすることができ, タグによって複数ユーザのブックマーク情報が関連付けされる.

システム概要

解決方法

類似度 0 の問題を防ぐため, ユーザと各タグとの親和度でユーザの嗜好を表現する. また, タグ表記のゆれを防ぐ方法については, 類似タグをクラスタリングすることでユーザと各タグクラスタとの親和度を計算して, ユーザの嗜好表現を抽象化させる.

システム概要

ソーシャルブックマーク上の公開されているブックマークデータの収集し, 各ユーザと各タグの親和度, 各タグ間の類似度を求める. 次に, タグ間の類似度からタグをクラスタ化し, 求めたユーザとタグ間の親和度をもとに, 各ユーザと各タグクラスタの親和度を求める. タブクラスタごとに推薦ページ群を計算し, 求めたユーザ・タグクラスタの親和度をもとに各ユーザに対する推薦ページ群を計算する.

システム概要図

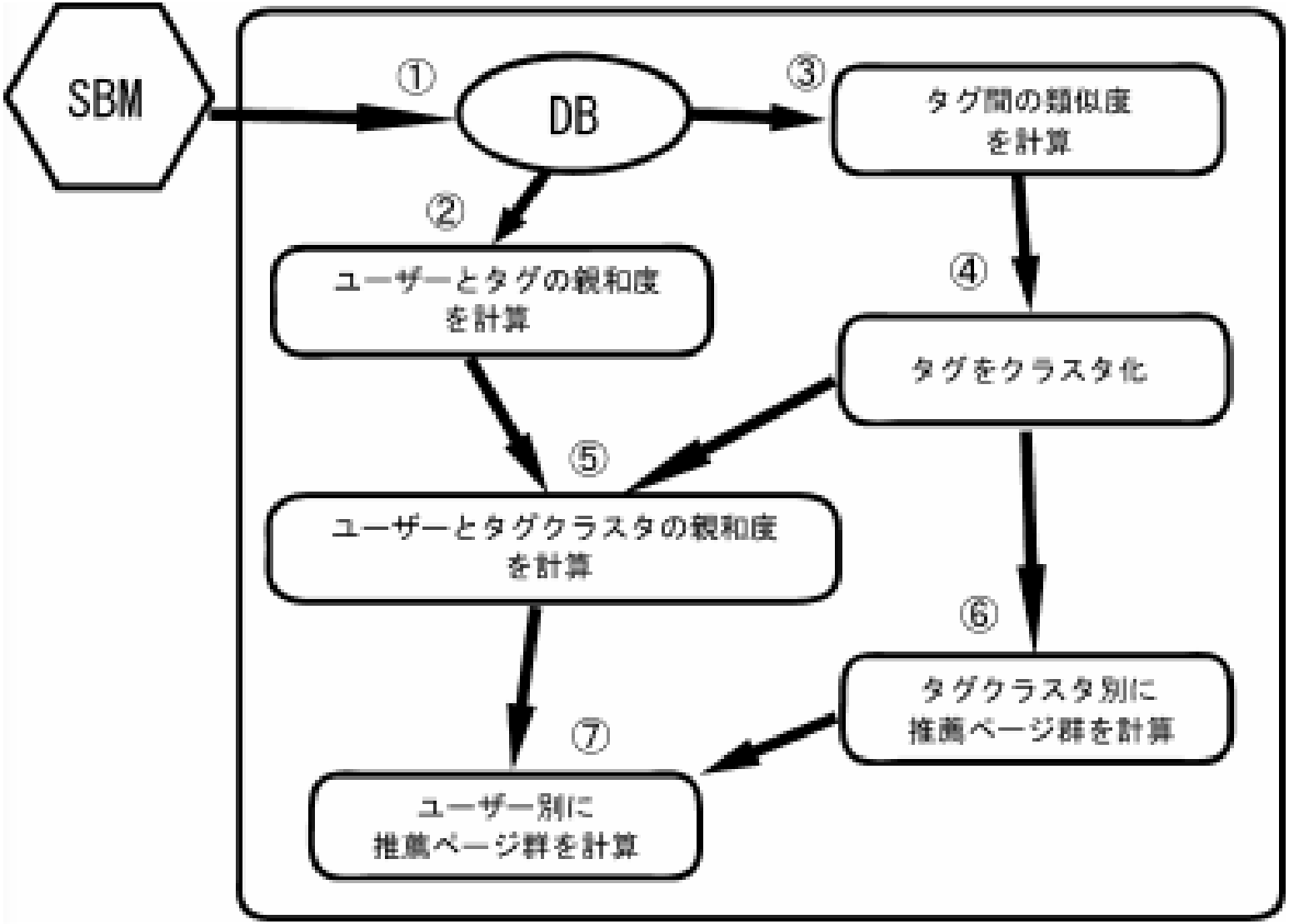

図 2　全体の処理の流れ

ユーザとシステム, タグの関係図

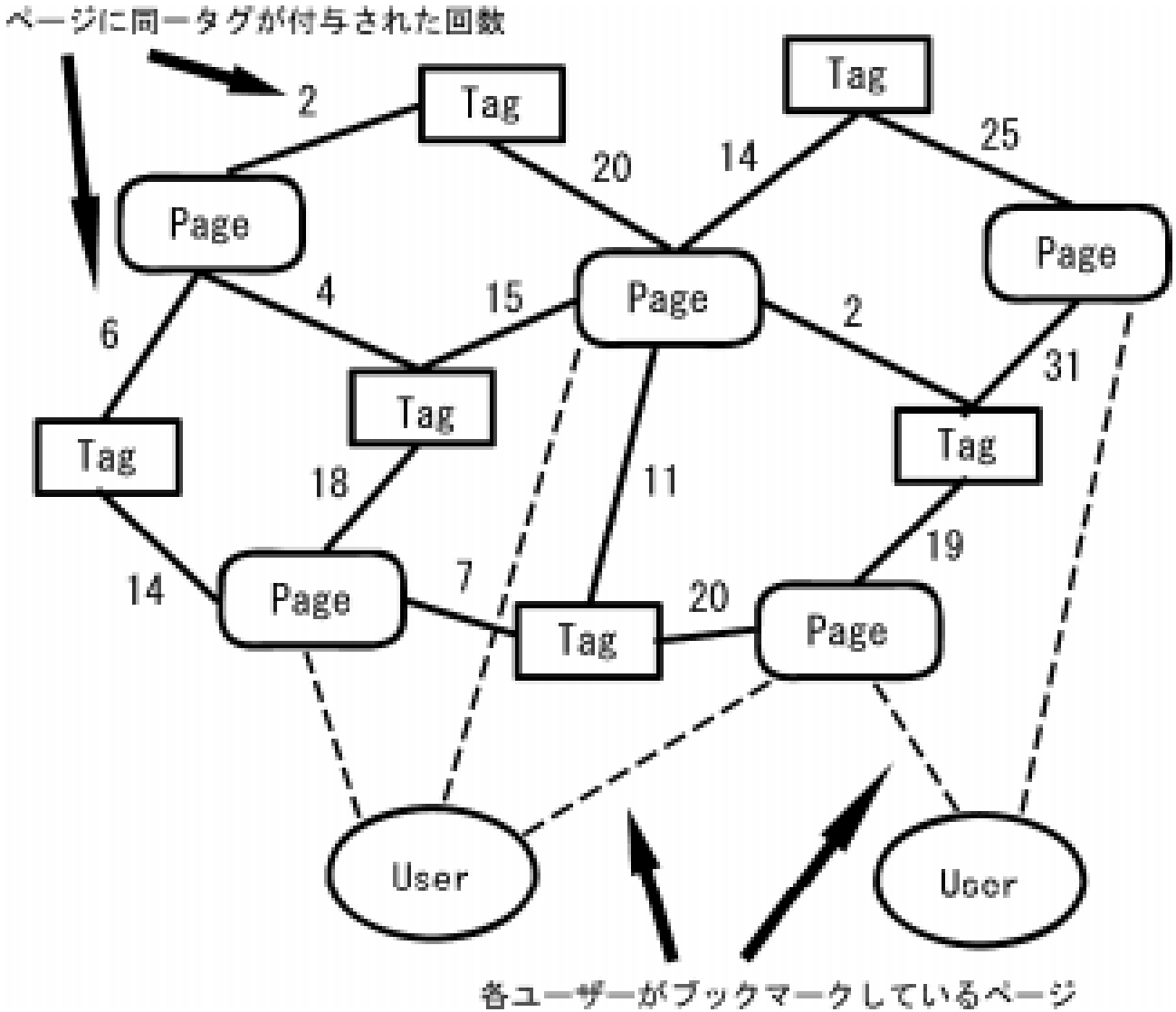

図 3　SBM の内部データのモデル

ユーザとタグの親和度の計算

ユーザとタグの親和度の計算

w(P,T):ページ P とタグ T の多重度 bookmark(A):ユーザ A のブックマークしているページ群
rel(A,T):ユーザ A とタグ T の親和度
TF(P,T):ページ P に関連付けられているすべてのタグに対してタグ T が占める割合
IDF(T):全体のページにおけるタグ T の希少性

$$rel(P,T) = TF(P,T) \times IDF(T) \quad (1)$$

$$TF(P,T) = \frac{w(P,T)}{\sum_{T_i \in TAGS} w(P,T_i)} \quad (2)$$

$$IDF(T) = \log \frac{\sum_{P_j \in PAGES} \sum_{T_i \in TAGS} w(P_j,T_i)}{\sum_{P_j \in PAGES} w(P_j,T)} \quad (3)$$

$$rel(A,T) = \sum_{P_i \in bookmark(A)} rel(P_i,T) \quad (4)$$

タグ間の類似度の計算

タグ間の類似度の計算

rel(T_1, T_2):タグ T_1 にとってのタグ T_2 の類似度

$$rel(T_1, T_2) = \sum_{P_i \in PAGES} w(P_i, T_1) \times rel(P_i, T_2) \quad (5)$$

タグのクラスタ化

タグのクラスタ化

タグ T に対して rel(T,T_i) の値が最も大きい T_i を選び,rel(T,T_i) が閾値 V_{limit} を超えていたら T の親タグとし, 超えていない場合は T 自身を T の親タグとする.
閾値 V_{limit} やクラスタサイズ C_{max} は変化させることで, 最終的に生成されるクラスタの粒度を調整できる.

ユーザとタグクラスタの親和度の計算

ユーザとタグクラスタの親和度の計算

rel(A,C):ユーザとタグクラスタ C の親和度

$$rel(A,C) = \sum_{T_i \in C} rel(A,T_i) \tag{6}$$

推薦ページの計算

タグクラスタ別の推薦ページの計算

point(C,P):クラスタ C から各ページ P に付与される推薦ポイント

$$point(C,P) = \sum_{T_i \in C} w(P,T_i) \qquad (7)$$

このポイントが高いページから順に, クラスタ C の推薦ページ群とする.

ユーザ別の推薦ページの計算

point(A,P):ユーザ A からページ P に付与される推薦ポイント

$$point(A,P) = \sum_{C_i \in CLASTERS} rel(A,C_i) \times point(C_i,P) \qquad (8)$$

このポイントが高いページから順に, ユーザ A に対する推薦ページ群とする.

実験

実験概要

実験対象のソーシャルブックマーク:はてなブックマーク
ユーザデータを約 5800 人分収集.
そのうち 5000 人を訓練ユーザとし, 訓練ユーザによってブックマークされたページデータ 57000 ページ分を収集.
またそれらのページに付与されたタグ約 9300 個分のデータを収集.(利用するのは合計 6 回以上出現したタグのみで 2175 個)
クラスタリングの粒度は 5 段階で行う.
実験は,SBM の大量のデータをもとに推薦精度を測定する実験 (A) と, ユーザに実際に推薦ページを評価してもらう実験 (B) の 2 種類行う.

実験

タグ表記のゆれの解決

	A	B	C	D	E
全クラスタ数	2175	1430	877	512	412
平均クラスタサイズ	1.00	1.52	2.48	4.25	5.28
Step(7)における1ユーザーあたりの平均処理時間(秒)	5.4	2.8	2.2	1.8	1.6
"english"クラスタのサイズ	1	3	5	7	10
"java"クラスタのサイズ	1	11	15	30	34
"blog"クラスタのサイズ	1	28	51	64	70

図 6　各クラスタリングケースの比較

タグ表記のゆれの解決

ケース B,C において図のように同義語クラスタが数多く見られた. 英語表記, 日本語表記, 略語など多様な表記のタグがうまくクラスタリングされているので, タグ表記のゆれの問題を解決できている.

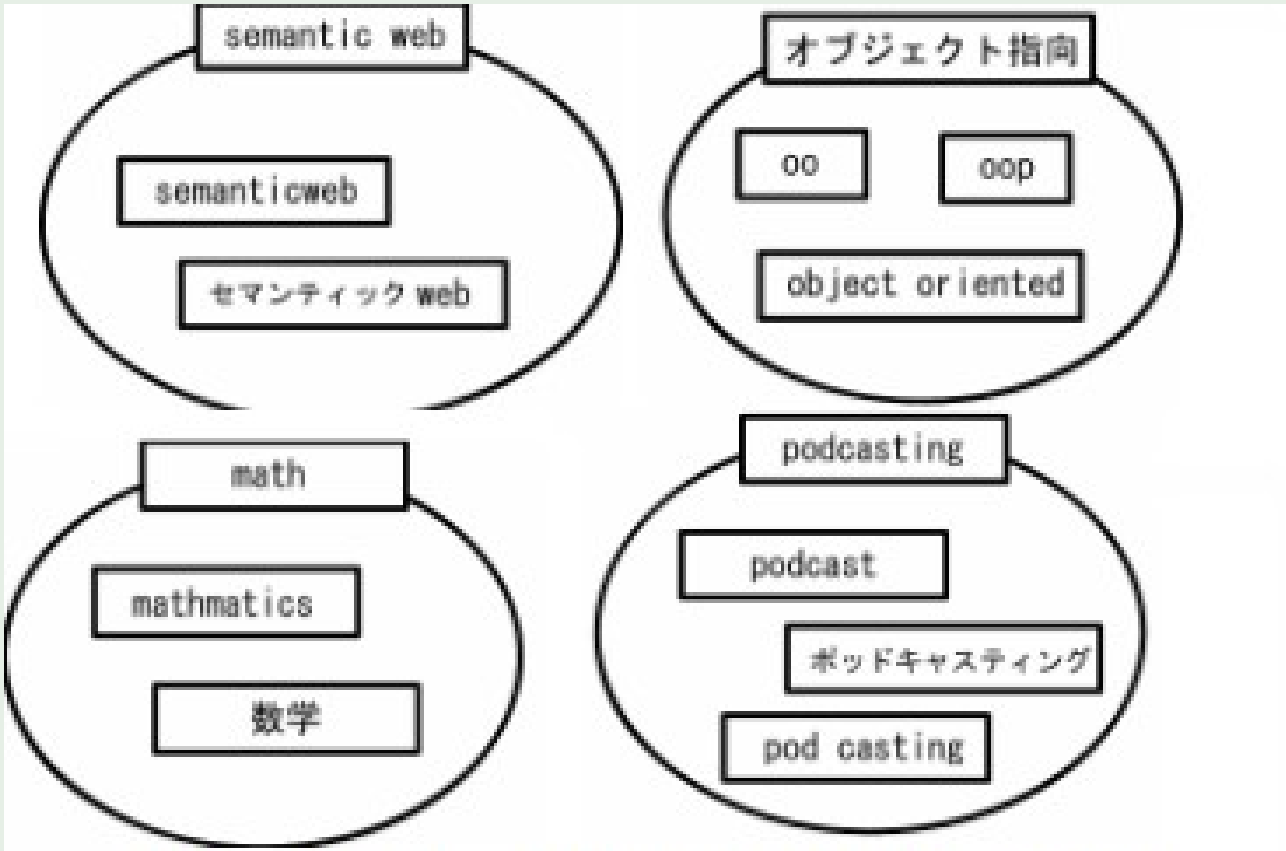

図 8　ケース B , C に見られた同義語クラスタの例

実験 A

実験 A の目的

クラスタリング粒度とシステムの推薦精度の関係を調べること.

実験 A のやりかた

ユーザを訓練ユーザとテストユーザの 2 種類に分け, 訓練ユーザに関するデータでページ推薦システムを構築する.
テストユーザのブックマークページ群をクエリ用ページ群とテスト用ページ群に分けてシステムにクエリを投げ, 帰ってきた推薦ページ群とユーザのテスト用ページ群との一致度を比較して評価する.
一致度の計算は, 再現率と適合率の 2 種類の評価指標を用いる.
R:推薦ページ数
T:テストページ数
H:一致したページ数

$$\text{再現率} = H / T \qquad (9)$$

$$\text{適合率} = H / R \qquad (10)$$

実験 A の結果

各クラスタリングごとの再現率の比較

1 ユーザのブックマークページ数が再現率や適合率に大きな影響を与えるため, ユーザをブックマークページ数の大きさに応じてクラス分けし, クラスごとに集計を行っている.

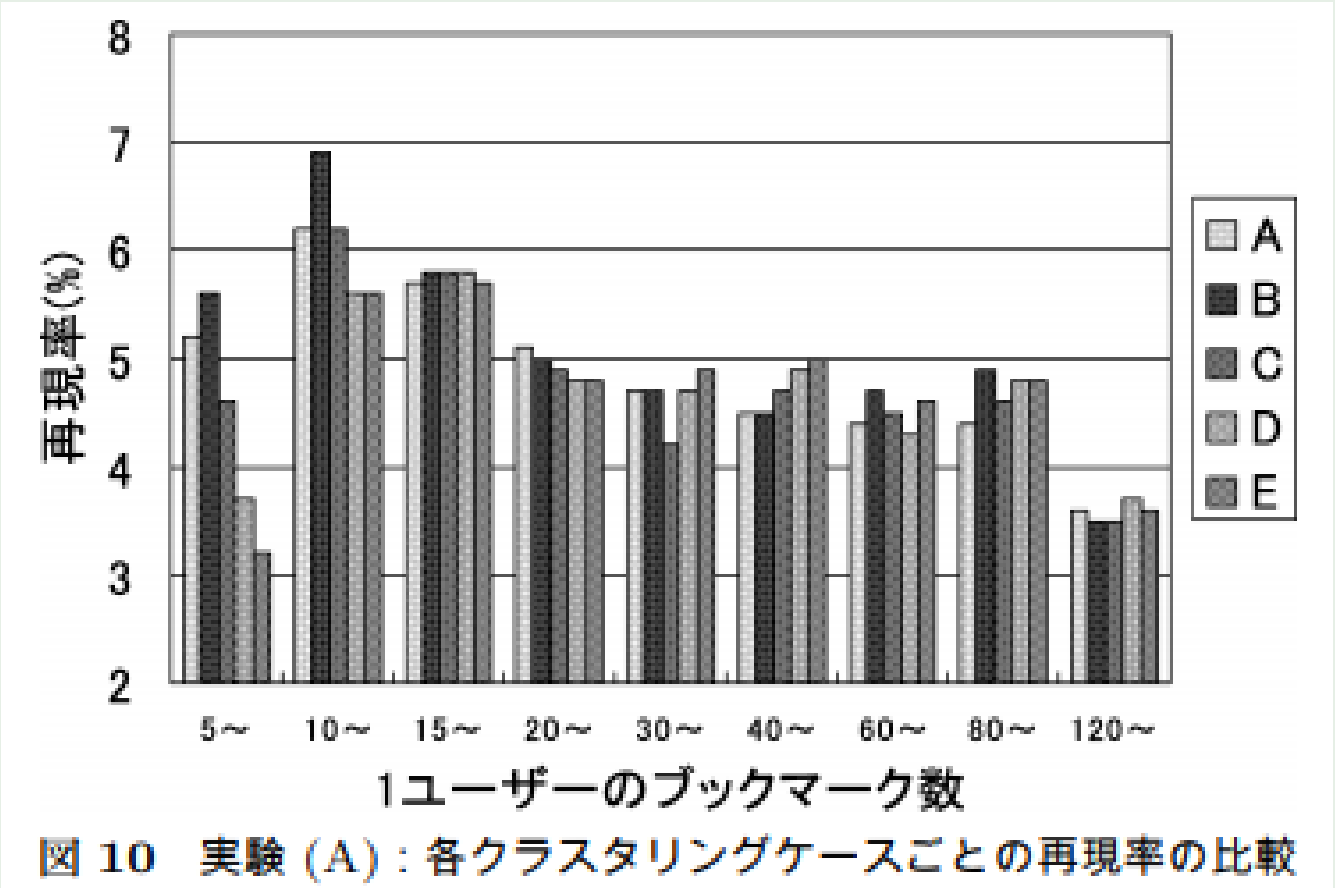

図 10　実験 (A) : 各クラスタリングケースごとの再現率の比較

実験Aの結果

各クラスタリングごとの適合率の比較

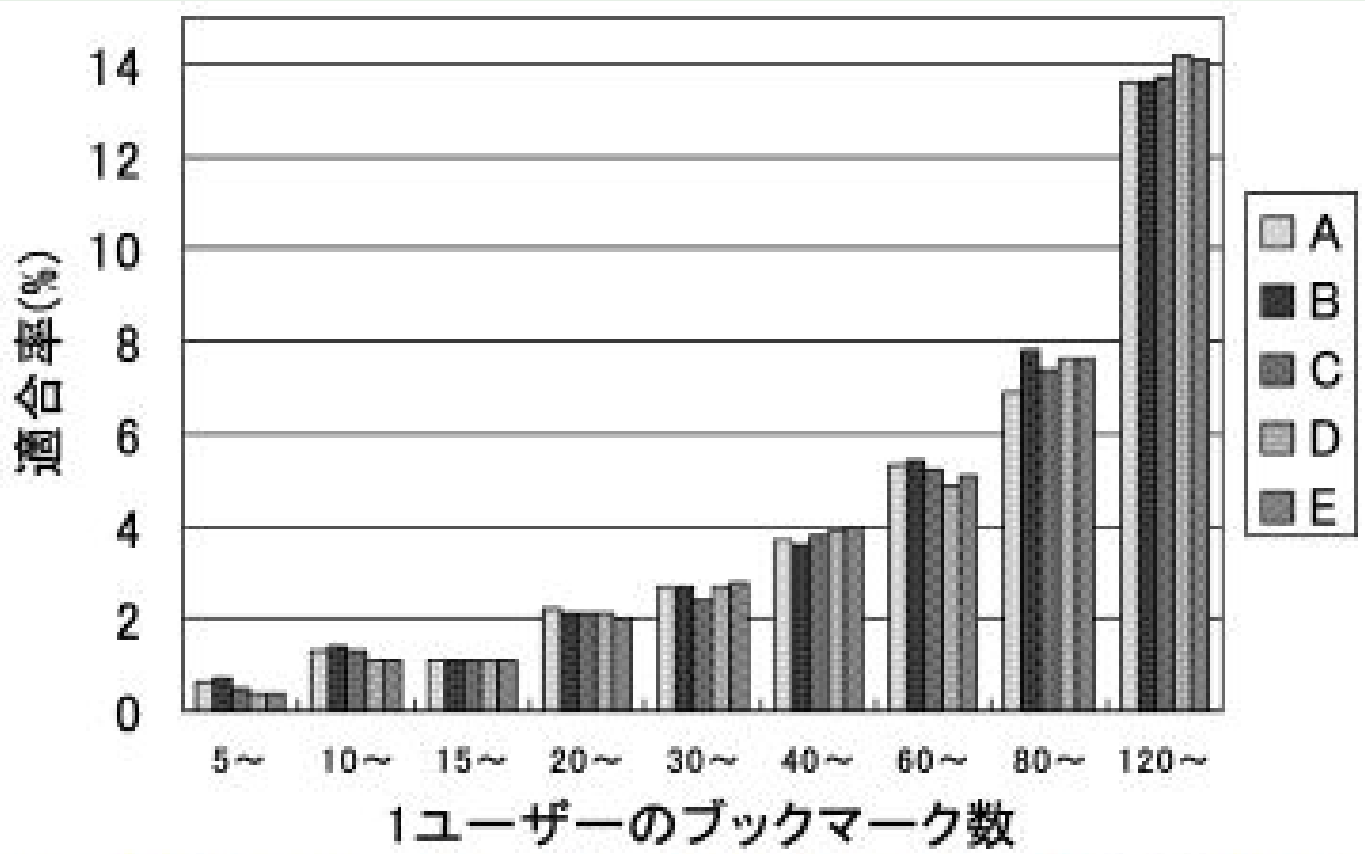

図 11　実験 (A)：各クラスタリングケースごとの適合率の比較

実験 A の結果

実験 A の考察

推薦制度は全体的にケース B が最も優れていた. ケース C,D,E の推薦制度があまり上がっていない原因として, 話題を抽象化しすぎたために個人の細かい嗜好が反映されず, 一般的に人気のあるページばかり推薦されてしまうという問題が起こっていると考えられる.

実験 B

実験 B の目的

ユーザが実際に興味を持つ Web ページが推薦されたかを調べること.

実験 B のやりかた

実験 A で相対的に最もいい結果を出したもの（クラスタリング B）を用いる.
ユーザは SBM の会員ではない 10 人のブラウザのブックマークを用いる.
このデータをインプットとして出力にユーザの最も親和度が高いタグと推薦ページ上位 30 個を表示する.
これに対して, ユーザは 3 段階の主観評価をする.
(a) とても興味があり, 内容が自分と関係が深い.
(b) 興味はあるが, 特に自分と関係が深いわけではなく, 万人向けの内容である.
(c) 興味はない. 内容も自分とは関係ない.

実験 B の結果

実験 B の結果

じっけん A と同様に 1 ユーザのブックマークページ数の大小が適合率に大きな影響を与えるため, ユーザのクラス分けを行っている.
ユーザ群 (1):ブックマーク数が 50 ページ未満のユーザ
ユーザ群 (2):ブックマーク数が 50 ページ以上のユーザ
(a) 適合率:主観評価 (a) が選択された割合
(a)(b) 適合率:主観評価 (a) または (b) が選択された割合

実験Bの結果

推薦タグの適合率の比較

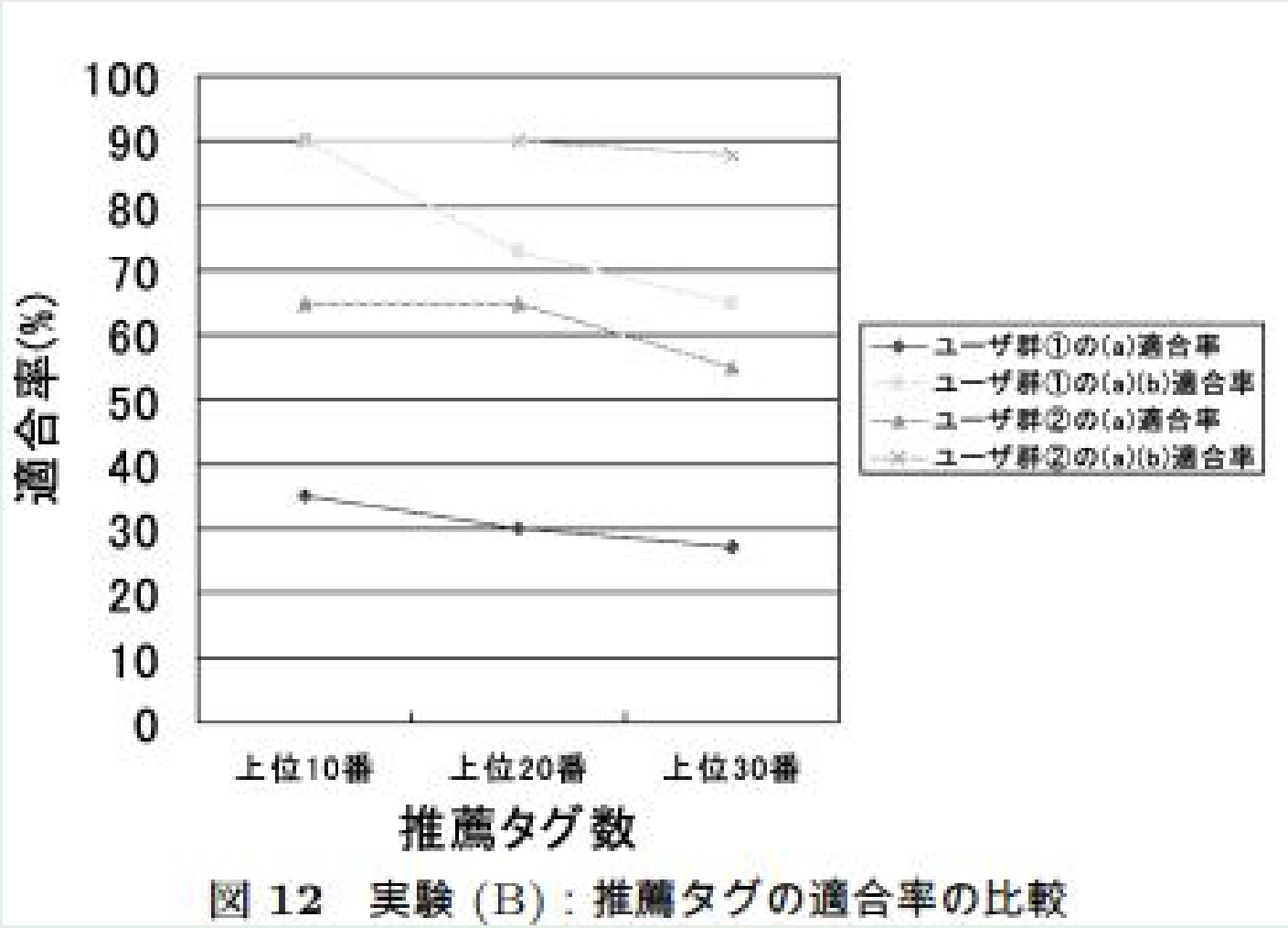

図 12　実験 (B)：推薦タグの適合率の比較

実験Bの結果

推薦ページの適合率の比較

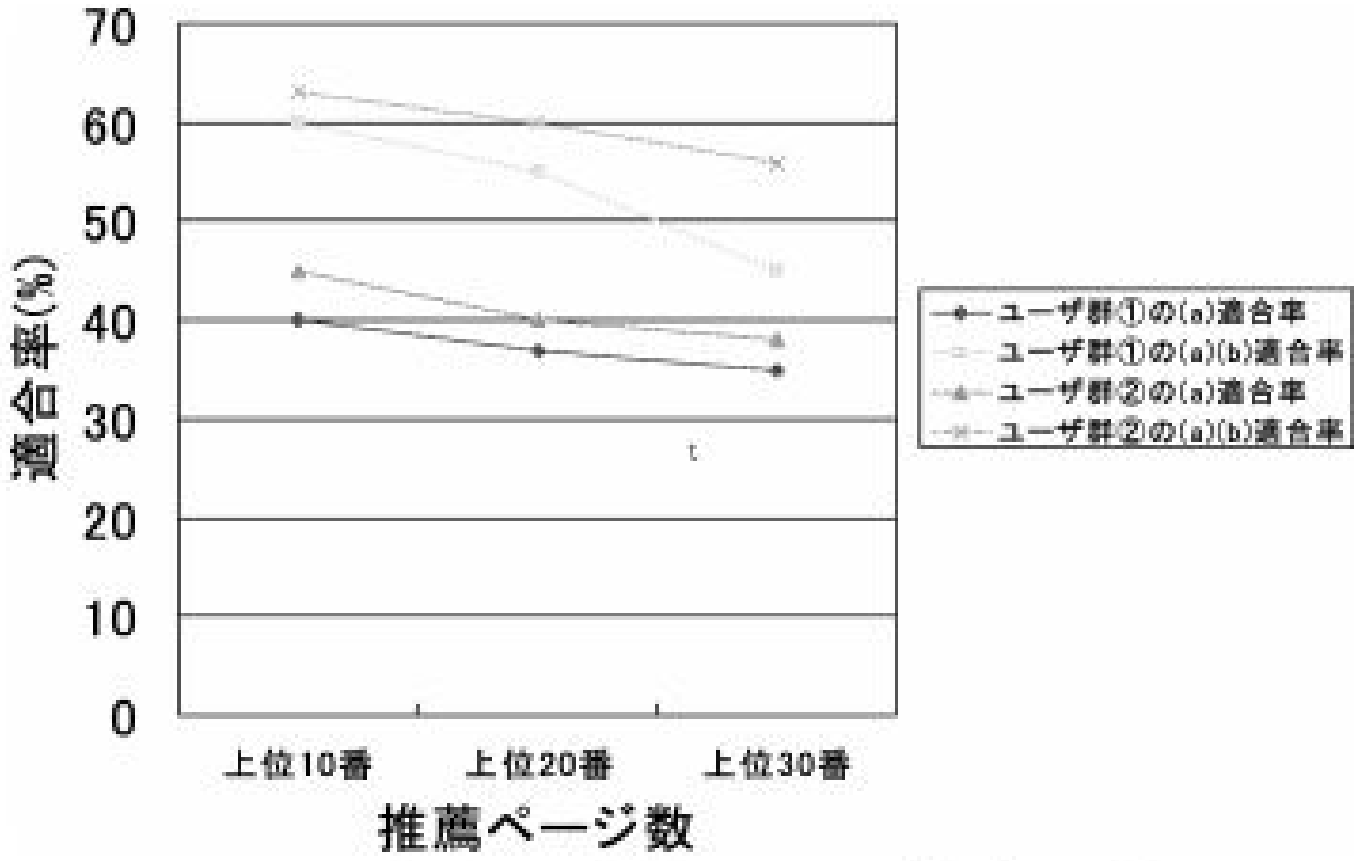

図 13　実験 (B)：推薦ページの適合率の比較

実験 B の結果

実験 B の考察

ブックマーク数の多いユーザのほうが適合率が高い.(a) 適合率で平均 40 %,(a)(b) 適合率の平均 60 %というのがこの Web ページ推薦システムの総合評価となった.

TOYAMA Prefectural University

おわりに

まとめ

1. TF・IDF を用いているので汎用的なページ (検索サイトなど) に低い重みがつくのでユーザの興味があるページが推薦されやすい.
2. 計算量的な観点での実現可能性については, ソーシャルブックマークの枠組みを超えて数億から数十億のページを扱うとしても, システムのアルゴリズムは各ノードからリンクを 1,2 段たどる仕組みなので計算量は $0(n^2)$ 程度. また, すべてのフェーズは分散することできるので計算時間を短縮させることができる.
3. タグのクラスタリングによってタグ表記のゆれ問題を解決できた.
4. 実験 B の推薦精度の結果 40〜60 %はインターネット全体を対象と既存の Web 推薦システムと比較しても遜色がない値.
5. 従来の協調フィルタリングを用いたシステムと比べて, 推薦対象を特定のサイト群に限定しないシステムができた.