

動画サイトのコメントを用いたタグ推定に関する研究

坂地泰紀 (東京大学), 小林暁雄 (理研AIP), 小花聖輝 (茨城大学)

使用データ: 「ニコニコデータセット」

1. ニコニコ動画



ユーザ数	39,360,000↑
動画数	12,266,299↑
コメント数	63,114,942,995↑

2. タグとは

タグ: 動画に関連する文字列 ※ 1動画に11個まで

例えば、

ゲーム, Skyrim, かわいいは正義, アニメ, BGM

カテゴリタグ

運営会社によって決められたタグ (31個) ゲーム, 自然, 旅, etc.

ニコニコ動画にとってタグは、動画の特徴づけ、検索、分類するのに重要な役割

例えば、

タグ: ゲーム

検索!



3. 問題

動画にとって重要なタグが付与されていない場合が存在



この動画は11個のタグがある。しかし、「初音ミク」タグがない。

新しく投稿された動画には十分なタグが付与されていない場合が存在



この動画は「料理」タグのみ

しかし!

4. 目的

タグ検索をもっと便利にするために、
タグ推定を行う手法を開発

5. 提案手法

- Step 1
- タグとコメントの組に対するスコア付与
 - 情報理論と統計情報を利用
- Step 2
- スコアとタグの共起情報を用いてタグを推定
 - アルゴリズム "Estimation of tags" を開発

6. スコア付与

スコア $S(t, c)$ はタグ t とコメント c の関連度を計算

$$S(t, c) = PMI(t, c) \times ITF(c) \times \left(0.5 + 0.5 \times \frac{tdf(t)}{\max_t tdf(t)} \right)$$

タグとコメントの関連度

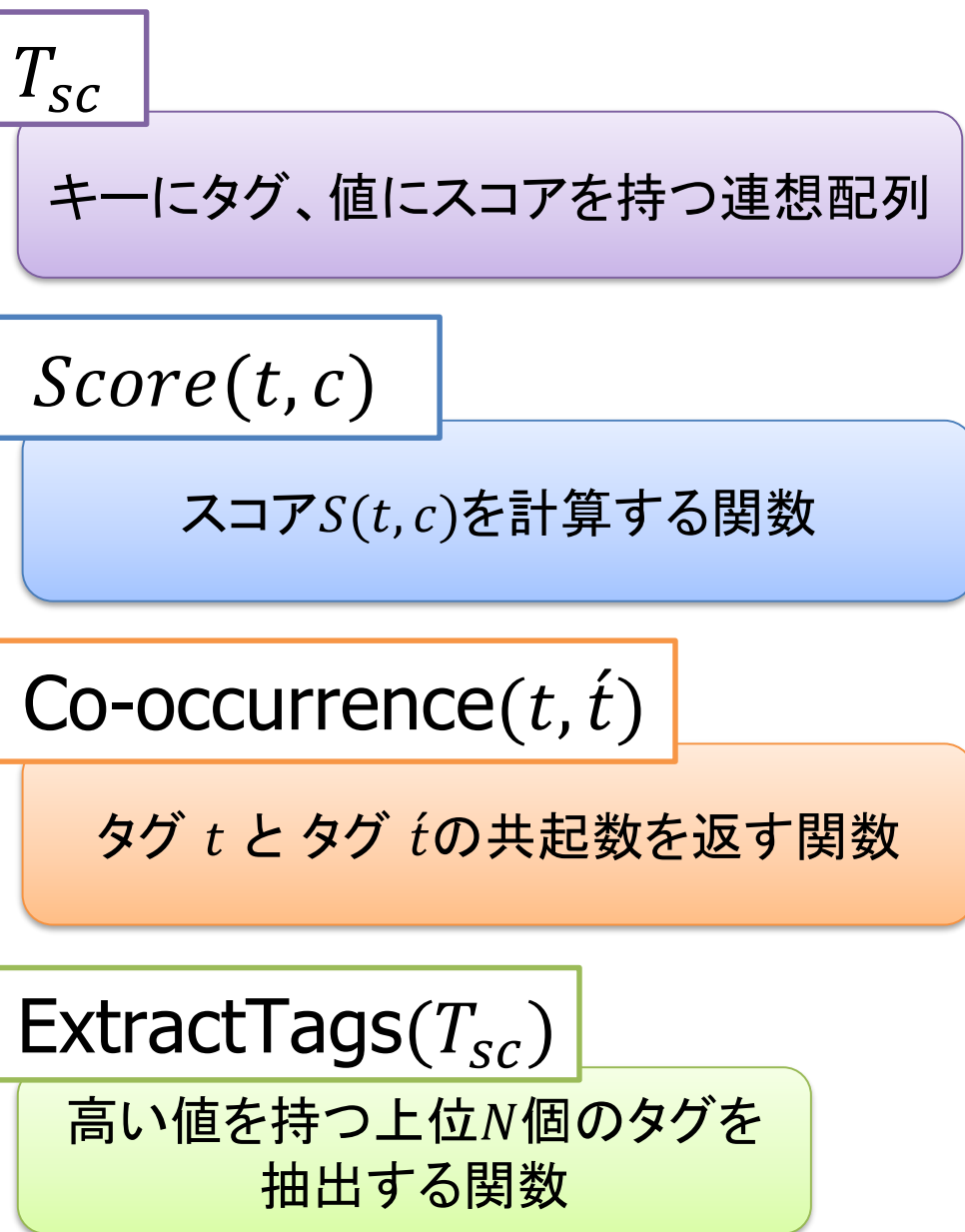
一般的なコメントを抑制

カテゴリタグとレアタグの制御

7. Estimation of tags

タグ推定のための新しいアルゴリズム

Input: A video $V = (C, T)$
C: Comment Set ($c_0, c_1, \dots, c_n$)
T: Tag Set ($t_0, t_1, \dots, t_m$)
Output: New Tag Set $\hat{T}$
1: $\hat{T} \leftarrow \emptyset$
2: $T_{sc} \leftarrow \emptyset$
3: for each $c \in C$ do
4: for each $t \in D_c(c)$ do
5: $T_{sc}[t] \leftarrow T_{sc}[t] + Score(t, c)$
6: end for
7: end for
8: for each $t \in T$ do
9: for each $\hat{t} \in D_t(t)$ do
10: if $\hat{t} \in T_{sc}$ then
11: $T_{sc}[\hat{t}] \leftarrow T_{sc}[\hat{t}] \times Co-occurrence(t, \hat{t})$
12: end if
13: end for
14: end for
15: $\hat{T} \leftarrow ExtractTags(T_{sc})$
16: return $\hat{T}$



本研究では、 $N = 5$

8. 評価実験

- スコア計算用のデータセット
 - 779,154 動画
 - 244,942,338 コメント
- テスト用のデータセット
 - 11 動画
 - 11,486 コメント
- Baseline
 - タグとコメントの共起頻度に基づく手法

	Precision		
	our method	baseline	Labeled LDA
First	0.82	0.36	0.09
Third	1.0	0.55	0.18
Fifth	1.0	0.64	0.18
Three	0.64	0.33	0.09
Five	0.58	0.33	0.05

9. タグ推定の例

Video ID	Existing tags	Estimated tags
sm10864587	自然	標識:1, 料理:0, BGM:1, エンターテイメント:1, 科学:1
sm13139215	VOCALOID, 鏡音レン, 鏡音リン, etc.	初音ミク:1, ゲーム:0, カイト:1, 巡音ルカ:1, グミ:1
sm900149	Avril Lavigne, グレンラガン, シモン, カミナ	音楽:1, 歌ってみた:0, ゲーム:0, VOCALOID:0, 西洋音楽:1

sm10864587



sm13139215

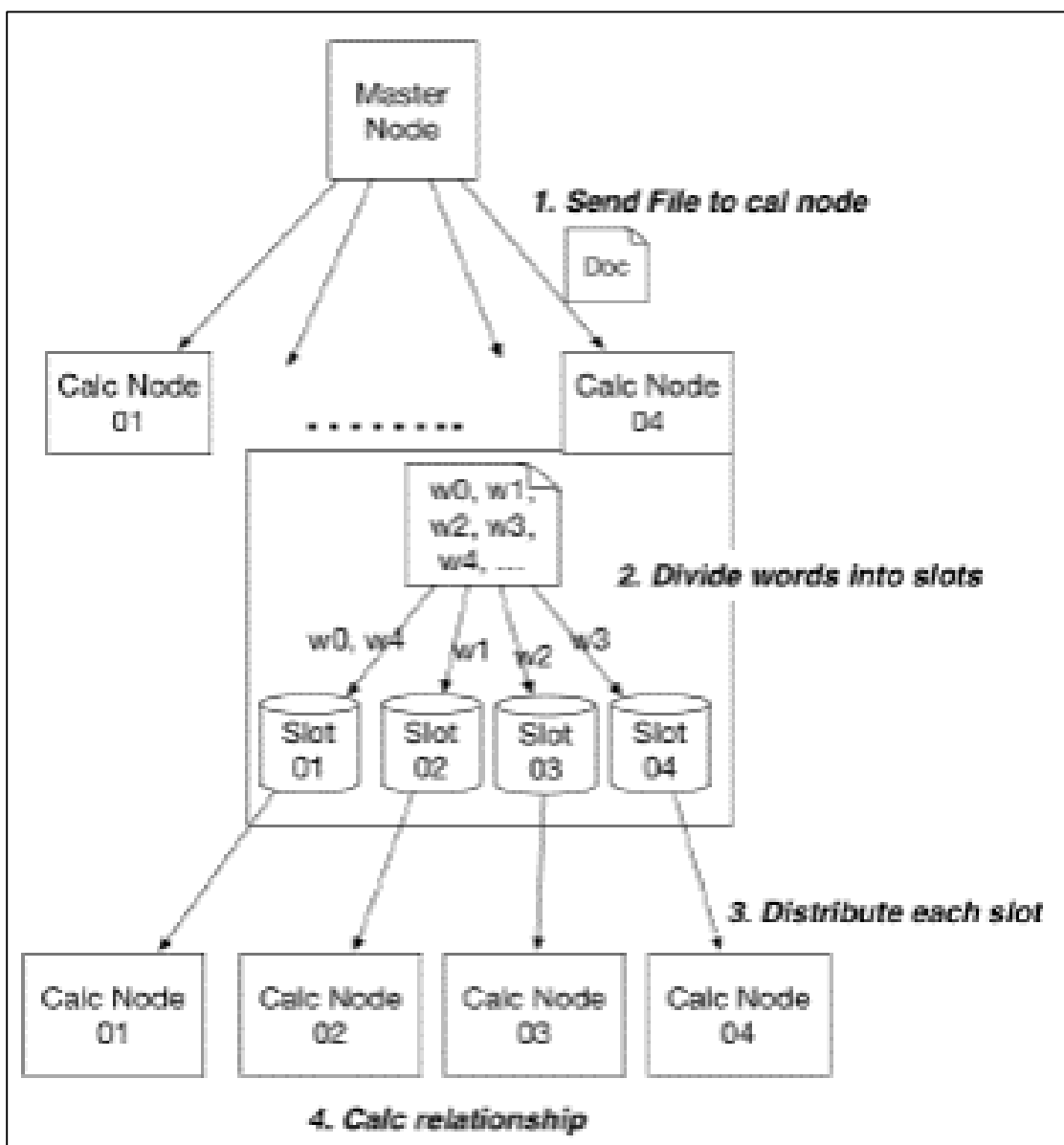


sm900149



10. 今後の発展1

- スコア計算の速度に時間がかかるため、並列分散による高速化に挑戦
- http通信を使うことで、データ分散を行いながら、並列計算を行う仕組みを開発
- 1マシンで計算するよりも、高速に計算できることを確認



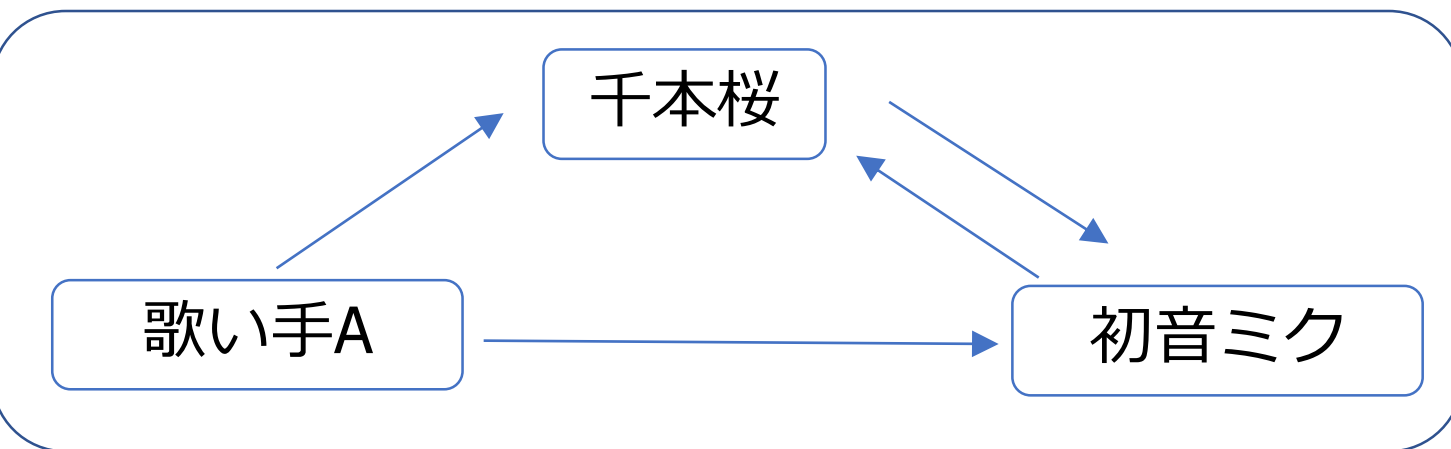
Masaki Kohana, Hiroki Sakaji, Akio Kobayashi, Shusuke Okamoto, A Distributed Calculation Scheme for Contents Categorization, The 31st IEEE International Conference on Advanced Information Networking and Applications (AINA-2017), pp.614-620, Taipei, Taiwan, March, 2017.

12. まとめ

- タグを推薦する手法を開発
 - 新しいスコアリング
 - 新規アルゴリズムによるタグ推薦
- 今度の課題
 - 示した今後の発展を本手法に組み込む
 - 他のデータへの応用
 - Instagram, Twitter, 学術論文

11. 今後の発展2

- ニコニコ大百科から辞書を構築することで、タグ推薦の性能向上を目指す
- 本ドメインに合わせて方向を持った辞書を構築



- 例えば、「歌手A」にとっては、楽曲「千本桜」や「初音ミク」の曲は歌っているため有向線を引く
- しかしながら、「千本桜」や「初音ミク」からは「歌手A」は数ある歌手の一人であるため、有向線を引く必要がない

Akio Kobayashi, Hiroki Sakaji, Kohana Masaki, A Method for Extracting Correct Links from Automatic Created Links on Folksonomy, The 6th International Workshop on Web Services and Social Media In conjunction with The 20th International Conference on Network-Based Information Systems (NBIS-2017), pp.1144-1150, Toronto, Canada, August, 2017.