

A Discriminative Model Approach for Suggesting Tags Automatically for Stack Overflow Questions

AVIGIT K. SAHA, RIPON K. SAHA[†] KEVIN A. SCHNEIDER*

2013 10TH WORKING CONFERENCE ON MINING
SOFTWARE REPOSITORIES PP. 73-76

背景

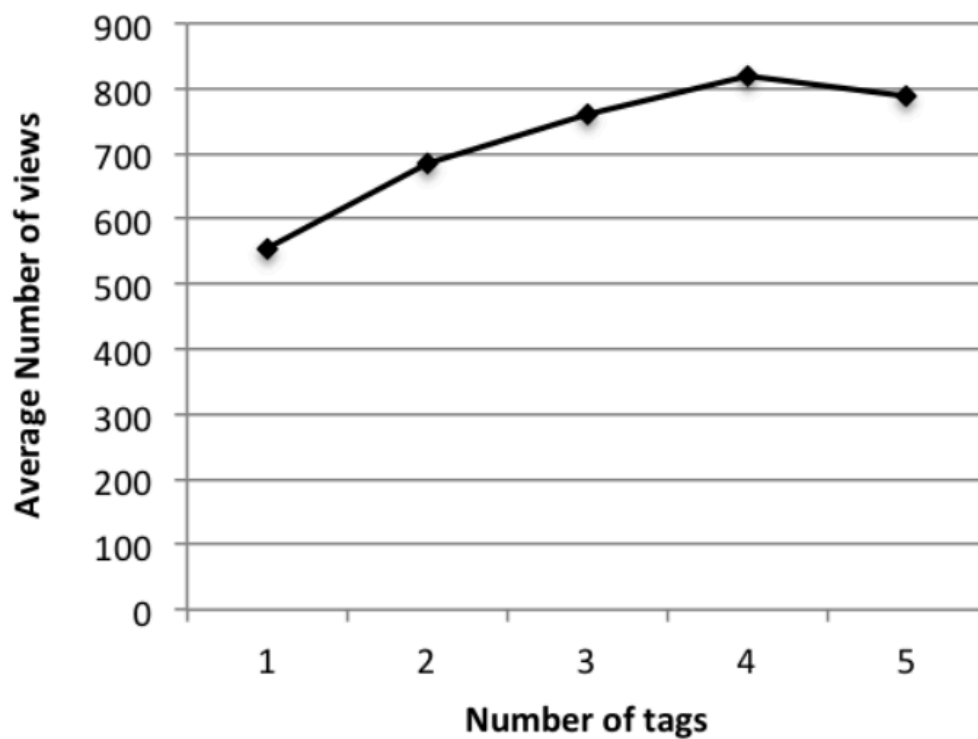
情報システムではタグ付けによる情報の分類・検索は一般的な方法である

StackOverflowでは質問に既存または新たなタグを5つまで追加できる

PROPORTION OF QUESTIONS BY NUMBER OF TAGS

Number of Tags	Number of Questions	[%]
1	438,475	12.70%
2	887,037	25.68%
3	997,906	28.89%
4	693,234	20.07%
5	437,090	12.66%

タグ数と閲覧回数の関係性



質問に答えてもらうには？

タグの数による視認性の向上

タグの増加による「承認された」回答の増加

質問に対してより適切に

タグを付けられるようになるべき

→ タグを付けるための自動システムを構築

自動システムのメリット

5つ未満のタグを持つ既存の質問に
タグを付けられる

適切なタグを提案することにより
新規質問者の手助けができる

質問者が不適切なタグを
選択する傾向があるかどうか注意してくれる

アプローチ

主に3ステップ

質問をベクトルに変換

識別モデルの訓練

与えられた質問のタグを提案

質問をベクトルに変換

各質問を高次のベクトル空間に変換

質問テキストをトークン分割するため字句解析を行う

トークン化後にストップワードを削除,

用語の重要度を測定するため

term frequencyを使用する

識別モデルの訓練

トレーニングデータセット

StackOverflowにある100万件以上の質問

テストデータセット

タグごとにランダムで質問を選択

834個のタグ × 20の質問

これらのデータを元にSVMで識別モデルを構築する

与えられた質問のタグを提案

Algorithm 1 Propose probable tags

Procedure proposeTags

Input: Q : a new question; M : all models

Output: *result* : a list of suggested tags for Q

Body:

- 1: *listOfTags* = a <key,value> map where key is a model tag m and value is its *similarity*
 - 2: **for each** model $m \in M$ **do**
 - 3: $similarity = predict(Q, m)$
 - 4: add < m , *similarity*> to *listOfTags*
 - 5: **end for**
 - 6: sort *listOfTags* in descending order of *similarity*
 - 7: **return** sorted *listOfTags*
-

与えられた質問のタグを提案

Algorithm 2 Calculate similarity between Q and a model

Procedure predict

Input: Q : a new question; m : a model

Output: max : similarity between Q and model m

Body:

- 1: $max = 0$
 - 2: $vectors = featureVector(Q)$
 - 3: $similarity = SVMPredict(vectors, m)$
 - 4: **return** $MAX(max, similarity)$
-

結果

3つのRQ

- このモデルがどれくらい正確に各タグに対応しているか？
- このシステムはある質問のタグをどれくらいうまく提案できるか？
- このモデルで欠けているタグを提案できるか？

このモデルがどれくらい正確に各タグに対応しているか？

$$\text{accuracy} = \frac{\text{number of correctly classified questions}}{\text{total number of questions}}$$

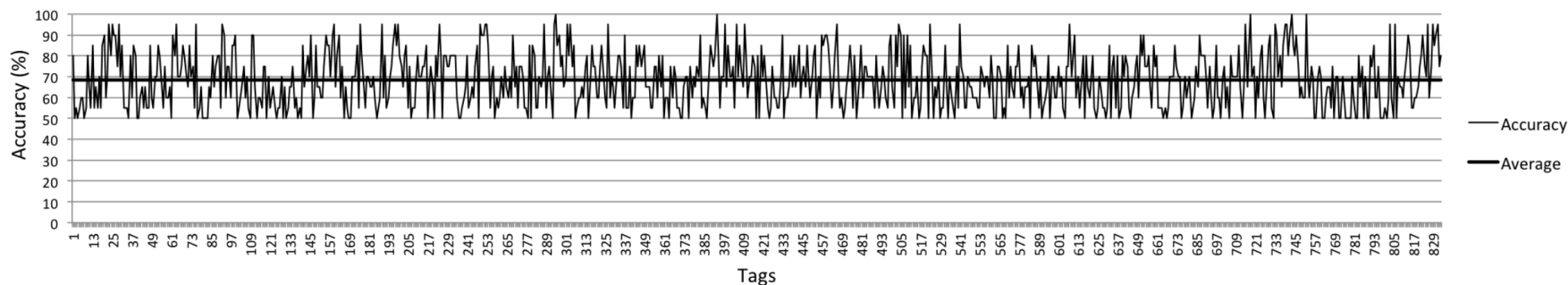


Fig. 2. Accuracy of Discriminative Modes for each Tag

このシステムはある質問のタグをどれくらいうまく提案できるか？

COMPARISON BETWEEN ORIGINAL TAGS AND SUGGESTED TAGS

Id	Original Tags	Suggested Tags
11744046	ios5, uiviewcontroller, uiscrollview	scrolling, uiscrollview, uikit, uiview, ipad, uiviewcontroller, cocoa-touch, uinavigationcontroller, ios5, textview, uitabbarcontroller, uitableviewcell, uiimageview, uitextfield
11709481	iphone, xcode, delegates, uitabbarcontroller, storyboard	uitabbarcontroller, uinavigationcontroller, tabs, xcode4.2, uikit, cocoa-touch, interface-builder, uiview, uiviewcontroller
11726620	ios, uibutton, uitabbarcontroller, tabbar, segue	uitabbarcontroller, tabs, uinavigationcontroller

このシステムはある質問のタグをどれくらいうまく提案できるか？

11730588	iphone, storyboard, uitabBarController, modalviewController, uistoryboardsegue	uiviewcontroller, uiview, uitabBarController, uINavigationController, cocoa-touch, interface-builder
11738761	cocos2d-iphone, box2d-iphone	cocos2d, cocos2d-iphone, textview, xmpp
11749370	html, css	css
11738714	android, android-listview	android-listview, android-intent, checkbox, android-widget
11745574	cocoa-touch, uitableview, core-data, nsfetchedresultscontrolle	core-data, nsarray, ios4, iphone-sdk-4.0, xsd, uitableviewcell, iphone-sdk-3.0
11733482	datepicker, sencha-touch-2	extjs, datepicker, extjs4, sencha-touch
11737147	google-maps, google-maps-api-3	gwt, google-maps-api-3, google-analytics, shell, google-app-engine, nullpointerexception, jvm

このモデルで欠けているタグを提案できるか？

MISSING TAGS

Id	Original Tags	Missing Tags
11748476	android	json
<u>11733330</u>	telerik	datetime, date
11732592	jquery	jquery-selectors, webforms
11718763	android	java
11699099	android	tcp, crash
11737147	google-maps, google-maps-api-3	gwt
11744046	ios5, uiviewcontroller, uiscrollview	uiview
11745574	cocoa-touch, uitableview, core-data, nsfetchedresultscontrolle	uitableviewcell

Title “Disabling Holidays in a Calendar”

結論

タグの数が増えるにつれて、問題ごとの平均ビュー数が徐々に増加し、受け入れられる回答が得られる可能性が高くなる

識別モデルを使用したタギングシステムにより平均68.47%の精度で新しい質問のタグを提案できた

今後の方針

正確性の向上の為手動調査による
モデルチューニング

タグとレスポンスの関係の探求

所感

提案タグが多すぎて数撃ちや当たるな感じがした

SVMPredictやfeatureVectorについてよくわからなかった