Tag Prediction for Posts on StackExchange Site

Author: Tao LI Supervisor: Prof. Dit-Yan Yeung

The Problem

Formally, the problem is that given a training set of a series of questions with user tagged tags, and a testing set of only questions without tags, it is required to predict the appropriate tags of the questions in the testing set

The Challenges

- Scalability and Efficiency
- The Over-fitting problem

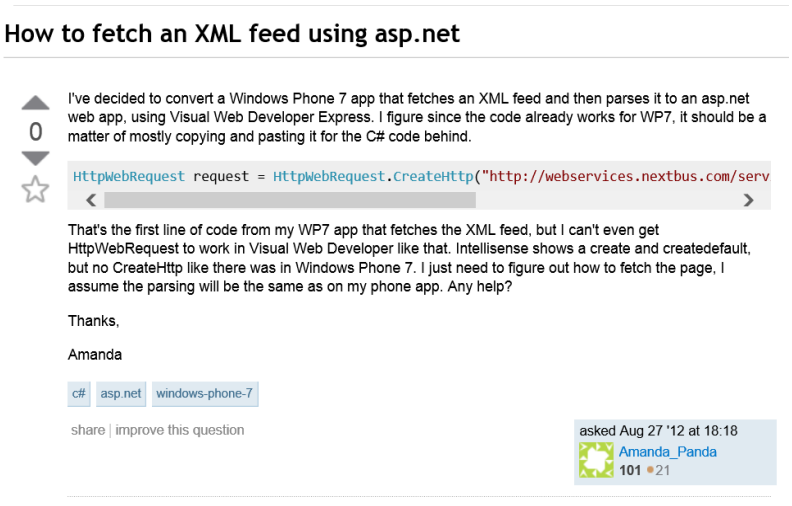
Abstract

In this project, we tackle the problem of automatically predicting the tags of user posts on the StackExchange website according to their topics. We model it as a mulit-label classification problem and innovatively use the K-Nearest-Neighbor like method with the help of inverted-index. As a result, we overcome the efficiency and over-fitting problem faced by most existing methods and achieve a high scalability with reasonable good results.

Tag Statistics and Top Tags

Type	Num
No. of Rows (training size)	6,034,195
No. of Unique Tags	42,048
No. of tag occurrences	17,409,994
Avg. No. of Tags	2.89

Tag	Count
C#	463,526
Java	412,189
Php	392,451
Javascript	365,623



How to fetch an XML feed using asp.net

0

I've decided to convert a Windows Phone 7 app that fetches an XML feed and then parses it to an asp.net web app, using Visual Web Developer Express. I figure since the code already works for WP7, it should be a matter of mostly copying and pasting it for the C# code behind.

```
HttpWebRequest request = HttpWebRequest.CreateHttp("http://webservices.nextbus.com/serv
```

That's the first line of code from my WP7 app that fetches the XML feed, but I can't even get HttpWebRequest to work in Visual Web Developer like that. Intellisense shows a create and createdefault, but no CreateHttp like there was in Windows Phone 7. I just need to figure out how to fetch the page, I assume the parsing will be the same as on my phone app. Any help?

Thanks,

Amanda

c# asp.net windows-phone-7

share | improve this question

asked Aug 27 '12 at 18:18
Amanda_Panda
101 ●21

A Screenshot of Sample Problem

A Sample Post

Each post consists of:

- The Title
- The body in HTML
- Tags

Data Preprocessing

- HTML tag cleaning
- Stemming
- Stop words removal
- Frequency fitltering

⇨

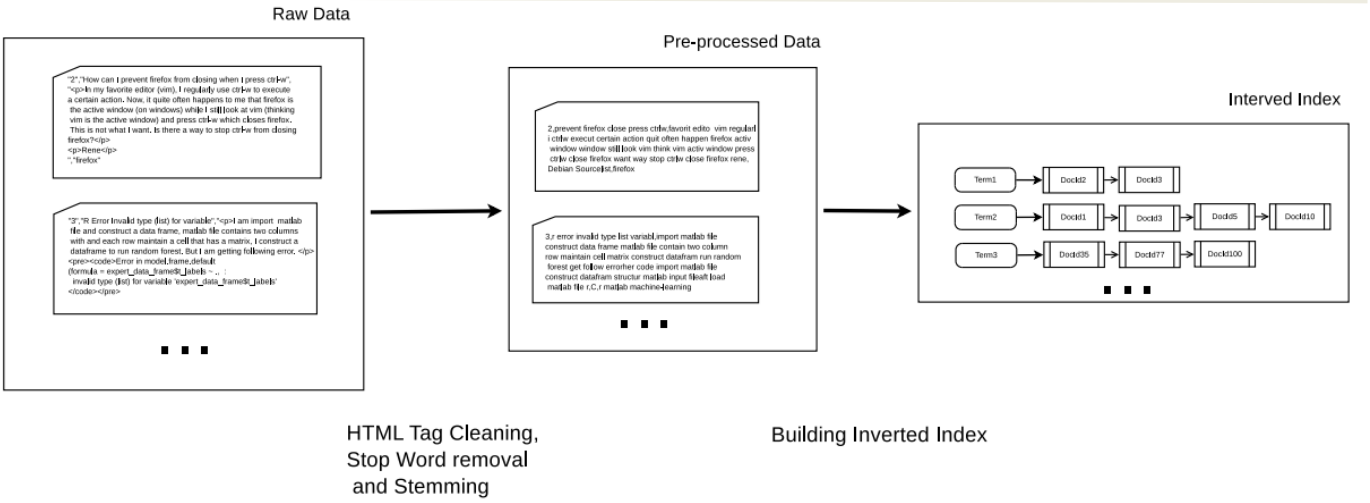
1. Shorter documents
2. Fewer terms

⇨

Processed Sample

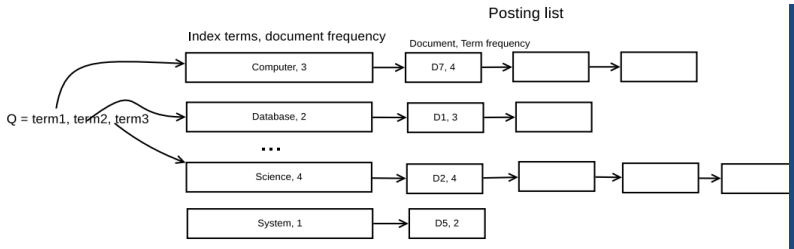
fetch xml feed use aspnet,

ve decid convert window phone app fetch xml feed pars aspnet web app use visual web develop express figur sinc code alreadi work wp matter mostli copi past c code behindthat first line code wp app fetch xml feed ca nt even httpwebrequest work visual web develop intellisens show creat createdefault createhttp window phone need figure fetch page assum pars phone app help thank amanda,

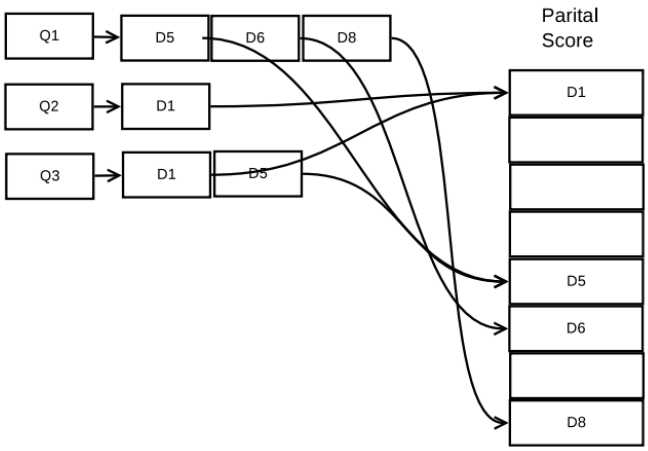


Data Preprocess and Index

The Index and K-Nearest Neighbor Extraction



The Structure of Inverted Index



Data Preprocess and Index

- Extracting the nearest neighbor is equivalent to querying the similar documents from inverted-index
- The inverted-index list documents containing the term.
- The inverted index calculating the score by iteratively adding up partial sums.

Classification

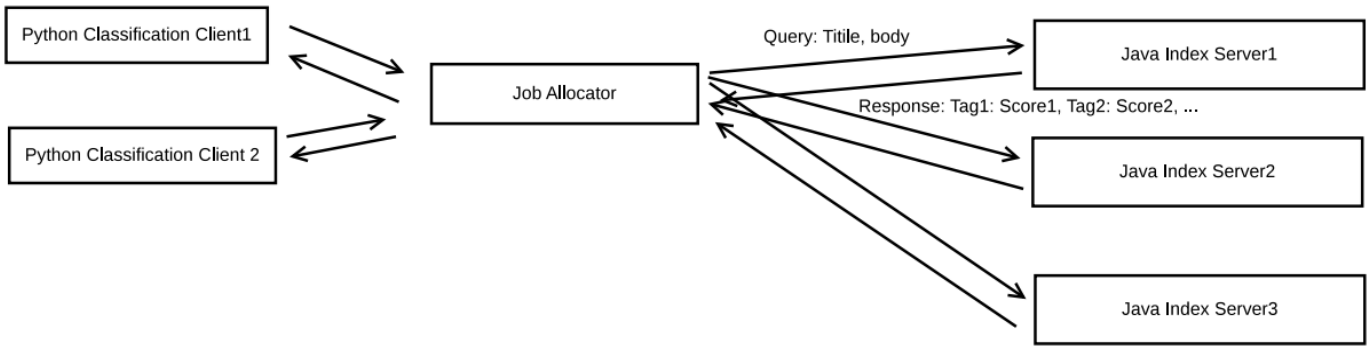
Voting Method

- The documents in the candidate space (the neighbors) essentially votes for their tags.
- The more relevant the document to the query, the more weight it contributes to its tags.
- Pick the tags with score higher than a threshold

BayeSian Method

- Assume that the scores of the tags given by the voting method follow Gaussian distribution
- Build a Gaussian model for each tag as likelihood
- Use Bayes' rule to make discriminant

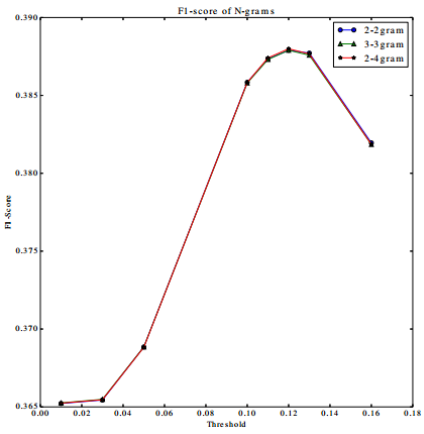
Implementation



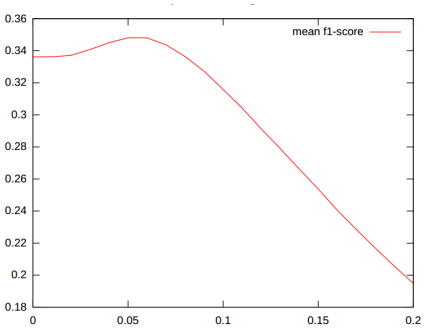
The Client-Server Architecture of Implementation

Experiments

- Use f1 score to measure the quality
- Index more than 6,000,000 item in less than 20mins
- Predict each instance in less than 0.2 sec
- Max score with voting method is 0.388
- Max score with Bayesian method is 0.398



The experiments on n-gram model with voting method. The F1-score changes with threshold



The change of f1 score with respect to the change of threshold. With bag -of-word model