Machine Learning webcontent categorization
Artificial Intelligence

Masumi Mutsuda Zapater

February 2012

Abstract

This is the abstract

Contents

Chapter 1

Introduction

Hello my name is here [1]

Chapter 2

Conclusions

Bibliography

[1] Ronald L. Graham, Donald E. Knuth and Oren Patashnik, *Concrete mathematics* Addison-Wesley, Reading, MA, 1995.

[2] alias-i, "Logistic Regression Tutorial", *Lingpipe Home*, 23 Nov. 2011, `http://alias-i.com/lingpipe/demos/tutorial/logistic-regression/read-me.html`

Chapter 3

Annex: Diary

This annex contains the diary notes I took during the project development. It is written in catalan and it may contain partial or disconnected information. Reader discretion is advised.

- Identificació del problema: On es determina la viabilitat de la construcció del SBC i la disponibilitat de les fonts de coneixement.
- Conceptualització: Descripció semiformal del coneixement del domini del problema i descomposició en subproblemes, segons la visió d'un expert.
- Formalització: Cal definir el mecanisme adequat de representació del coneixement, en aquest cas segons la visió de l'enginyer de coneixement.
- Rapidminer per testejar diferents solucions, naive bayes rapid, maxent lent pero + accuracy
- Categoritzacio URL per contingut -¿ MaxEnt
- Classificador basat en categories de Yahoo Directory (training)
- Testing inicial amb petites dades
- Testing posterior amb 140.000 URLs d'Irlanda O2
- Descartar idiomes estrangers -¿ Trigraphs
- Fitxers amb trigrams d'idiomes generats via randomly crawling wikipedia
- Crawling a traves de proxy amb headers tema idioma
- Afegida categoria adult
- Afegit categoria other, massa generica
- Desmembrat categoria other en altres per posterior mapeig a other
- Generat script per calcular mitja de certesa de classficacio i distancia mitja amb segona opcio

- Millorats fitxers de train en base al punt anterior

El primer approach per la classificació va ser fent servir les categories que demanava O2 irlanda. La categoria entertainment englobava molts temes, movies, tv, music, literature..., i es requeria a més una categoria "Other". El classificador sempre assigna una categoria, la que té la màxima probabilitat de ser, i per tant s'havia de generar la categoria other a partir de categories que no tinguessin res a veure. El problema de crear categoria generica other o entertainment, es que passaven a tenir molt ambigüitat, i moltes coses passaven a considerar-se other i entertainment. La solucio va ser crear subcategories sense tenir en compte entertainment i other, i fer un post tractament del ficher resultant de la classificació assignant other o entertainment a allò que realment ho era. D'aquesta manera es té més granularitat, tot i que l'accuracy del classificador baixa al tenir més categories. Un problema comú és que algunes urls no es deixen crawlejar. Detecten que no ets un navegador corrent i et donen un contingut que no és significatiu. A l'hora de classificar aquestes urls acaben sent categoritzades a categories que no tenen res a veure. Per exemple wikipedia retorna simplement una llista de països, que fa que la categoria passi a ser "Adult". Coses similars passen amb facebook. A youtube el problema és que el text crawlejat són els títols i descripcions de videos que la gent puja, per tant depenent de quan es produeixi el crawling, la classificacio de Youtube pot canviar.

29-12-2011

En afegir categories concretes per a ser englobades posteriorment per Entertainment, l'accuracy del classificador ha baixat fins al 73.33 adult education food_drink health literature music real_estate religion social_networking sport travel automotive email games history maps news reference science social_science television weather crime financial government instant_messaging movies photos regional shopping software theme_parks

Si la confusió és entre categories que seran englobades per "Entertainment" no hi ha problema. Observem que la distància disminueix lògicament perquè ara hi ha més categories.

05-01-2012

Proves a jabato

Figure 3.1: Gràfic de blabla