Website Classification

Akshay Kumar(10-CSS-06)

Niyas C(10-CSS-44)

Introduction

Website classification or website categorization is the process of assigning a website to one or more category labels. E.g. “News”, “Sport” , “Business”

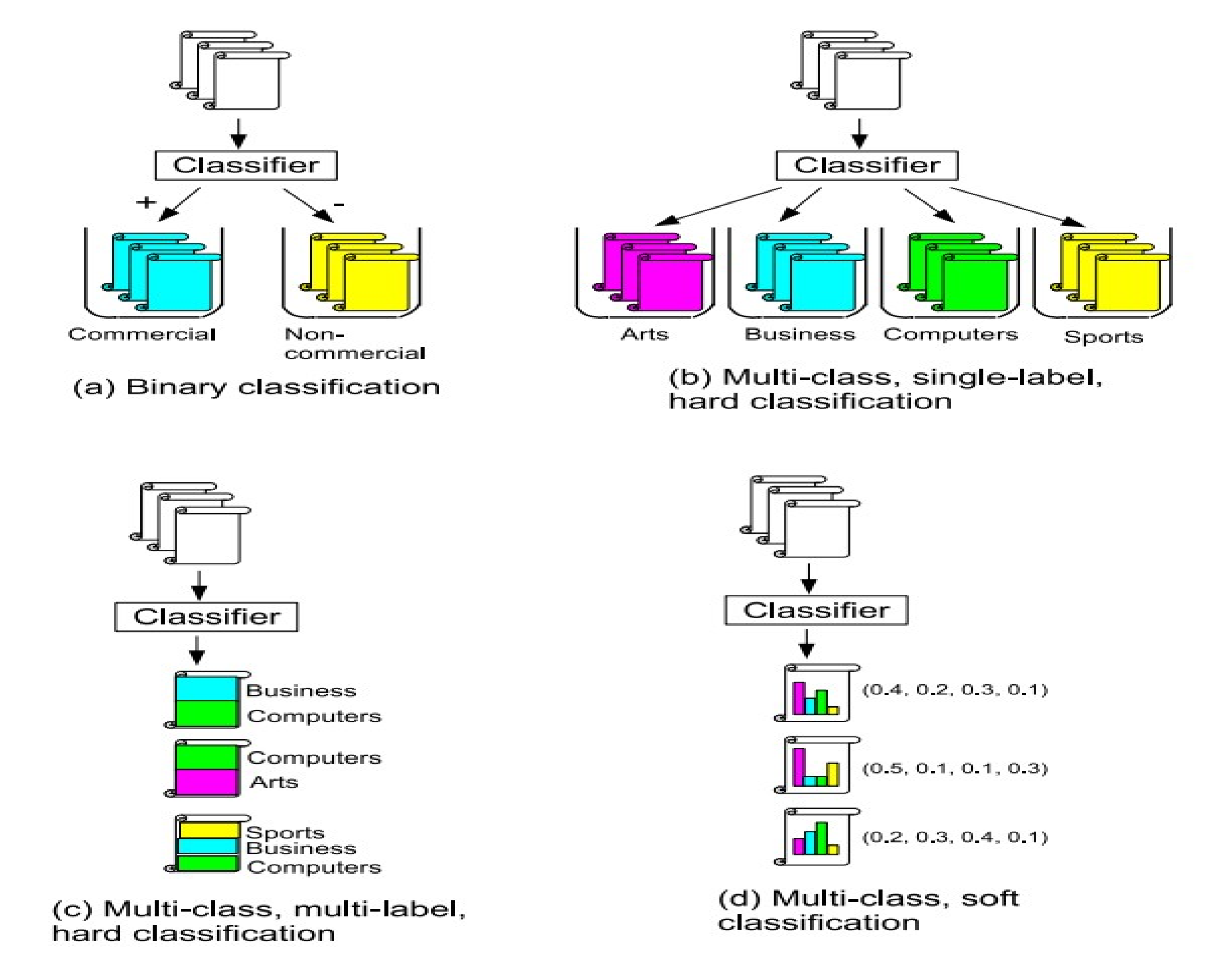
- Website classification can be divided into two categories
  - **Subject Classification:** On the basis of subject or topic of website. Eg: “sports,politics,technology...etc”
  - **Functional Classification:** The role that website play. Eg: “Personal website,Enrollment website ...etc”

- Based on number of classes website classification can be divided into
  - **Binary classification :** Two classes. For ex:

    “political and non-political”,”commercial and non-commercial”
  - **Multi class classification:** Multiple classes
    - Soft classification
    - Hard classification

- Based on the number of classes that can be assigned to an instance, classification can be divided into
  - **Single Label :** One instance will be strictly assigned to one category only
  - **Multi Label :** One instance can be classified into multiple categories

Overview of classification



(a) Binary classification

(b) Multi-class, single-label, hard classification

(c) Multi-class, multi-label, hard classification

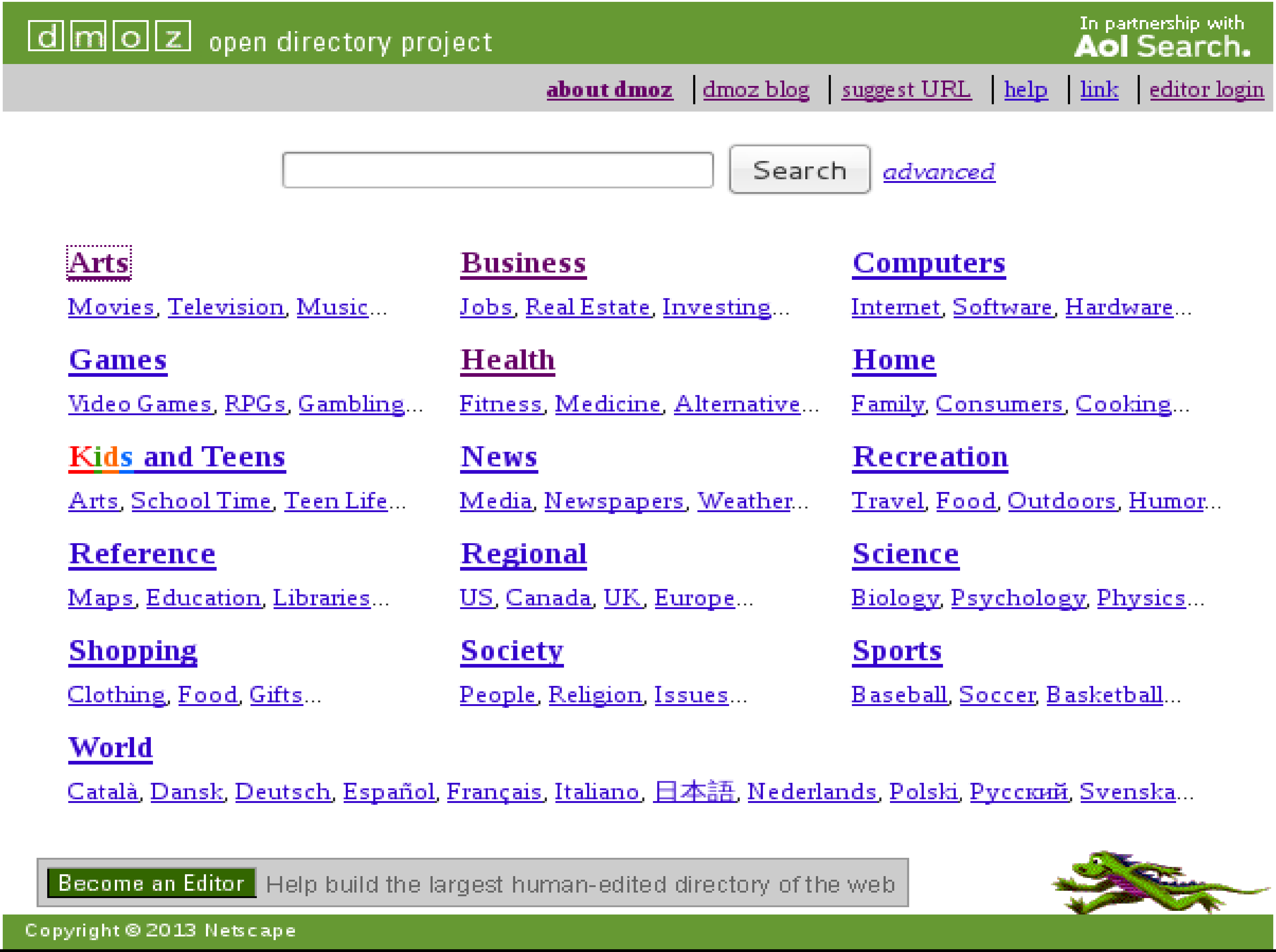
(d) Multi-class, soft classification

Applications

- **Assist web information gathering:**
  - Let to search websites belonging to particular category.
  - Help to automate web directory projects

Web directory projects

- They are directories in which websites are organized into different categories by human effort
- Eg: open web directory, yahoo directory



dmoz open directory project

In partnership with Aol Search.

about dmoz | dmoz blog | suggest URL | help | link | editor login

Search *advanced*

Arts
Movies, Television, Music...

Business
Jobs, Real Estate, Investing...

Computers
Internet, Software, Hardware...

Games
Video Games, RPGs, Gambling...

Health
Fitness, Medicine, Alternative...

Home
Family, Consumers, Cooking...

Kids and Teens
Arts, School Time, Teen Life...

News
Media, Newspapers, Weather...

Recreation
Travel, Food, Outdoors, Humor...

Reference
Maps, Education, Libraries...

Regional
US, Canada, UK, Europe...

Science
Biology, Psychology, Physics...

Shopping
Clothing, Food, Gifts...

Society
People, Religion, Issues...

Sports
Baseball, Soccer, Basketball...

World
Català, Dansk, Deutsch, Español, Français, Italiano, 日本語, Nederlands, Polski, Русский, Svenska...

Become an Editor Help build the largest human-edited directory of the web

Applications

- Content filtering
  - Parent Control:
    - Let parents to impart restriction on internet access of children by blocking non-desirable categories.
  - Manage organization/institution network
    - Let organization/institutions to block access to particular categories of websites to prevent students/employees from wasting their time and band width of institution.

      For example Jamia use cyberoam portal to control internet usage of staff/students and it contains a semi automated website classifier.

Features used in website classification

- URL
  - “Fast webpage classification using URL features”
    - Min-Yen Kan,Hoang Oanh Nguyen Thi
- HyperLink
  - Part inside <a>...</a> tag
- Textual Content
- Powerful html tags
- Structure of website
- Statistical information

What we plan to do?

- Desktop application using python for predicting category of input websites
- Prepare training set and test set using sample database provided by open directory project.
- Achieve a classification accuracy more than 80%
- Demonstrate classification accuracy

How we do?

- Based on Naive Bayes Classification method
- Base research paper
  - Automated Classification of Web Sites using Naive Bayesian Algorithm-2012
    - Ajay S. Patil, B.V. Pawar
    - IMECS March 2012,Honkong

Limitations

- There will be several websites that do not belongs to predefined categories
- Multimedia contents in websites.

Thanks