

# Automatic Categorization Web sites

Lida Zhu



## Presentation Outline

- Introduction
- Problem Description
- Background Information
- Proposed Solution
- Result
- Conclusion

## Introduction

- Explosively growth of World Wide Web.
- Great challenges for data mining.
  - Size too huge
  - Complexity of web pages is too difficult
  - Constantly updating
  - Diversity of communities
- Traditional data mining become inadequate.
- This project focus on web mining field.
- Use keyword-based classification to solve automatic web sites classification problem.

Introduction

Problem Description

Background Information

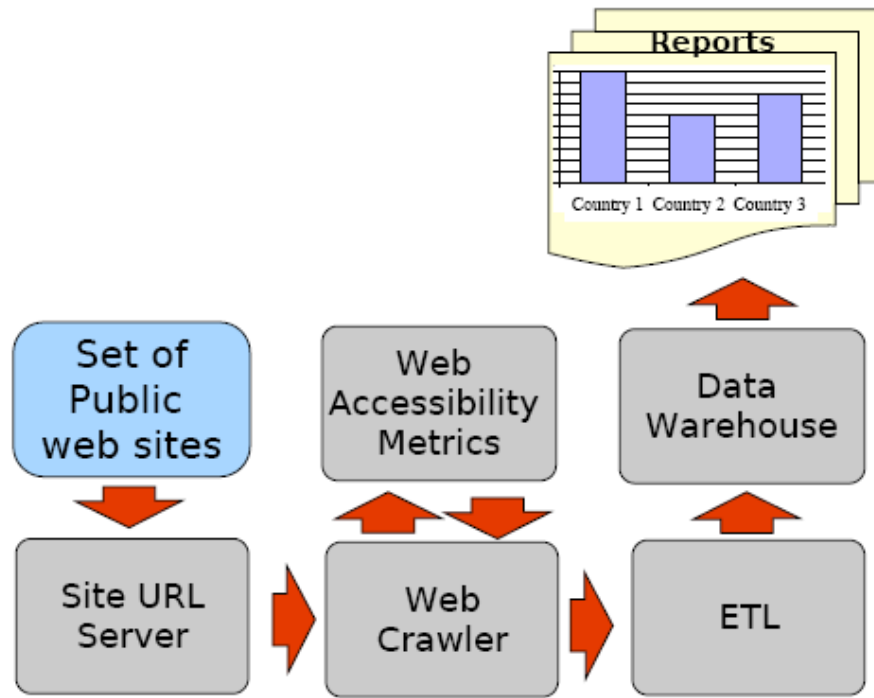
Proposed Solution

Result

Conclusion

## Problem Description

- The goal is to perform automatic Web sites categorization for EIAO machinery.



Introduction

Problem Description

Background Information

Proposed Solution

Result

Conclusion

## Problem Description cont.

- Classify web sites into NACE?
- Classify web sites into NUTS?
  - [Manual classification list from EIAO.](#)
- How to deal with large, complex form, data set?
  - [Preprocessing the website.](#)
- How to lower down the high-dimensional vector space?
  - [Extracting most useful features.](#)
- Determine the most appropriate classifier?
  - [Compare and evaluate common classifiers.](#)

Introduction

**Problem Description**

Background Information

Proposed Solution

Result

Conclusion

## Background

- NACE
  - NACE is a statistical classification of economic activities used within European Community.
- NUTS
  - NUTS is a statistical standard classification at a regional level for EU members and EFTA countries in geography.

Introduction

Problem Description

Background Information

Proposed Solution

Result

Conclusion

## Background cont.

- Remove Stopword
  - Stop words are set of non-informative words, such as “a, the, of, for, with”, and so on.
  - Save spaces for storing document contents
  - Improve efficiency and accuracy of classification.
- Skip html
  - “skip-html” skips all the words in “<>”
  - Useful for tokenizing (X)HTML files.

Introduction

Problem Description

Background Information

Proposed Solution

Result

Conclusion

## Background cont.

- Feature Selection method:
- Mutual Information:
  - Mutual Information measures the associativity between terms and categories.

$$I(t, c) \approx \log \frac{A \times N}{(A + C) \times (A + B)}$$

Introduction

Problem Description

Background Information

Proposed Solution

Result

Conclusion

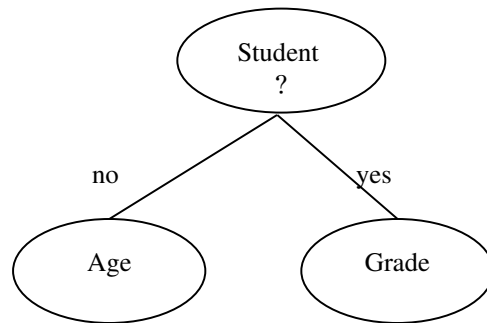
## Background cont.

Classification methods:

- Naive Bayes
- A typical statistical classifier

$$h_{NB} = \operatorname{argmax}_{h_j \in H} P(h_j) \prod_i P(o_i | h_j)$$

- Decision Tree



Introduction

Problem Description

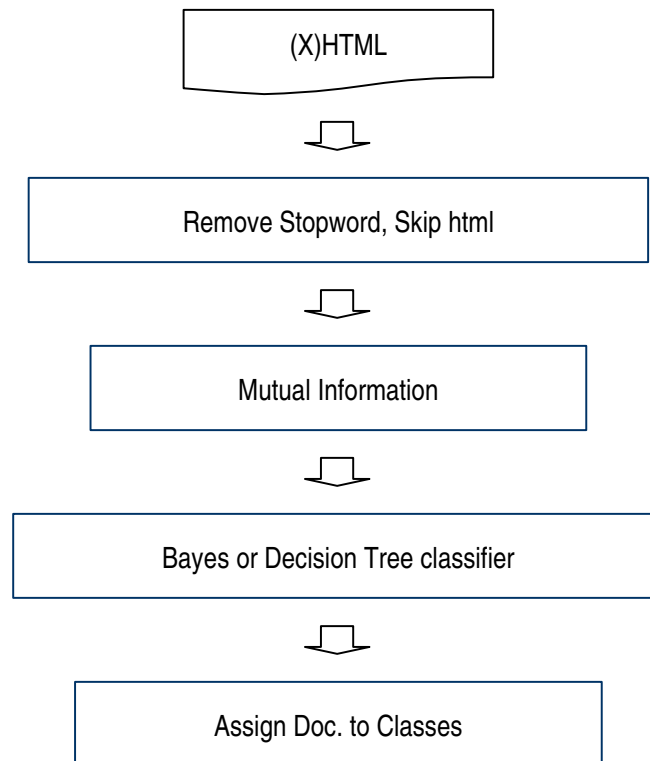
Background Information

Proposed Solution

Result

Conclusion

## Proposed Solution



Introduction

Problem Description

Background Information

**Proposed Solution**

Result

Conclusion

## Result

- NACE

Introduction

Problem Description

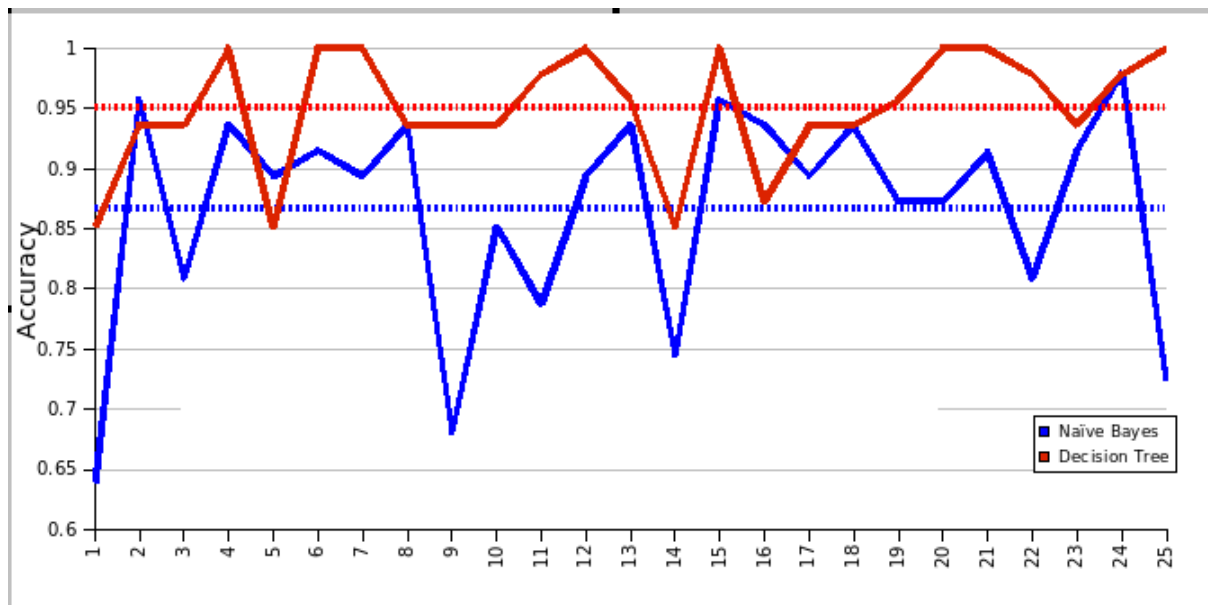
Original class label	Classified class label			
	J&L	K	Total	%
J&L	27.7	0.8	28.5	97%
K	4.5	14	18.5	76%
Total	32.2	14.8	<b>47</b>	
%	86%	95%		<b>88.72%</b>

Naive Bayes

Original class label	Classified class label			
	J&L	K	Total	%
J&L	27.8	0.7	28.5	98%
K	0.5	18	18.5	97%
Total	28.3	18.7	<b>47</b>	
%	98%	96%		<b>97.45%</b>

Decision Tree

## Result cont.



Introduction

Problem Description

Background Information

Proposed Solution

Result

Conclusion

## Result cont.

- NUTS

Introduction

Problem Description

Original class label	Classified class label			
	Ireland	UK	Total	%
Ireland	1.1	23	24.1	5%
UK		64.9	64.9	100%
Total	1.1	87.9	<b>89</b>	
%	100%	74%		<b>74.16%</b>

Naive Bayes

Original class label	Classified class label			
	Ireland	UK	Total	%
Ireland	20.3	4	24.3	84%
UK	2	62.7	64.7	97%
Total	22.3	66.7	<b>89</b>	
%	91%	94%		<b>93.26%</b>

Decision Tree

## Conclusion

- The proposed strategy had shown good performance.
- NACE has 97% accuracy in Decision Tree, 88% in Naive Bayes
- NUTS has 93% accuracy in Decision Tree, 73% in Naive Bayes
- The proposed solution offers an accurate, reliable classification tool for EIAO.

Introduction

Problem Description

Background Information

Proposed Solution

Result

Conclusion