MLeXAI Workshop

Ingrid Russell, Zdravko Markov

Sanibel Island, FL, May 18, 2009

Web Document Classification

Originally developed in summer 2005 by Ingrid Russell, Zdravko Markov, and Todd Neller

Revised with material from the

Probabilistic Resoning Project

Developed in summer 2007 by Zdravko Markov and Ingrid Russell

Introduction

- Topic directories (dmoz.org)
- Automatic classification of web pages
- Expanding and creating new directory structures
- Investigating the process of tagging (labeling) web pages using topic directory structures
- Applying Machine Learning techniques for automatic tagging

Objectives

- Learn basic concepts and techniques of machine learning
- Implement a learning system
- Understand the role of learning for improving performance and allowing a system to adapt based on previous experiences
- Understand the importance of data preparation and feature extraction in machine learning
- Learn and apply the vector space model for representing web documents

Project Phases

- Collect web documents
- Extract text and select features
- Represent documents as feature vectors (term-document matrix)
- Prepare data for Weka
- Create and evaluate ML models

Resources

- AI course web page (Prolog programs)
- Weka (software)
- DMW book (sample data)
- Related projects (Probabilistic Reasoning)
- Other (web crawling, text stat)

Reading

- Stuart Russell, Peter Norvig. Artificial Intelligence: A Modern Approach, 2003.
- Tom Mitchell, Machine Learning, 1997.
- Ian H. Witten and Eibe Frank. Data Mining: Practical ML Tools and Techniques, 2005.
- Zdravko Markov and Daniel T. Larose. Data Mining the Web: Uncovering Patterns in Web Content, Structure, and Usage, Wiley, 2007.

Data Collection

- Collect web pages from 5 different topics with at least 20 documents in each
- Choose a more elaborated topic structure (not necessarily a tree)
- Each document should have enough text content
- Each document must include enough terms to represent the topic

Data Collection Tools

- Topic directory (dmoz.org)
- Web browsing
- Web search
- Web crawler (WebSPHINX)

Feature Extraction

- Remove stopwords, apply stemming
- Compute term frequencies in the corpus
- Select 100 most representative terms (consider TF and IDF factors)
- Create term document matrix (binary, TF, TFIDF).

Feature Extraction Tools

- Prolog programs (described in project)
- Specialized text editors
- Weka (described in project)
- Custom-made programs

Term-Document Matrix

- Create a feature vector for each document
  - Binary (0/1, nominal)
  - Term frequency (counts)
  - TFIDF representation (numeric)
- Use Prolog programs or Weka

Data Preparation

- Create data files for Weka
  - CSV format
  - ARFF format
- Use different representations
  - Binary
  - TF
  - TFIDF
- Use Weka for conversions between formats and representations

Machine Learning and Model Evaluation

- Attribute ranking and selection
- Decision trees
- Naïve Bayes
- KNN
- Clustering
- Classification of new documents

Sample Project 1 (UH)

- 5 topics, 116 documents, 1000 terms
  - Machine Learning
  - Agents
  - Sorting
  - MPEG
  - History of computing
- Feature extraction (binary representation) by using TextSTAT, Excel and VB
- ML models and error analysis: Decsion tree, Naive Bayes, KNN

Sample Project 2 (CCSU)

- Two separate topic structures:
  - Musical instruments (5 topics)
  - Four general topics: Non-profit, Government, Personal, Commercial
- Data preparation using Prolog and Weka
- ML models created by Weka
  - Increasing number of features (10,20,30,40)
  - Naïve Bayes, KNN, WKNN, Decision tree (best)
  - Predicting class of new documents

Sample Project 3 (CCSU)

- 5 topics, 100 documents, 100 terms
  - Computer Science
  - Artificial Intelligence
  - Machine Learning
  - Data Mining
- Data preparation using Prolog and Weka
- ML models created by Weka
  - Increasing number of features (25,50,75,100)
  - Naïve Bayes, KNN, Decision tree
  - Predicting class of 15 new documents