Classification of Websites
as
Sets of Feature Vectors

Outline of the Talk

Introduction

Classification of Websites
  Classification as single documents
  Classification as Sets of Feature Vectors
  Centroid Set Optimization

Experimental Evaluation

Conclusion
What is a Website?

**Webpage:** One HTML document accessible via an URL (e.g. [http://www.dbs.informatik.uni-muenchen.de/dbprakt.html](http://www.dbs.informatik.uni-muenchen.de/dbprakt.html)).

**Website:** All HTML documents published by the same person/group serving a common purpose.

**Here:** All webpages published within one domain/subdomain. (e.g. [www.dbs.informatik.uni-muenchen.de](http://www.dbs.informatik.uni-muenchen.de))

---

What is Webpage Classification?

**Feature Transformation of Webpages:**
- feature selection chooses relevant words
- 1 word is 1 dimension in the feature space
- text used as "Bag of Words"
- normalize vectors by the number of words in the text

**Classification of Word Vectors by:**
- naive Bayes Classifiers
- KNN-Classification
- Support Vector Machines
- …

---

Database Group
Institute for Computer Science
University of Munich, Germany
**What is Website Classification?**

Website Classification is the task to predict the purpose of a website, i.e., the common purpose of all webpages published under the same domain.

**Application:** Directory Services like DMOZ and YAHOO list commonly websites and not single documents.

---

**Outline of the Talk**

- Introduction
- Classification of Websites
  - Classification as single documents
  - Classification as Sets of Feature Vectors
  - Centroid Set Optimization
- Experimental Evaluation
- Conclusion
Website Classification (1)

**Idea:** Derive one representative Webpage and apply Webpage Classification

1. **Approach: Homepage Classification**
   - Classify Homepages as representative for each domain.
     (e.g. [http://www.lmu.de](http://www.lmu.de))
   - Homepages are not always meaningful (frames, intros, …).

2. **Approach: Superpage Classification**
   - Merge all Webpages into one common “superpage” / wordvector.
   - Local context is lost.
     \[
     \text{superpage}(W) = \sum_{page \in W} f(page) 
     \]

Website Classification (2)

**Idea:** Transform a website into a set of word vectors.

+ local context of the documents is preserved.

- most classifiers are not directly applicable.

⇒ kNN classification directly uses the trainingset.
  e.i. all data types are applicable as long as there is a suitable *distance measure*. 

Database Group
Institute for Computer Science
University of Munich, Germany
**Distance Measure for Websites**

**Sum of Minimum Distances (SMD):**
Given Websites $W_1, W_2 \in 2^{FV}$, with FV is the feature space of word vectors.

$$SMD(W_1, W_2) = \frac{\sum_{v \in W_1} \min_{w \in W_2} \text{dist}(v, w)}{|W_1|}$$

---

**Centroid Sets**

**Problem:** Bad time performance.

**Idea:** reduce the training set to one representative object for each class  
$\Rightarrow$ What is a good representation of a set of Sets of Feature Vectors?

A Centroid Set represents several sets of feature vectors.
- Merge all vector sets: $W_{all} = \bigcup W_i$
- cluster all the resulting set $W_{all}$ (using e.g. GDBSCAN)
- reduce each cluster to its centroid
Incremental Classification using half-SMD:

\[
\text{half} - \text{SMD}(CS, W) = \frac{\sum \min_{v \in CS} \text{dist}(v, c)}{|W|}
\]

- half-SMD is a fairer distance measure (websites must not contain all concepts)
- fast incremental classification
- achieved best classification accuracy

Outline of the Talk

- Introduction
- Classification of Websites
  - Classification as single documents
  - Classification as Sets of Feature Vectors
  - Centroid Set Optimization
- Experimental Evaluation
- Conclusion
Classification Accuracy

Classification accuracy for a 7-class problem

- 234 websites (18,000 webpages)
- 7 Classes (horse dealer, game retailer, business schools, ghosts, astronomy, snowboard, other)

Classification Time

Classification Time in seconds for 3 two class problems

<table>
<thead>
<tr>
<th></th>
<th>SNN</th>
<th>Centroid Cl.</th>
<th>superpage</th>
<th>homepage</th>
</tr>
</thead>
<tbody>
<tr>
<td>busin.schools</td>
<td>39.2</td>
<td>0.4</td>
<td>0.3</td>
<td>0.1</td>
</tr>
<tr>
<td>horse dealers</td>
<td>22.4</td>
<td>0.3</td>
<td>0.2</td>
<td>0.1</td>
</tr>
<tr>
<td>astronomy</td>
<td>37</td>
<td>0.4</td>
<td>0.3</td>
<td>0.1</td>
</tr>
</tbody>
</table>
Outline of the Talk

- Introduction
- Classification of Websites
  - Classification as single documents
  - Classification as Sets of Feature Vectors
  - Centroid Set Optimization
- Experimental Evaluation
- Conclusion

Conclusion

Summary

- Website Classification is more complex than webpage classification.
- Sets of Feature Vectors and SMD provide a suitable representation space for kNN-Classification of Websites.
- Incremental classification using centroid sets is fast and accurate.

Future Work

- Applying the introduced classifier to a hierarchy of website classes.
- Apply clustering algorithms to websites to dynamically extend class- hierarchies of websites.