Telling Apples From Oranges

Enhanced Web Page Classification

Xiaoguang Qi
• Background
• Utilizing features of neighbors
• Using fielded features
Problem definition

- Classification
  - A set of labeled data is used to train a classifier which can be applied to label future examples.

- Web page classification
  - The process of assigning a web page to one or more predefined category labels.
  - aka. web page categorization
Why important?

- Web page classification is essential to
  - Improving quality of retrieval
  - Maintaining web directories
  - Helping question answering systems
  - Many more ...
Why important? (Cont.)

- Tradition text classification approaches don’t perform well on web pages
  - An experiment on dmoz ODP dataset
    - 12 topical categories
    - 228,000 training documents, 12,000 for testing
  - Naïve Bayes: 35% error rate
  - Support vector machine: 27% error rate
Text is not enough
Text is not enough (Cont.)

- Textual features could be missing, misleading, or unrecognizable.
  - Web page temporarily unavailable, robot exclusion, picture, flash, frame, etc …

- There could be too much irrelevant text.
  - Advertisement, navigational panel, spam follow-ups, etc …

- Solution?
  - Besides text, web pages have other features.
  - Make use of them!
On-page features

http://www3.lehigh.edu/

Living large in the nano world
Gast draws on Packer’s vision to meet today’s challenges
Exploring new horizons
Girardot receives Romania’s Presidential Medal of Scholarship
Meet Lehigh President Alice P. Gast
Lehigh ranks among nation’s top universities

Tours, Chats, and Information Sessions >
Zoellner Arts Center’s Current Season >
The Environmental Initiative at Lehigh >
On-page features (Cont.)

```html
<html>
<head>
  <title>Lehigh University - Campus Homepage for Students, Faculty & Staff</title>
  <link rel="stylesheet" type="text/css" href="/includes/nsstyle.css" title="Default Styles">
  <link rel="stylesheet" type="text/css" href="/includes/iestyle.css" title="Default Styles">
  <META NAME="keywords" CONTENT="student faculty staff start homepage information services events news">
  <META NAME="Description" CONTENT="Start page for information and services for Lehigh Students, Faculty & Staff">
</head>
<body leftmargin="0" topmargin="0" marginwidth="0" marginheight="0" bgcolor="#ffffff" onload="preload">
  <table width="664" border="0" cellspacing="0">
    <tr>
      <td><a href="/"><img src="/images/sitewide/toplogo.gif" width="162" height="74" alt=""></a></td>
    </tr>
  </table>
  <option value="#" selected>Popular Pages</option>
  <option value="http://www3.lehigh.edu/about/news/default.asp">News</option>
  <option value="http://www3.lehigh.edu/academics/colleges.asp">Colleges</option>
</body>
</html>
```
Using on-page features

- HTML tags
  - structural info embedded in HTML documents
  - Golub and Ardo [2005]
    - Assign significance indicators for different HTML tags
    - Title, headings, metadata, and main text

- Kwon and Lee [2000, 2003]
  - Divide all the HTML tags into three groups
  - Assign each group an arbitrary weight
Using on-page features (Cont.)

- Summarize then classify
  - Classify web pages based on their **summarization** [Shen et al., 2004].

- **URL**
  - A web page can be reasonably classified just based on its **URL**! [Kan, 2004; Kan and Thi, 2005]
Using on-page features (Cont.)

- Visual analysis
  - Most approaches focus on text, ignoring visual info.
  - Sometimes, it might be more expensive than using text.
  - Analyze a web page’s **visual layout**, represent the recognized parts in an adjacency graph, apply heuristic rules on the graph [Kovacevic et al., 2004].
Using features of neighbors

- Directly incorporate text from parent or child into the target page
  - It does more harm than good [Chakrabarti et al., 1998; Ghani et al., 2001; Yang et al., 2002].

- Parent and child pages == useless?
  - No! We can make it useful!
  - Using a subset of parent/child pages (not all of them)
  - Using a portion of content of parent/child pages (not full content)
Using features of neighbors (Cont.)

- Use a neighbor’s feature only if its content is similar enough to the target page [Oh et al., 2000].

- Use anchor text, the surrounding text of anchor text [Attardi et al., 1999; Furnkranz, 1999; 2001; Sun et al., 2002; Glover et al., 2002].

- Sibling pages are more useful than parent/child pages [Chakrabarti et al., 1998; Slattery & Mitchell, 2000; Qi & Davison, 2006].
Using features of neighbors (Cont.)

- **Labels** [Chakrabarti et al., 1998; Slattery & Mitchell, 2000; Calado et al., 2003]
  - Human-generated, accurate
  - Only a small portion of the web is labeled

- **Partial content** [Glover et al., 2002; Cohen 2002]

- **Full content**
Using artificial links

- Besides hyperlinks, we can create other types of links!
  - Content similarity [Kurland and Lee, 2005; 2006]
  - Pages co-occur in top results of a query
  - Co-occur + clicked [Shen et al., 2006]
  - ...

• Background
• Utilizing features of neighbors
• Using fielded features
Outline of Neighboring algorithm

- Four types of neighbors are used.
  - Parent, child, sibling, spouse.

- Information is linearly combined.
  - Weights are adjusted through experiments.

- Both page content and human labeling are considered. Human labeling is used whenever available.
A web page and its neighbors
A probability distribution vector

\[ \vec{v}_p = (v_{p,1}, v_{p,2}, \ldots, v_{p,i}, \ldots, v_{p,n}) \]

Each component \( v_{p,i} \) is the normalized probability of the page \( p \) being in the category \( c_i \).
\[ \vec{v}_i = (v_{i,1}, v_{i,2}, \cdots, v_{i,n}) \]
Neighboring pages

Page level weighting

Grouping

Parent

Child

Sibling

Spouse

Group level weighting

Topic vector of neighbors

Weighting between target page and neighbors

Final topic vector
Relative contributions of labeled pages and those unlabeled.

- Any classifier produces an approximation to the desired human labeling.
- Use human judgment whenever it is available.
- Decisions of classifiers are down-weighted.

\[ \tilde{v}_p' = \tilde{v}_p \times w(p) \]

where

\[ w(p) = \begin{cases} 1 & \text{if } p \text{ is labeled} \\ \eta & \text{if } p \text{ is not labeled} \end{cases} \quad (0 \leq \eta \leq 1) \]
Counting the multiple paths
Tuning the page level bias

- Weighted path works better than unweighted path.
- Human labeling is important.

\[ \vec{v}_p' = \vec{v}_p \times w(p) \]

where

\[ w(p) = \begin{cases} 1 & \text{if } p \text{ is labeled} \\ \eta & \text{if } p \text{ is not labeled} \end{cases} \]

(0 ≤ η ≤ 1)
Neighboring pages

Page level weighting

Grouping

Parent → Child → Sibling → Spouse

Group level weighting

Topic vector of neighbors

Weighting between target page and neighbors

Final topic vector
Bias among neighboring groups

**Diagram:**

- Parent
- Child
- Sibling
- Spouse

**Equation:**

\[
\vec{v}_{neighbors} = \beta_1 \times \frac{\sum_{p \in \text{Parent}} \vec{v}_p''}{\sum_{p \in \text{Parent}} w(p)} + \beta_2 \times \frac{\sum_{p \in \text{Child}} \vec{v}_p''}{\sum_{p \in \text{Child}} w(p)} + \beta_3 \times \frac{\sum_{p \in \text{Sibling}} \vec{v}_p''}{\sum_{p \in \text{Sibling}} w(p)} + \beta_4 \times \frac{\sum_{p \in \text{Spouse}} \vec{v}_p''}{\sum_{p \in \text{Spouse}} w(p)}
\]

\[
0 \leq \beta_i \leq 1, \quad \sum_{i=1}^{4} \beta_i = 1, \quad i = 1,2,3,4
\]
Individual contribution of the four types of neighbors

- Sibling pages contribute the most.
Weight of sibling pages

\[ \beta = (0.2, 0.2, 0.4, 0.2) \]
Combining neighbors with target page

- A combined vector: a weighted average of the topic vector of the target page and that of the neighboring pages

- The category corresponding to the major component is chosen as the classification result

\[ \vec{v} = \alpha \times \vec{v}_{\text{target}} + (1-\alpha) \times \vec{v}_{\text{neighbors}}, \quad \text{where} \ 0 \leq \alpha \leq 1 \]
$\alpha$: combining the neighbors with the target page

Diagram:

- **Accuracy**
- **$\alpha$**

Lines:
- **with labels**
- **without labels**
Experimental setup

- Dataset
  - Open Directory Project (ODP)
    - A web directory, September 2004, contains 0.6 million categories, 4.4 million web pages.
    - We used 12 out of the 17 top-level categories.
  - Training: 19,000 documents * 12 categories
  - Tuning: 500 documents * 12 categories
  - Test: 500 documents * 12 categories
  - Three disjoint sets of documents randomly selected from ODP dataset
  - Plus 6.5 million neighboring documents
Experimental setup (Cont.)

• Textual classifiers
  • A Naïve bayes classifier (*Rainbow*)
  • A SVM classifier (*SVM Light*)
Experimental result of labeled ODP dataset

- Naïve bayes: ~70%
- SVM: <80%
- K+C: <80%
- IO-bridge: <80%
- Neighboring Alg.: ~90%
- Neighboring + NB: ~90%
- Neighboring + SVM: ~90%
With and without human labeling

- Neighboring alg. still has a fairly good performance w/o labeling.
- Prior approaches can’t perform well.
Section summary

• Our Neighboring algorithm
  • is able to improve the accuracy of common textual classifiers from 70% to more than 90%.
  • can improve accuracy even when neighboring pages are unlabeled.

• While all neighbor types can contribute, sibling pages are the most important.
• Background
• Utilizing features of neighbors
• Using fielded features
Motivation

- Most existing classification work uses web page as a whole.
- Should all the content on a web page be treated the same?
  - In retrieval, BM25f [Robertson et al., 2004] is a successful fielded extension to BM25 [Robertson & Walker., 1994].
  - In classification, information from different parts of web pages should also have different importance.
  - E.g., anchor text should be of extra value.
F-Neighbor Algorithm

- A fielded extension to Neighboring Algorithm.
- Break up web pages (target page, as well as its neighbors) into text fields.
- Combine them linearly with different weights.
- Perform classification on the combined representation.
Text fields

- Title of the target page;
- Full text of the target page;
- Titles of parent, child, sibling, and spouse pages (as four separate fields);
- Full text of parent, child, sibling, and spouse pages;
- Anchor text (referring to target) on parent page;
- Surrounding text of anchor text on parent pages (referred to as “extended anchor text”).
Text representation

- Each field is represented by a TFIDF vector.
  \[ \vec{d}_{f_i}, \quad i = 1, \ldots, K \]

- The combined representation of a target document is the weighted combination of all the associated fields.
  \[ \vec{d}_{comb} = \sum_{i=1}^{K} w_{f_i} \cdot \vec{d}_{f_i} \quad \text{where} \quad \sum_{i=1}^{K} w_{f_i} = 1 \]

- The weights will be determined through experiments.
Parameter tuning

- Exploring the full K-dimensional parameter space is expensive!
- Divide and conquer!
  - Lower layer: tune weights within each neighbor type
  - Upper layer: tune weights among neighbor types
Lower-layer optimization results

- Emphasizing titles benefits all neighbor types.

- Significant improvement for parent, child, and spouse.

- Marginal benefit for target and siblings.
Section summary

- We can further improve performance based on Neighboring algorithm.

- Fielded information is useful in web page classification.
Thank you!