Proposing a New Term Weighting Scheme for Text Categorization

LAN Man
School of Computing
National University of Singapore
12th July, 2006
Outline

• Introduction
• Our Position
• Survey and Analysis: Supervised and Unsupervised Term Weighting Schemes
• Experiments
• Concluding Remarks
Introduction: Background

• Explosive growth of Internet and quick increase of textual information available
• Organizing and accessing these information in flexible ways
• **Text Categorization** is the task of classifying natural language documents into a predefined set of semantic categories
Introduction:
Applications of TC

- **Categorize** web pages by topic (the directories like Yahoo!);
- **Customize** online newspapers to different labels according to a particular user’s reading preferences;
- **Filter** spam emails and forward incoming emails to the target expert by content;
- **Word sense disambiguation** is also taken as a text categorization task once we view word occurrence contexts as documents and word sense as categories.
Introduction

• Two Subtasks in Text Categorization:
  – 1. Text Representation
  – 2. Construction of Text Classifier
Introduction: Construction of Text Classifier

• Approaches to learn a classifier:
  – No more than 20 algorithms
  – Borrowed from *Information Retrieval*: Rocchio
  – *Machine Learning*: SVM, kNN, decision Tree, Naïve Bayesian, Neural Network, Linear Regression, Decision Rule, Boosting, etc.

• SVM is the best for TC.
Introduction:
Text Representation

• Various text format, such as DOC, PDF, PostScript, HTML, etc.

• Can Computer read them like us? No.

• Convert them into a compact format in order to be recognized and categorized for classifiers or a classifier-building algorithm in a computer.

• This indexing procedure is also called text representation.
Introduction:
Vector Spaces Model

Texts are vectors in the term spaces.
Assumption: Documents that are “close together” in space are similar in meaning.
Introduction: Text Representation

• Two main issues in Text Representation:
  – 1. What a term is
  – 2. How to assign a weight for a term

• 1. What a term is
  – Sub-word level – syllables
  – Word level – single token
  – Multi-word level – phrases, sentences, etc
  – Syntactic and semantic – sense (meaning)

• Word level is best
Introduction: Text Representation

• 2. The weight of a term represents how much the term contributes to the semantic of document.

• How to weight a term (Our current focus)
  – Simplest method – binary
  – Most popular method – tf.idf
  – Combination with the linear classifier
  – Combination with information-theory metrics or statistics method – tf.chi2, tf.ig, tf.gr, tf.rf (we presented)
Our Position

- Leopold (2002) stated that text representation dominates the performance of text categorization rather than the kernel functions of SVMs.
- Little room to improve the performance from the algorithm aspect:
  - Excellent algorithms are few.
  - The rationale is inherent to each algorithm and the method is usually fixed for one given algorithm
  - Tuning the parameter has limited improvement
Our Position

• Analyze term’s discriminating power for text categorization
• Present a new term weighting method
• Compare supervised and unsupervised term weighting methods
• Investigate the relationship between different term weighting methods and machine learning algorithms
Survey and Analysis

• Salton’s three considerations for term weighting:
  – 1. **Term occurrence**: binary, tf, ITF, log(tf)
  – 2. **Term’s discriminative power**: idf

  *Note*: chi^2, ig (information gain), gr (gain ratio), mi (mutual information), or (Odds Ratio), etc.
  – 3. **Document length**: cosine normalization, linear normalization
Survey and Analysis

• “Text categorization” is a form of \textit{supervised} learning

• The prior information on the membership of training documents in predefined categories is useful
  – Feature selection
  – \textit{Supervised learning for classifier}
Survey and Analysis

• Supervised term weighting methods
  – Use the prior information on the membership of training documents in predefined categories

• Unsupervised term weighting methods
  – Does not use.
  – binary, \( tf \), \( \log(1+tf) \), ITF
  – Most popular is \( tf.idf \) and its variants: \( \log tf.idf \), \( tf.idf-prob \)
Survey and Analysis: Supervised Term Weighting Methods

• 1. Combined with information-theory functions or statistic metrics
  – such as chi2, information gain, gain ratio, Odds ratio, etc.
  – Used in feature selection step
  – Select the most relevant and discriminating features for the classification task, that is, the terms with higher feature selection scores
  – The results are inconsistent and/or incomplete
Survey and Analysis: Supervised Term Weighting Methods

• 2. Interaction with supervised linear text classifier
  – Linear SVM and Perceptron
  – Text classifier selects the positive test documents from negative test documents by assigning different scores to the test samples, these scores are believed to be effective in assigning more appropriate weights to terms
Survey and Analysis: Analysis of term’s discriminating power

- Assume they have same \( tf \) value. \( t_1, t_2, t_3 \) share the same \( idf \) value; \( t_4, t_5, t_6 \) share same \( idf \) value.

- Clearly, the six terms contribute differently to the semantic of documents.
Survey and Analysis:
Analysis of term’s discriminating power

\[ idf = \log \left( \frac{N}{b+c} \right) \]

\[ idf - prob = \log \left( \frac{a+d}{b+c} \right) \]

\[ \chi^2 = N \cdot \frac{(a \cdot d - b \cdot c)^2}{(a+d)(b+c)(a+b)(c+d)} \]

\[ rf = \log \left( 2 + \frac{b}{c} \right) \]
Survey and Analysis:
Analysis of term’s discriminating power

• **Case 1.** t1 contributes more than t2 and t3; t4 contributes more than t5 and t6.

• **Case 2.** t4 contributes more than t1 although $idf(t4) < idf(t1)$.

• **Case 3.** for t1 and t3, if $a(t1)=d(t3)$, $b(t1)=c(t3)$, $chi2(t1)=chi2(t3)$ but t1 may contribute more than t3.
Survey and Analysis: Analysis of term’s discriminating power

\[ rf = \log(2 + \frac{b}{\max(1, c)}) \]

- \( rf \) - relevance frequency
- The \( rf \) is only related to the ratio of \( b \) and \( c \), not involve \( d \)
- The base of \( \log \) is 2.
- in case of \( c = 0, c = 1 \)
Survey and Analysis:
Analysis of term’s discriminating power

<table>
<thead>
<tr>
<th>rf</th>
<th>c</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.1</td>
<td>1</td>
<td>1.58</td>
<td>2</td>
<td>2.32</td>
<td>2.58</td>
<td>2.81</td>
<td>3</td>
</tr>
<tr>
<td>2</td>
<td></td>
<td>1</td>
<td>1.32</td>
<td>1.58</td>
<td>1.81</td>
<td>2</td>
<td>2.17</td>
<td>2.32</td>
</tr>
<tr>
<td>3</td>
<td></td>
<td>1</td>
<td>1.22</td>
<td>1.42</td>
<td>1.58</td>
<td>1.74</td>
<td>1.87</td>
<td>2</td>
</tr>
<tr>
<td>4</td>
<td></td>
<td>1</td>
<td>1.17</td>
<td>1.32</td>
<td>1.46</td>
<td>1.58</td>
<td>1.70</td>
<td>1.81</td>
</tr>
</tbody>
</table>
Survey and Analysis:
Comparison of $idf$, $rf$ and $chi^2$ value of four features in two categories of Reuters Corpus

<table>
<thead>
<tr>
<th>feature</th>
<th>Category: 00_acq</th>
<th>Category: 03_earn</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$idf$</td>
<td>$rf$</td>
</tr>
<tr>
<td>Acquir</td>
<td>3.553</td>
<td>4.368</td>
</tr>
<tr>
<td>Stake</td>
<td>4.201</td>
<td>2.975</td>
</tr>
<tr>
<td>Payout</td>
<td>4.999</td>
<td>1</td>
</tr>
<tr>
<td>dividend</td>
<td>3.567</td>
<td>1.033</td>
</tr>
</tbody>
</table>
Experimental Methodology:
Methodology: eight term weighting schemes

<table>
<thead>
<tr>
<th>Methods</th>
<th>Denotation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unsupervised</td>
<td>binary</td>
<td>0: absence, 1: presence</td>
</tr>
<tr>
<td>Term Weighting</td>
<td>tf</td>
<td>Term frequency alone</td>
</tr>
<tr>
<td></td>
<td>tf.idf</td>
<td>Classic tf.idf</td>
</tr>
<tr>
<td>Supervised</td>
<td>tf.rf</td>
<td>tf * rf</td>
</tr>
<tr>
<td>Term Weighting</td>
<td>rf</td>
<td>binary * rf</td>
</tr>
<tr>
<td></td>
<td>tf.chi2</td>
<td>Chi square</td>
</tr>
<tr>
<td></td>
<td>tf.ig</td>
<td>ig: information gain</td>
</tr>
<tr>
<td></td>
<td>tf.or</td>
<td>or: Odds Ratio</td>
</tr>
</tbody>
</table>
Experimental Methodology

• Methodology
  – 8 commonly-used term weighting schemes
  – 2 benchmark data collections
    • Reuters News Corpus: skewed category distribution
    • 20 Newsgroups Corpus: uniform category distribution
  – 2 popular machine learning algorithms
    • SVM and kNN
  – Micro- and Macro- averaged F1 measure
  – Significance tests
Experimental Results:

1. Results on Reuters News Corpus using SVM
Experimental Results:
2. Results on Reuters News Corpus using kNN
Experimental Results:
3. Results on 20Newsgroups Corpus using SVM
Experimental Results:
4. Results on 20Newsgroups Corpus using kNN
### Experimental Results: McNemar’s Significance Tests

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Corpus</th>
<th>#_fea</th>
<th>Significance Test Results</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVM</td>
<td>R</td>
<td>15937</td>
<td>(tf.rf, tf, rf) &gt; tf.idf &gt; (tf.ig, tf.chi2, binary) &gt;&gt; tf.or</td>
</tr>
<tr>
<td>SVM</td>
<td>20</td>
<td>13456</td>
<td>(rf, tf.rf, tf.idf) &gt; tf &gt;&gt; binary &gt;&gt; tf.or &gt;&gt; (tf.ig, tf.chi2)</td>
</tr>
<tr>
<td>kNN</td>
<td>R</td>
<td>405</td>
<td>(binary, tf.rf) &gt; tf &gt;&gt; (tf.idf, rf, tf.ig) &gt; tf.chi2 &gt;&gt; tf.or</td>
</tr>
<tr>
<td>kNN</td>
<td>20</td>
<td>494</td>
<td>(tf.rf, binary, tf.idf, tf) &gt;&gt; rf &gt;&gt; (tf.or, tf.ig, tf.chi2)</td>
</tr>
</tbody>
</table>
Experimental Discussion:
Effects of feature set size on algorithms

• For SVM, almost all methods achieved the best performance when in putting the full vocabulary (13000-16000 features)
• For kNN, the best performance achieved at a smaller feature set size (400-500 features)
• Possible reason: different noise resistance
Experimental Discussion:
Summary of different methods

• Generally, supervised and unsupervised methods have not shown universally consistent performance

• Except for $tf.rf$, it shows the best performance consistently

• $rf$ alone, shows a comparable performance to $tf.rf$ except for on Reuters using kNN
Experimental Discussion:
Summary of different methods

- Specifically, the three typical supervised methods based on information theory, \texttt{tf.chi2} (chi square), \texttt{tf.ig} (information gain) and \texttt{tf.or} (Odds ratio), are the worst methods.
Experimental Discussion:
Summary of different methods

• The three unsupervised methods, \( tf \), \( tf.idf \) and \( binary \), show mixed results with respect to each other.
  – Example 1, \( tf \) better than \( tf.idf \) on Reuters using SVM and but it is the other way round on 20Newsgroups using SVM.
  – Example 2, these three are comparable to each other on 20 Newsgroups using kNN
Experimental Discussion:
Summary of different methods

• 1. kNN favors *binary* while SVM does not
• 2. The good performance of *tf.idf* on 20Newsgroups using SVM and kNN may be attributed to the natural property of corpus, uniform category distribution
• 3. *tf* outperforms many other methods although it does not perform comparably to *tf.rf*
Concluding Remarks 1

• Not all supervised term weighting methods have a consistent superiority over unsupervised term weighting methods.

• Specifically, three supervised methods based on information theory, i.e. $tf.chi^2$, $tf.ig$ and $tf.or$, perform rather poorly in all experiments.
Concluding Remarks 2

• On the other hand, newly proposed supervised method, *tf.rf* achieved the best performance consistently and outperforms other methods substantially and significantly.
Concluding Remarks 3

• Neither *tf.idf* nor *binary* shows consistent performance.
  – Specifically, *tf.idf* is comparable to *tf.rf* on the uniform category distribution corpus
  – *binary* is comparable to *tf.rf* on the kNN-based text classifier
Concluding Remarks 4

• $tf$ does not perform as well as $tf.rf$, but it performs consistently well and outperforms other methods consistently and significantly
Concluding Remarks

• We suggest *tf.rf* be used as term weighting method for TC task.

• The observations above are made based on the controlled experiments
Future Work

• Can we observe the similar results on more general experimental settings, such as different learning algorithms, different performance measures and other benchmark collections?

• Term weighting is the most basic component of text preprocessing methods, can we integrate it into various text mining tasks, such as information retrieval, text summarization, etc.?