

Hybrid DIAAF/RS: Statistical Textual Feature Selection for Language-Independent Text Classification

Yanbo J. Wang¹, Fan Li¹, **Frans Coenen**², Robert Sanderson³, and Qin Xin⁴

¹Information Management Center, China Minsheng Banking Corporation Ltd., China

²Department of Computer Science, University of Liverpool, UK

³Los Alamos National Laboratory, USA

⁴Simula Research Laboratory, Norway

icdm

Industrial Conference on Data Mining

July 12–14, 2010, Berlin, Germany

Outline

- **Background**
 - Text Classification
 - Textual Data Pre-processing
 - Data Classification
 - Summary of Background
- **Motivation**
 - Language-Independent Text Classification
 - Availability of Language-Independent “Bag of Words” & “Bag of Phrases”
 - Summary of Motivation
- **Language-Independent Feature Selection**
 - Previous Studies
 - Proposed “Hybrid DIAAF/RS” Approach
- **Experimental Results**
 - Text Collections
 - Setting of Experiments
 - Classification Accuracy
- **Conclusions & Future Work**

Background



Text Classification

- What is Text Classification (TC)?
 - TC is the task of assigning one or more predefined categories to natural language text documents, based on their contents.
 - Early studies of TC can be dated back to the early 1960s.
 - Broadly speaking, TC studies can be separated into two divisions: single-label vs. multi-label.
 - With regard to the single-label TC, three distinct approaches can be identified: one-class TC, binary TC & multi-class TC.
 - **Our study is concerned with the single-label multi-class TC.**
 - The overall TC process can be divided into two stages: data pre-processing & data classification.

Textual Data Pre-processing

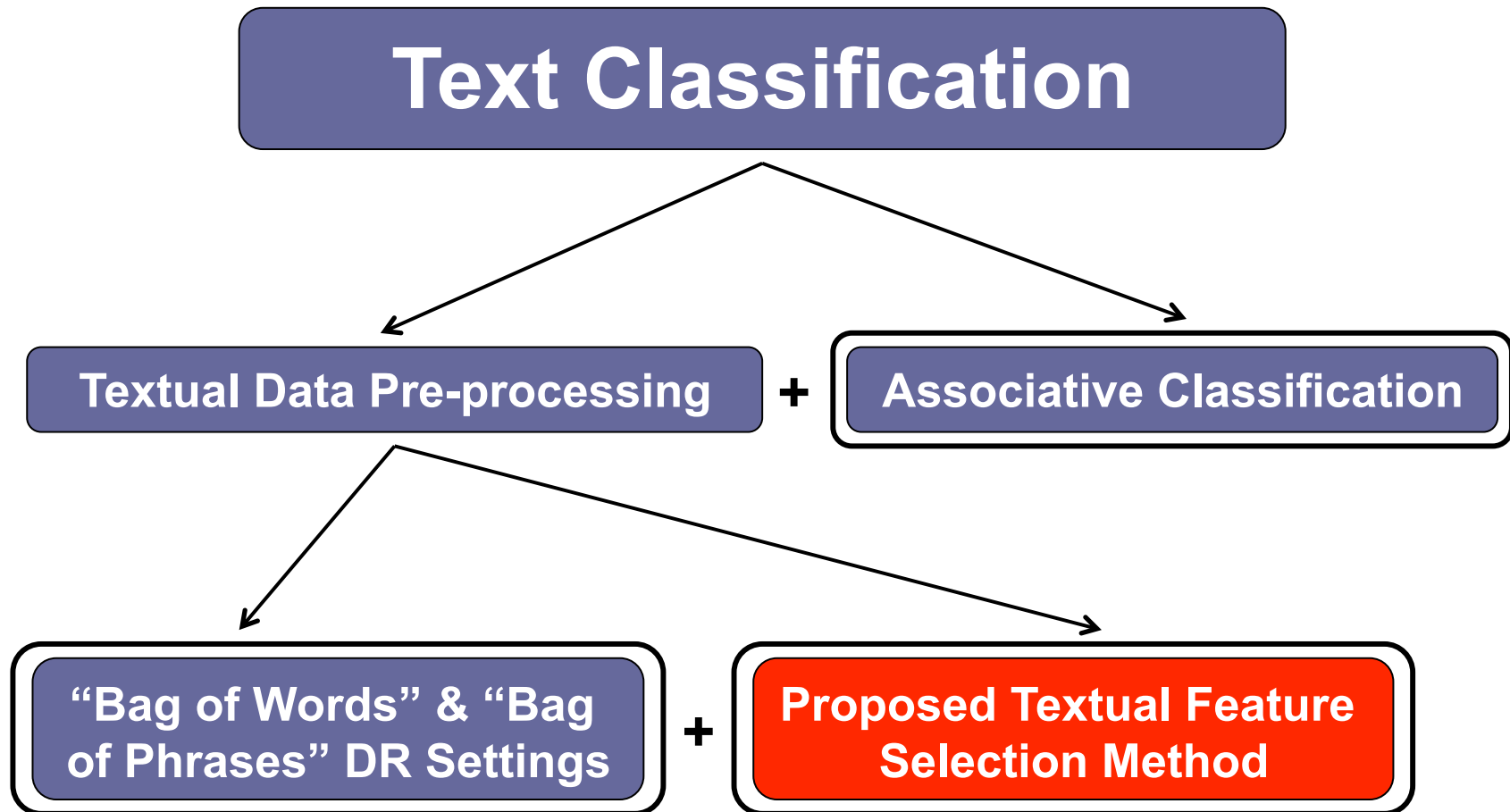
- Textual data pre-processing comprises: (i) Document-base Representation (DR) & (ii) Textual Feature Selection (TFS).
 - During the **DR** stage the input data is translated into an application oriented data structure.
 - In TC, the “bag of *” approach or vector space model is popular.
 - The “*” symbol stands for the type of text-units, i.e. words, word-sets, phrases, concepts, etc.
 - **In our study, we select to use both the “bag of words” and the “bag of phrases” representations.**
 - TFS aims to identify the most significant text-features (i.e. *key words/ phrases*) in the document-base.
 - **In this study, we propose a statistical textual feature selection method.**

Data Classification

- ❑ Mechanisms on which data classification algorithms have been based include: *decision trees, naive bayes, k-nearest neighbour, support vector machine, association rules, genetic algorithm, neural networks, etc.*
- ❑ Previous studies indicate that in many cases data classification based on *association rules* (i.e. **associative classification**) offers good classification accuracy.
- ❑ Associative classification have the following advantages:
 - They are fast during both the training and categorisation phases, especially when handling large document-bases; and
 - Such text classifiers can be read, understood and modified by humans, so that users are able to see why the classification predictions have been made.
- ❑ **Thus, an associative classification approach is adopted in this study.**

Summary of Background

In our study:



Motivation



Language-Independent TC

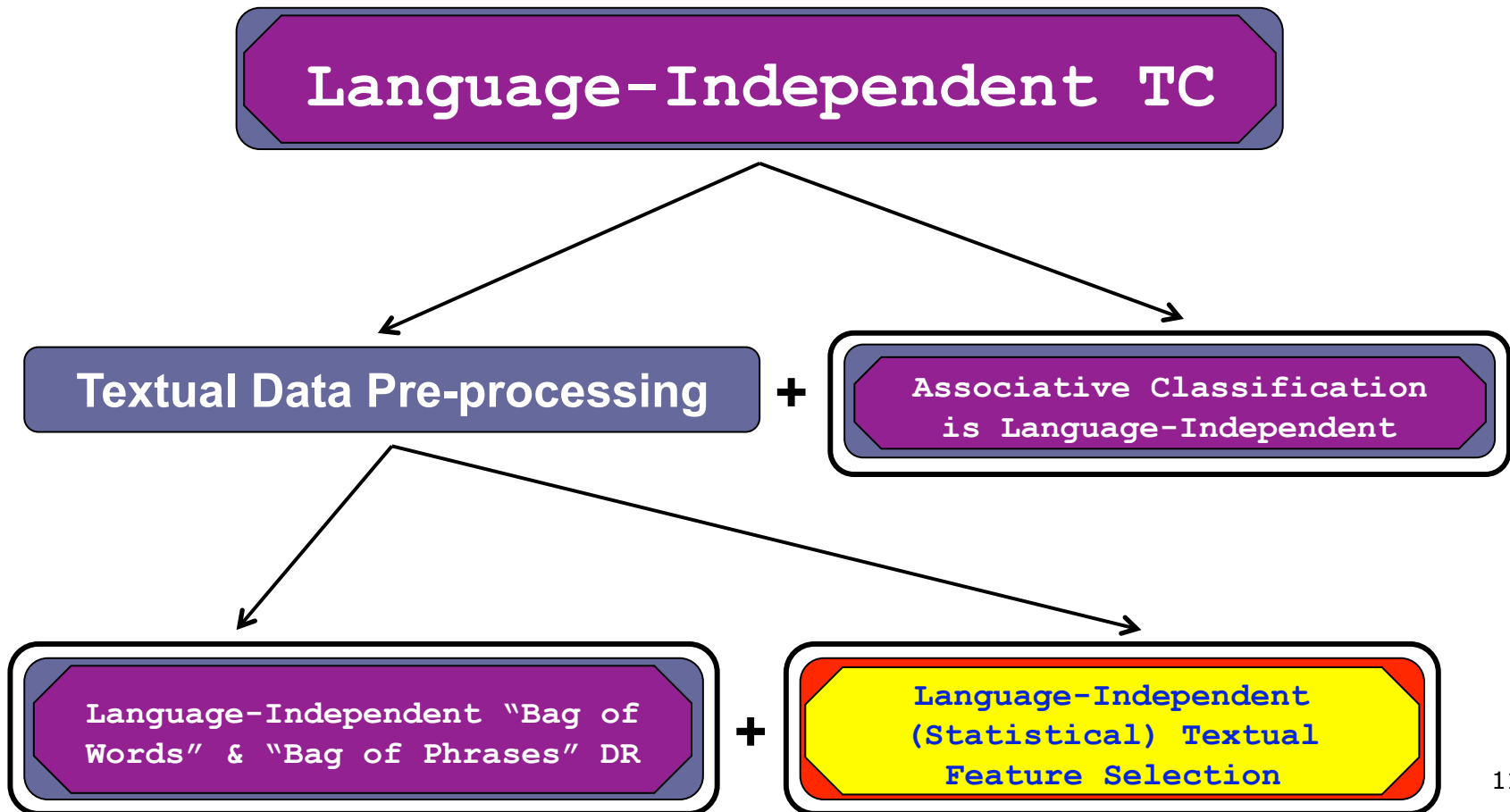
- Many textual data pre-processing mechanisms use language-dependent ideas to identify *key* words and phrases (e.g. *stop word lists*, *synonym lists*, *stemming*, *part-of-speech tagging*, *word sense disambiguation*, etc).
- These techniques operate well but are designed with particular target languages in mind.
- They are therefore not generally applicable to all languages (e.g. *Chinese*, *Arabic*, *Spanish*, etc).
- **We are interested in language-independent TC, which aims to address the above issues.**
- Such text classifier can also be applied to *cross-lingual*, *multi-lingual* and/or *unknown lingual* textual data collections.

Language-Independent “Bag of Words” & “Bag of Phrases”

- Some definitions
 - **Words:** Words in a document-base are defined as *continuous sequences of alphabetic characters* delimited by non-alphabetic characters.
 - **Noise Words (N):** *Common* and *rare* words are collectively defined to be *noise* words in a document-base.
 - **Potential Significant Words:** A potential significant word, also referred to as a *key* word/feature, is a non-noise word.
 - **Significant Words (G):** The first k words for each predefined class, selected from the ordered list of potential significant words. **(This is referred to as the language-independent “Bag of Words”).**
 - **Ordinary Words (O):** Other non-noise words that have not been selected as significant words (Pot. Sig. Wrds. = Sig. Wrds. Union Ord.Wrds.)
 - **Stop Marks (S):** Not actual words but six key punctuation marks (, . : ; ! and ?). All other non-alphabetic characters are ignored.
- Language-Independent “Bag of Phrases” Generation
 - This approach is named as *DelSNcontGO*: **phrases are Delimited by stop marks (S) or noise words (N), and (as phrase *contents*) made up of sequences of one or more significant words (G) and ordinary words (O); sequences of ordinary words delimited by stop marks or noise words that do not include at least one significant word (in the contents) are ignored.**

Summary of Motivation

In our study:



Language-Independent Feature Selection



Previous Studies

- Previous language-independent (statistical) textual feature selection mechanisms below. Each is used to calculate how significantly a word/feature (u_h) determines a predefined text-category (C_i) in a document-base (D_R).

Name	Probabilistic Form	Calculation	Description
DIA (Darmstadt Indexing Approach) Association Factor (DIAAF)	$\text{diaaf_score}(u_h, C_i) = P(C_i u_h)$	$\frac{\text{count}(u_h \in C_i)}{\text{count}(u_h \in D_R)}$	This score expresses the proportion of the word's occurrence in the given class divided by the word's document-base occurrence.
Relevancy Score (RS)	$\text{rs_score}(u_h, C_i) = \log((P(u_h C_i) + d) / (P(u_h \neg C_i) + d))$	$\log \left[\frac{\frac{\text{count}(u_h \in C_i)}{ C_i } + d}{\frac{\text{count}(u_h \in (D_R - C_i))}{ D_R - C_i } + d} \right]$ <p>where d is a constant damping factor</p>	This score expresses the proportion (in a logarithmic term) of the frequency with which the word occurs in documents of the given class divided by the word's frequency in the complement of the class.
Mutual Information (MI)	$\text{mi_score}(u_h, C_i) = \log(P(u_h C_i) / P(u_h))$	$\log \left[\frac{\frac{\text{count}(u_h \in C_i)}{ C_i }}{\frac{\text{count}(u_h \in D_R)}{ D_R }} \right]$	This score expresses the proportion (in a logarithmic term) of the frequency with which the word occurs in documents of the given class divided by the word's document-base frequency.

Hybrid DIAAF/RS

- In this study, we propose a hybrid statistical textual feature selection approach that integrates the DIAAF and the RS mechanisms, namely “Hybrid DIAAF/RS”.
- The rationale of the “Hybrid DIAAF/RS” approach is that a significant textual feature (term) with respect to a particular text class should have:
 - A high ratio of the class term support (document frequency) to the document-base term support; and/or
 - A low ratio of the class term support of non-appearance to the document-base term support of non-appearance.
- The calculation of this proposed approach can be shown as follows.

Name	Probabilistic Form	Calculation
DIA Association Factor based Relevancy Score (DIAAF-RS)	$\text{diaaf-rs_score}(u_h, C_i)$ $= \log\left(\frac{\mathbf{P}(C_i u_h) + d}{\mathbf{P}(C_i \neg u_h) + d}\right)$	$\log \left[\frac{\frac{\text{count}(u_h \in C_i)}{\text{count}(u_h \in D_R)} + d}{\frac{\text{count}(\neg u_h \in C_i)}{\text{count}(\neg u_h \in D_R)} + d} \right]$ <p>where d is a constant damping factor</p>

Experimental Results



Text Collections (1)

- We evaluate the proposed “Hybrid DIAAF/RS” approach with respect to the accuracy of classification, using three well-known text collections:
 - **Usenet Articles (20 Newsgroups)**
 - **Reuters-21578**
 - **MedLine-OHSUMED**
- In our experiments, five individual document-bases (textual datasets) were extracted from above text collections.

Text Collections (2)

1. **20NG.D10000.C10:** Document base describing first 10 groups of documents (10,000 documents in 10 classes) from the 20 Newsgroups collection.
2. **20NG.9997.C10:** This document-base comprises the rest of the 20 Newsgroups collection (9,997 documents in 10 classes).
3. **Reuters.D6643.C8:** We first of all selected the top-10 most populous classes from Reuters-21578. Then we removed those multi-labelled and/or non-text documents from each class. As a consequence, 2 of the 10 classes were empty. The resulting Reuters.D6643.C8 document-base comprises 6,643 documents in 8 classes.
4. **OHSUMED.D6855.C10:** We select the top-100 most populous classes from this collection. We then selected 20 target-classes from these 100 classes by hand, so as to exclude obvious super-and-sub class-relationships. Finally, we remove documents which were either multi-labelled or without a proper content from each target-class. We randomly separate the 20 target-classes into two parts, the first document-base created here comprises 6,855 documents in 10 classes.
5. **OHSUMED.D7427.C10:** The second (part) comprises 7,427 documents in 10 classes.

Setting of Experiments

- Experiments designed to evaluate the proposed “Hybrid DIAAF/RS” textual feature selection approach, in comparison with previous mechanisms (i.e. DIAAF, RS and MI), with regard to both the (language-independent) “bag of words” and the *DelSNcontGO* “bag of phrases” approaches.
- Evaluation conducted using the *TFPC (Total From Partial Classification)* associative classifier although any other similar classifier could equally well have been used
- Accuracy figures were obtained using *Ten-fold Cross Validation (TCV)*.
- *support threshold value* = 0.1% (for TFPC)
- *confidence threshold value* = 35% (for TFPC)
- *lower noise threshold value* = 0.2% (for *rare* words)
- *upper noise threshold value* = 20% (for *common* words)
- In both RS and Hybrid DIAAF/RS, 0 was used as the *constant damping factor* value (*d*).
- The parameter *K* (*maximum number of selected final key features*) was chosen to be 1,000. (Note: the value of *K* was changed to be 900 for OHSUMED document-bases for the “bag of phrases” version because 1,000 *key* features generated more than 2^{15} *key* phrases; for operational reasons the TFPC associative classifier limits the total number of identified attributes (significant phrases) to 2^{15}).

Classification Accuracy

	Document-bases	DIAAF	RS	MI	Hybrid DIAAF/RS
BAGOFW	20NG.D10000.C10	76.72	76.72	76.72	<u>77.01</u>
	20NG.D9997.C10	80.61	80.61	80.61	<u>80.75</u>
	Reuters.D6643.C8	85.40	86.34	86.56	<u>86.81</u>
	OHSUMED.D6855.C10	77.54	<u>79.28</u>	79.27	79.17
	OHSUMED.D7427.C10	<u>78.97</u>	77.21	77.45	78.12
	Average Accuracy	79.85	80.03	80.12	<u>80.37</u>
BAGOFW	20NG.D10000.C10	76.96	76.96	76.96	<u>77.32</u>
	20NG.D9997.C10	81.72	81.72	81.72	<u>82.09</u>
	Reuters.D6643.C8	87.63	87.94	87.99	<u>88.53</u>
	OHSUMED.D6855.C10	79.20	<u>80.16</u>	80.04	80.03
	OHSUMED.D7427.C10	<u>78.24</u>	75.80	75.75	77.07
	Average Accuracy	80.75	80.52	80.49	<u>81.01</u>
	# of Best Accuracies	2	2	0	<u>6</u>

Classification Accuracy (continue...)

- The number of instances of best classification accuracies obtained throughout the 5 document-bases, with regard to both the “bag of words” and the “bag of phrases” settings, can be ranked in order as follows.
- The average accuracy of classification throughout the 5 document-bases in the “bag of words” DR setting can be ranked in order as follows.
- The average accuracy of classification throughout the 5 document-bases in the “bag of phrases” DR setting can be ranked in order as follows.

1	Hybrid DIAAF/RS	6
2	DIAAF	2
2	RS	2
4	MI	None

1	Hybrid DIAAF/RS	80.37
2	MI	80.12
3	RS	80.03
4	DIAAF	79.85

1	Hybrid DIAAF/RS	81.01
2	DIAAF	80.75
3	RS	80.52
4	MI	80.49

- These results demonstrate the good performance and the stability of the “Hybrid DIAAF/RS” approach.

Conclusions & Future Work



Conclusions & Future Work

- An alternative language-independent textual feature selection technique (Hybrid DIAAF/RS), which integrates the ideas of DIAAF and RS, has been introduced.
- From the experimental results, it can be seen that the proposed “Hybrid DIAAF/RS” approach outperforms existing mechanisms with respect to language-independent “bag of words” and the *DelSNcontGO* “bag of phrases” approaches.
- Our study improves the performance of language-independent TC.
- The results presented in this study corroborate previously reported results that the TC problem can be solved, with good classification accuracy, in a language-independent manner.

The End



Thank You!