The 2008 DaWaK Conference, Turin, Italy

Document-base Extraction for Single-label Text Classification

Yanbo J. Wang, Robert Sanderson, Frans Coenen, and Paul Leng

The University of Liverpool

September 2008

MOTIVATION

- We are interested in data preparation, particularly document-base extraction, for single-label text classification.

1. Many text collections (textual data-sources) are available for text mining research: Reuters-21578, Usenet Articles (20 Newsgroups), MedLine-OHSUMED, etc.
2. These data-sources in their original form have some limitations. Hence it is difficult to run TC experiments using the original form of these text collections (especially when dealing with the single-label TC problem).
3. Our study aims to address the data-source limitations, and automatically extract "*qualified*" document-bases (subsets) from a given text collection that can be used more effectively and more reliably in single-label TC experiments.

DATA-SOURCE LIMITATIONS

- With regard to single-label TC, most textual data-sources, in their original form, have at least one of the following limitations:

1. The overall volume of textual data is too large for ease of experimentation;
2. There are many predefined classes involved;
3. Most of the classes consist of only a very few documents;
4. Some documents are labelled with a single class whereas others have multiple classes; and
5. There are documents found without any actual textual content, i.e. a document containing less than *q* recognized words, where *q* is usually a small constant.

PREVIOUS WORK IN DOCUMENT-BASE EXTRACTION

- The Reuters-21578 Collection

Description of Reuters-21578:

Comprises 21,578 documents collected from the Reuters newswire with 135 predefined classes.

13,476 documents are labelled with at east one class; 7,059 have no class; and 1,043 documents have their class-label as "*bypass*".

Within these 13,476 classified documents, 2,308 appear to have a class but on further investigation that class turns out to be spurious.

This leaves 11,168 documents, of which 9,338 are single-labelled and 1,830 are multi-labelled.

The RE.D10247.C10 Document-base:

There are total 135 predefined classes in Reuters-21578. Many TC studies, e.g. (Zaiane and Antonie, 2002) and (Li and Liu, 2003), have used only the 10 most populous classes for their experiments and evaluations.

There are 68 classes that consist of fewer than 10 documents, and many others consist of fewer than 100 documents. The extracted document-base, suggested b y Zaiane and Antonie (also Li and Liu) comprises 10,247 documents with 10 classes.

However RE.D10247.C10 includes multi-labelled documents that are inappropriate for a single-label TC environment.

The Reuters_100 Document-base:

Deng et al. (2002) introduce the Reuters_100 document-base that comprises 8,786 documents with 10 classes.

Deng et al. assign "*one document (to) one category and adopt categories that contain training documents (of) more than 100*".

However, many documents in Reuters_100 may be found without a "*proper*" text-content — the document contains less than q recognized words, where q is usually a small constant (20 in our study).

- The Usenet Articles

Description of Usenet Articles: Compiled by Lang (1995) from 20 different newsgroups and often referred to as the “20 Newsgroups” text collection.

Each newsgroup represents a predefined class. There are exactly 1,000 documents per class with one exception that the class “soc.religion.christian” contains 997 documents only.

The structure of the “20 Newsgroups” collection is relatively consistent — every document within this collection is labelled with one class only and almost all documents (higher than 95% of all documents) have a proper text-content ($q \geq 20$).

The NG.D10000.C10 and NG.D9997.C10 Document-bases:

Deng et al. (2002) divide the entire “20 Newsgroups” into two non-overlapping and (almost) equally sized document-bases covering 10 classes each: NG.D10000.C10 and NG.D9997.C10.

Other Previously Extracted “20 Newsgroups” Document-bases:

Wu et al. (2002) use four smaller document-bases extracted from the “20 Newsgroups” collection: NG.Comp.D5000.C5, NG.Rec.D4000.C4, NG.Sci.D4000.C4, and NG.Talk.D4000.C4.

- The MedLine-OHSUMED Collection

Description of MedLine-OHSUMED: Collected by Hersh et al. (1994).

Consists of 348,566 records relating to 14,631 predefined MeSH (Medical Subject Headings) categories (classes).

The OHSUMED collection accounts for a subset of the MedLine text collection for 1987 to 1991.

Characteristics of OHSUMED include: (i) many multi-labelled documents; (ii) total of 14,631 named classes arranged in hierarchies; and (iii) the text-content of each document comprises either a title on its own (without a text-content), or a "*title-plus-abstract*" (with a text-content) from various medical journals.

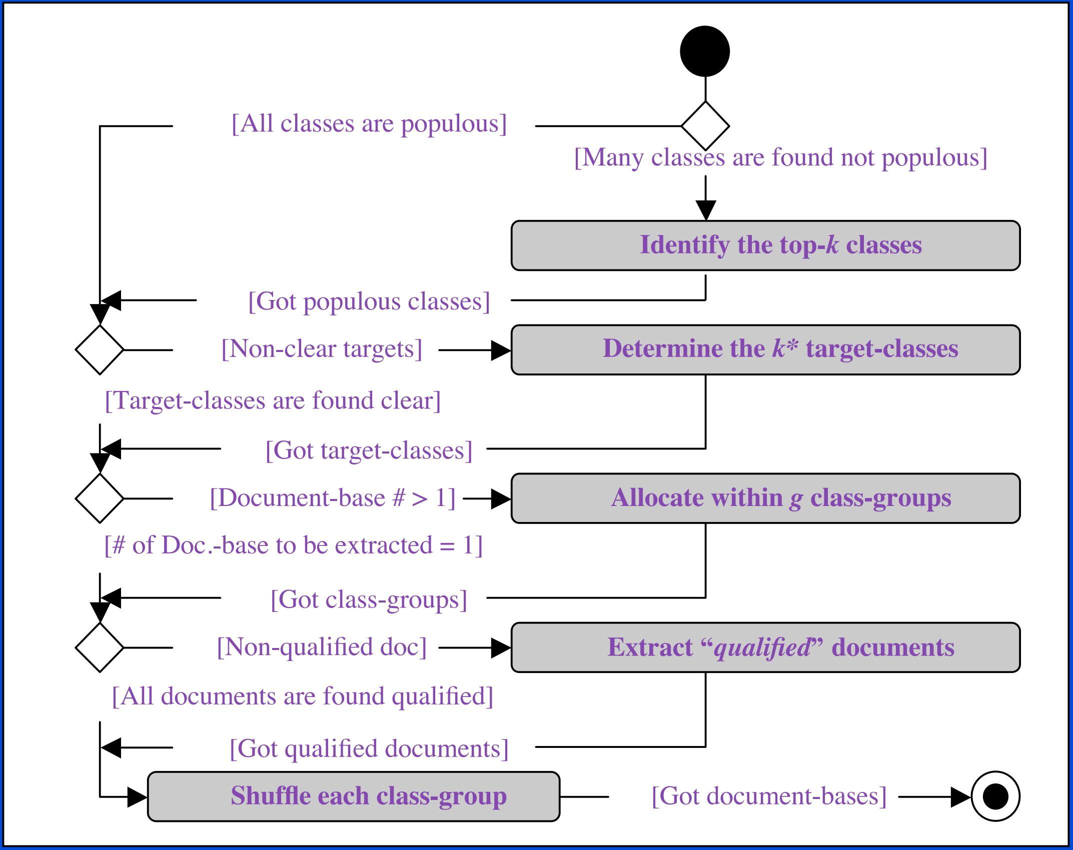
The OH.D20000.C23 Document-base: With the goal of investigating the multi-label TC problem, Joachims (1998) used the first 10,000 title-plus-abstracts texts of the 50,216 documents for 1991 as the training instances, and the second 10,000 such documents as the test instances. This defines the OH.D20000.C23 document-base, in which the classes are 23 MeSH “diseases” categories. Each document may be labelled with more than one class.

The OH.Maximal Document-base: Zaiane and Antonie (2002) introduced the OH.Maximal document-base, which consists of all OHSUMED classes incorporating all 233,445 title-plus-abstract documents.

Both the above are not well suited to single-label TC investigation.

PROPOSAL

- In this study we propose a standard textual data preparation approach that automatically extracts qualified document-bases from a given large textual data-source. The entire process of the proposed approach consists of five component-functions (states).


[All classes are populous]
[Many classes are found not populous]
Identify the top-*k* classes
[Got populous classes]
[Non-clear targets]
Determine the *k** target-classes
[Target-classes are found clear]
[Got target-classes]
[Document-base # > 1]
Allocate within *g* class-groups
[# of Doc.-base to be extracted = 1]
[Got class-groups]
[Non-qualified doc]
Extract "*qualified*" documents
[All documents are found qualified]
[Got qualified documents]
Shuffle each class-group
[Got document-bases]

1. Top-k Class Identification: Given a text collection there may be many (hundreds, sometimes thousands or even more) predefined classes. However, many of them may be assigned to only one or very few (10) documents. Hence it is appropriayte to identify the *k* most populous classes.

2. Target Class Determination: Given a set of documents, some classes may be within a taxonomy-like form (sharing a super-and-sub class-relationship). Note that al documents that are included in a predefined (sub) class, are considered to be also involved in its super-class. Hence, retaining both super-and-sub classes within a created document-base would cause a conflict when running a single-label TC experiments — each single document should not be assigned more than one class. With regard to this super-and-sub class-relationship problem, a smaller group of *k** target-classes are suggested to be further extracted from the *k* most populous classes.

3. Class Group Allocation: This phase aims to equally and randomly allocate these *k** target classes into *g* non-overlapping class-groups, where *g* is a small constant (integer) defined by the user.

4. Qualified Document Extraction: In this stage, we extract all "*qualified*" documents. We define a qualified document as a document that (i) belongs to only one predefined class; and (ii) consists of at least *q* recognized words.

5. Document Shuffle: Finally, we "shuffle" these qualified documents within each class-group, and construct a document-base. Note that when investigating single-label TC, especially the multi-class problem — it simultaneously deal with all given categories and assign the most appropriate category to each "unseen" document, the cross-validation procedure is suggested to be addressed in a further training-and-test experimental phase. The cross-validation procedure usually requires inputting a sufficiently shuffled document-base, where documents sharing a common predefined class should be evenly and dispersedly distributed within the entire document-base.

EXAMPLE GENERATIONS

- Based on the proposed document-base extraction approach, we generate four document-bases regarding the case of single-label multi-class TC, where one is extracted from Reuters-21578, two from “20 Newsgroups”, and another one from MedLine-OHSUMED.

Class	# of documents	Class	# of documents
acq	2,108	interest	216
crude	444	money	432
earn	2,736	ship	174
grain	108	trade	425

Reuters-21578 Document-base Description (RE.D6643.C8)

Class	# of documents	Class	# of documents
amino_acid_sequence	333	kidney	871
blood_pressure	635	rats	1,596
body_weight	192	smoking	222
brain	667	tomography,_x-ray_computed	657
dna	944	united_states	738

OHSUMED Document-base Description (OH.D6855.C10)

(a) NG.D9482.C10		(b) NG.D9614.C10	
Class	# of documents	Class	# of documents
comp.windows.x	940	comp.graphics	919
rec.motorcycles	959	comp.sys.mac.hardware	958
talk.religion.misc	966	rec.sport.hockey	965
sci.electronics	953	sci.crypt	980
alt.atheism	976	sci.space	977
misc.forsale	861	talk.politics.guns	976
sci.med	974	comp.os.ms-windows.misc	928
talk.politics.mideast	966	rec.autos	961
comp.sys.ibm.pc.hardware	955	talk.politics.misc	980
rec.sport.baseball	932	soc.religion.christian	970

“20 Newsgroups” Document-base Description (NG.D9842.C10 & NG.D9614.C10)

Document-base Evaluation

Confidence: 35%	Support: 0.1%
Accuracy Sitting: Ten-fold Cross Validation	
Document-base	Accuracy
RE.D6643.C8	86.23%
NG.D9482.C10	77.49%
NG.D9614.C10	81.26%
OH.D6855.C10	79.27%

CONCLUSIONS

- In this study, we investigated the problem of textual data preparation, and introduced a standard document-base extraction approach for single-label TC.
- Based on three well-known textual data-sources (Retuers-21578, “20 Newsgroups”, and MedLine-OHSUMED), we extracted four document-bases with a simple text pre-processing approach.
- The experimental results demonstrate the effectiveness of our approach.
- Further single-label TC related studies are invited to utilize our proposed document-base extraction approach or directly make use of our generated document-bases in their result and evaluation part.