Available online at www.sciencedirect.com
ScienceDirect
Information Processing and Management 43 (2007) 344–352
INFORMATION PROCESSING & MANAGEMENT
www.elsevier.com/locate/infoproman

Contextual feature selection for text classification

Francois Paradis *, Jian-Yun Nie

DIRO, Université de Montréal, Montreal, Que., Canada

Received 28 May 2006; accepted 25 July 2006

Abstract

We present a simple approach for the classification of ''noisy'' documents using bigrams and named entities. The approach combines conventional feature selection with a contextual approach to filter out passages around selected features. Originally designed for *call for tender* documents, the method can be useful for other web collections that also contain non-topical contents. Experiments are conducted on our in-house collection as well as on the 4-Universities data set, Reuters 21578 and 20 Newsgroups. We find a significant improvement on our collection and the 4-Universities data set (10.9% and 4.1%, respectively). Although the best results are obtained by combining bigrams and named entities, the impact of the latter is not found to be significant.

Keywords: Classification; Named entities; Feature selection; Text filtering

1. Introduction

Text classification techniques rely heavily on the presence of a good *feature set*, or indexing terms, and the selection of discriminant features with regards to the classes. The quality of the feature set depends on the ''cleanliness'' of the documents: the presence of non-relevant or repetitive contents, as is often found on the web, degrades classification performance. In our work, we are especially interested in a particular kind of Web documents, *call for tenders*, in which a contracting authority invites contractors to submit a tender for their products and/or services. These documents can be found on the contracted organisation Web site, or on dedicated tendering sites. In earlier work (Paradis & Nie, 2005) we hypothesized that the noise in such documents was caused by the use of a *sublanguage* (Biber, 1993; Lehrberger, 1982) that describes the procedural aspects of the tenders submission, rather than their topic.

While feature selection undoubtedly brings a significant improvement to some classification methods (Yiming Yang, 1997), it is not clear whether it is adequate to filter such ''procedural'' noise. Indeed in our experiments with call for tenders we have found it difficult to extract either the procedural language (i.e. non-relevant features), or the tenders topic language (i.e. relevant features). There is a significant overlap

* Corresponding author. Tel.: +1 514 343 7484.
E-mail addresses: paradifr@iro.umontreal.ca (F. Paradis), nie@iro.umontreal.ca (J.-Y. Nie).

0306-4573/$ - see front matter © 2006 Published by Elsevier Ltd.
doi:10.1016/j.ipm.2006.07.006

between the two vocabularies, and the procedural language is often also relevant in the topics language. For example, the term "solicitation" can refer to the a call for tenders notice, or to the topic in the case of retail trade or telephone call centers. A difference is that the procedural language uses some common phrases or language constructs, which can be a direct indication of the relevance of the surrounding context.

In this paper we propose to combine conventional feature selection with a *contextual* approach to filter out words or passages in the documents. That is, we first select some *n*-grams features or named entities, and accept or reject passages based on their presence or absence. Our aim is to improve classification removing the "noise" from documents.

First, we briefly review related work and the context of our study. We then present our approach, first with bigram selection alone, then with the addition of named entities. We obtain a significant increase of the micro-F1 measure: +10.9% on our collection of call for tenders, and +4.1% on the 4-Universities data set. We also find that named entities alone can bring some improvement, but that results are only marginally better when combined with bigrams.

2. Related work

The search for a better feature set for classification is hardly new. It has been demonstrated that feature selection is central to some algorithms such as Naive Bayes (Yiming Yang, 1997), and therefore several techniques have been proposed, the most popular being *InfoGain*. In early work by Lewis (1992) the use of *phrases*, i.e. terms syntactically connected, was considered as a replacement for single-term features. The results were discouraging, which could be partly explained by the fact that there were too many index terms, with low frequency and high redundancy. Still, the idea was revisited by many. More recently Tan, Wang, and Lee (2002) find an improvement on the classification of Web pages, by using a combination of bigrams and unigrams selected on the merits of their InfoGain score. However the same technique applied to the Reuters collection did not yield the same gain, mostly because of its over-emphasis on "common concepts". Since their method favours recall, the authors conclude it was harder to improve Reuters because it already had high recall.

The traditional use of the term *filtering* in classification refers the selection of documents relevant to a user profile. There has been much interest lately with spam filtering (Zhang & Yao, 2003). Content filtering, such as discussed in this paper, is also not a new idea, although it has not often been linked with classification. Early work with filtering based on character *n*-grams met with surprising success (Cavnar, 1993). In Denoyer, Zaragoza, and Gallinari (2001) the notion of non-relevant passages in a document is exploited: a document is classified based on the relevance of its passages and their sequence as modeled with hidden Markov models. The area of automatic summarisation (Orasan, Pekar, & Hasler, 2004) is also related, since one of its subgoals is also to identify the most meaningful sentences. For example, the relevancy of a sentence can be defined based on its position, length, the frequency of the terms and its similarity with the title (Nobata et al., 2001).

Text classification is often used in the process of named entity extraction (Jansche, 2002) but rarely the other way around. Its use in classification is mostly restricted to replacing common strings such as dates or money amounts with tokens, to increase the ability of the classifier to generalise.

3. Classification of call for tenders

3.1. The MBOI project

This study is part of the MBOI project (matching business opportunities on the internet), which deals with the discovery of business opportunities on the internet (Paradis et al., 2004). The project aims to develop tools for business watch, including spidering, information extraction, classification, and search. The aspect of interest here, classification, consists of classifying call for tenders by industry type, according to one of the existing norms: SIC (standard industrial classification), NAICS (North American industry classification system), FCS (federal supply codes), CPV (common procurement vocabulary), etc.

A difficulty in the classification of call for tenders is to identify the relevant information amongst submission instructions, rules, requirements, etc. Sometimes the notice posted on the Web has little information to determine the subject, and the applicant is required to order a full description from the contracting authority.

Furthermore, we are spidering sites with great variations in style and format. Since a given organisation tend to reuse the same patterns, a first version of our prototype used some fixed information extraction rules: however this made it difficult to add new sites, or to know when rules for existing sites should be updated.

3.2. The test collection

For our experiments, we created a collection of call for tenders documents by downloading the XML daily synopsis from the FedBizOpps Web site (tenders solicited by American government agencies, available at http://www.fedbizopps.gov/). The XML documents have the same contents as the HTML documents found on the same site. The period downloaded ranged from September 2000 to October 2003. We kept only one document per tender, i.e. chose a document amongst pre-solicitations and amendments. Our collection, thereafter called *FBO*, is available at http://rali.iro.umontreal.ca/mboi/fbo/.

An example of call for tender is shown in Fig. 1. Not shown are some meta-data such as the date of publication (“21 May 2001”), classification codes (NAICS “424120” and FCS “75”), the contracting authority (“Office of Environmental Studies”), etc. The body of the document is composed of the subject line and the description; only these fields will be used for classification. Only a portion of the body is indicative of the tender subject. The rest concerns dates and modalities for submission.

We considered only documents with two classification codes, FCS and NAICS (although FCS will not be used here). Since the NAICS codes were not tagged in XML at the time (as they now are), they were extracted from the free text description. This resulted in 21945 documents (72 Megs), which were splitted 60% for training, and 40% for testing.

The NAICS codes are hierarchical: every digit of a six-digit code corresponds to a level of the hierarchy. For example, for industry code 424120 (stationery and office supplies merchant wholesalers) the sector code is 424 (merchant wholesalers, nondurable goods). Each of the three participating countries, the US, Canada and Mexico, have their own version of the standard, which mostly differ at the level of industry codes (5th or 6th digit). We reduced the category space by considering only the first three digits, i.e. the corresponding “sector”. This resulted in 92 categories (vs. 101 for FCS). We did not normalise for the uneven distribution of categories: for NAICS, 34% of documents are in the top two categories, and for FCS, 33% are in the top five. Our collection thus has similar characteristics to Reuters in terms of size, number and distribution of categories.

Our baseline for this collection is a Naive Bayes classifier trained and tested on the unfiltered documents. Naive Bayes is a common choice in the literature for baseline (Jason, Rennie, Lawrence Shih, & Karger, 2003), and it is known to be sensitive to feature selection, which makes it appropriate to our study. Furthermore, some of the better performing but costlier techniques, such as SVM, do not scale up to our project requirement of handling a large document base and feature set.

The 8000 top terms were selected according to their InfoGain score. The following thresholds were applied: a rank cut of 1 (*rcut*), a fixed weight cut of 0.001 (*wcut*), and a category cut learnt after cross-sampling 50% of the test set over 10 iterations (*scut*). More details about these thresholding techniques can be found in Yang (1999, 2001). The rainbow software (McCallum, 1996) was used to perform our experiments. The results for our baseline classifier are shown in Table 1, under the label “baseline”.

Subject: **Office supplies and devices**
Description: The office of Environmental Studies intends to procure printer toner cartridges and supplies for the Naval Inventory Control Point in Mechanicsburg, PA. Request for Quotation (RFQ) N00140-04-Q-4555 contemplates an indefinite delivery type firm fixed price order. This is a combined synopsis/solicitation for commercial items prepared in accordance with the format in FAR Subpart 13.5, Test Program for Certain Commercial Items, as supplemented with additional information included in this notice. This announcement constitutes the only solicitation; proposals are being requested, and a written solicitation will not be issued. This is a 100% Total Small Business Set-Aside. etc.

Fig. 1. A call for tender.

Table 1
FBO classification

Method	Macro-F1	Micro-F1
Baseline	.3297	.5498
Trained	.3223 (−2.2%)	.5918 (+7.6%)
Sentence bigram	.3585 (+8.7%)	.5891 (+7.1%)
Window bigram	.3583 (+8.7%)	.6075 (+10.5%)
Window entity	.3325 (≈)	.5640 (+2.6%)
Window bigram and entity	.3657 (+10.9%)	.6096 (+10.9%)

4. Passage filtering with bigrams

Two levels of passage filtering are considered: sentences or *windows* (i.e. sequence of words). Window filtering is appealing on our collection, because sentences can be long, and relevant and non-relevant information is often mixed in a sentence. Also, segmenting into sentences is not trivial in this collection, because it is not well formatted: for example the end-of-sentence period could be missing, or a space could appear inside an acronym (e.g. ''U._S.'').

4.1. Supervised filtering of sentences

In a first experiment we manually labeled 1000 sentences from 41 documents of FBO. The label was ''positive'' if the sentence was indicative of the tender's subject, or ''negative'' if not. Sentences with descriptive contents were labeled positive, while sentences about submission procedure, rules to follow, delivery dates, etc. were labeled negative. In the example of Fig. 1, only the first sentence would be labeled positive. Overall, almost a quarter of the sentences (243) were judged positive.

Intuitively, one would think that the first sentence(s) would often be positive, i.e. the author would start by introducing the subject of the tender, and then explain the rules and requirements. However this is not always the case. In the 41 documents we manually labeled, 25 documents had their first sentence positive, and 16 did not. In *combined* tenders, the text often starts with background information, and then define each item. In some cases, the subject is scattered amongst negative sentences.

We trained a Naive Bayes classifier on the 1000-sentence collection, for the positive and negative classes. The task seems to be relatively simple, since when we tested the classifier on a 40/60 split we obtained a micro-F1 measure of 85%. We thus filtered the whole collection with this classifier, keeping only the positive sentences. The collection size went from around 600,000 sentences to 96,811. The new, filtered documents were then classified with another Naive Bayes classifier.

Table 1 shows that this classification (''trained'') gives an increase of the micro-F1 measure, 7.6% over the baseline. Although this result in itself is interesting, our real aim is to achieve unsupervised filtering, i.e. not requiring a training collection and labeled sentences. We propose in the next section a technique to select sentences based on the presence of vocabulary.

4.2. Unsupervised filtering of sentences

Our approach to unsupervised filtering of sentences is to build a list of *n*-grams from the collection, and then filter out either a sentence or a window of terms around each of their occurrences in the documents. We define an *n*-gram as a consecutive sequence of *n* words, after removal of stop words. For example, we have found the following top five *n*-grams in FBO:

- unigrams: ''commercial'', ''items'', ''acquisition'', ''government'' and ''information'',
- bigrams: ''items-commercial'', ''business-small'', ''conditions-terms'', ''fedbizopps-link'' and ''document-fedbizzopps'',
- trigrams: ''link-fedbizopps-document'', ''supplemented-additional-information'', ''additional-information-included'', ''information-included-notice'', ''prepared-accordance-format''.

We first note that using *n*-grams as features has an adverse effect on our collection. For example, indexing only bigrams features results in a major drop of −20% macro-F1 and −12% micro-F1. Its seems the larger vocabulary has an ill effect on classification. Selecting less than 64,000 bigrams was not enough to capture the information, however selecting more introduced too much noise.

Selection of *n*-grams can be performed using the InfoGain measure, where the lowest scores, i.e. least discriminant *n*-grams, are the best candidates for filtering. However we have found that in this particular collection, choosing the high-frequency terms works just as well. This is because these are relatively uniformly distributed in the classes. On the other hand the InfoGain seems to capture some unfrequent features, whether they are distributed evenly or not.

Table 1 shows results of sentence filtering with bigrams (''sentence bigram''). Only 1250 bigrams were selected (this parameter was determined manually). The criterion for a sentence to be filtered out was the following: a sentence was rejected if 1/8 of its bigrams were in the reject list (again this parameter was determined empirically). The result, 0.5891, or an increase of 7.1% over the baseline, is similar to the trained classifier of the preceding section.

We have also tried unigrams and trigrams, but found bigrams to give the best results. Trigrams come close with a 6.6% increase of micro-F1, but the macro-F1 increase only by 2.9%.

4.3. Window filtering

As mentioned before, although the sentence seems like a good logical unit to perform filtering, it is a bit problematic in our collection because it is not so well delimited, and can contain both relevant and non-relevant information. Another approach is to ignore punctuation and sentence markers, and to filter a window around a term.

We select bigrams as above, and filter out a region of *m* words preceding, up to *m* words after the bigram. Additionally, two regions to be filtered out are connected if ''close'' enough.

Table 1 shows the results of window filtering for bigrams (''window bigram''). A window of size two was used, i.e. the region filtered out started with the two terms preceding the *n*-gram, up to the two succeeding terms. Two regions to be filtered out were ''connected'' if less than six terms apart. This filter gives our best results so far: a micro-F1 of 0.6075 (+10.5%) and a macro-F1 of .3583 (+8.7%).

We have also tried other *n*-grams: again a combination of bigrams and trigrams come close (0.3552 macro and 0.5997 micro-F1) but not as good as bigrams alone.

We have presented results of filtering *out* sentences or windows, based on non-relevant features. We have also tried the opposite, i.e. selecting relevant features and keeping only those sentences or windows where they appeared. As expected, the required feature set is much larger, but the results are similar.

5. Named entities

5.1. Entities as indicators of relevance

Named entities are expressions containing names of people, organisations, locations, time, etc. These often appear in call for tenders, but are rarely indicative of the subject of the tender. Therefore, we hope that by identifying these expressions, we can either filter out passages that contain them, or reduce their impact on the classifier.

We take a somewhat broad definition of named entities, to include the following:

- *Geographical location*. In a call for tender, this can be an execution or delivery location. A location can also be part of an address for a point of contact or the contracting authority (although these are often tagged as meta-data in FBO, they often appear in the text body).
- *Organisation*. Most often the organisation will be the contracting authority or one its affiliates. For predetermined contracts it can be the contractor.
- *Date*. This can be a delivery date or execution date (opening and closing dates are often explicitly tagged as meta-data, and therefore do not need to be extracted).

- *Time*. A time limit on the delivery date, or business hours for a point of contact.
- *Money*. The minimum/maximum contract value, or the business size of the contractor.
- *URL*. The web site of the contracting authority or a regulatory site (e.g. a link to CCR – central contract registry).
- *Email*, *phone number*. The details of a point of contact.

Although these entities have a particular use in our collection, they are generic in the sense that they also apply to many other domains. We have also considered the following entities, specific to our collection:

- *FAR* (federal acquisition rules). These are tendering rules for US government agencies. A call for tender may refer to an applicable paragraph in the FAR (e.g. ''FAR Subpart 13.5'').
- *CLIN* (Contract Line Item Number). The line item define a part or sub-contract of the tender. Line items usually appear as a list (e.g. ''CLIN 0001: ...'').
- *Dimensions*. In the context of a tender, a dimension almost always refers to the physical characteristics of a product to deliver (e.g. ''240 mm × 120 mm'').

All entities except CLIN and dimensions are *negative* indicators: their presence is an indication of a negative passage or sentence, i.e. not relevant to the subject of the tender. CLIN and dimensions on the other hand are *positive* indicators, since they introduce details about the contract or product.

The entities were identified in the collection using a mix of regular expressions and *Nstein NFinder*, a tool for the extraction of named entities. Table 2 shows the accuracy of the entities as positive/negative indicators on the 1000 training sentences. For example, dates (a negative indicator) appeared in 62 sentences, 59 of which were labeled negative. Dimensions (a positive indicator) appeared in eight sentences, all of which were labeled positive.

Locations, persons and organisations are the most ambiguous entities, with an accuracy around or lower that of an ''always-negative'' classifier (which would be correct 75.7% of the time on our 1000 sentences). That is partly because they often appear along with the subject in an introductory sentence. For example in Fig. 1 the first sentence contains an organisation, ''Office of Environmental Studies'', a location, ''Mechanicsburg, PA'', as well as the subject, ''toner cartridges and supplies''. Furthermore, these entities are inherently more difficult to recognise than date and time, which only require a few simple patterns, and can achieve near-perfect recognition accuracy. To make matters more difficult, some documents are all in capital letters, which make the task more difficult because there are no clues to distinguish proper and common nouns. Some examples of recognition errors were: ''Space Flight'' identified as a person, ''FOB'' as an organisation, or ''184 BW Contracting Office'' as a location.

5.2. Classification with entities

As noted above, a common use of named entities in text categorisation is to replace each instance in the text with a generic token. Bigrams are computed using these tokens, with the hope to find more generic patterns. For example, the bigrams now include patterns such as 'exceed-[money]' (as in ''business size should not

Table 2
Named entities in 1000 sentences

Type	Accuracy	Type	Accuracy
−Location	72% (252/348)	−Person	82% (351/429)
−Organisation	75% (357/479)	−Date	95% (59/62)
−Time	98% (42/43)	−money	100% (18/18)
−URL & email	100% (38/38)	−Phone number	98% (39/40)
−FAR	100% (56/56)		
+CLIN	80% (4/5)	+Dimensions	100% (8/8)

exceed $10.4 M''). Such patterns could not be picked up before because money amounts, as other numbers, would be rejected by the tokeniser. Furthermore, using an entity tag should increase the frequency of the bigram and therefore its chance to be included in the filter list.

That strategy does not pay off on FBO, since it brings a decrease of 1.5% to the macro-F1 and no change to the micro-F1. This can be explained partly by the fact that the best predictors were all negatives (except for dimensions) and included numeric attributes which were already ignored by the classifier. We have tried different combinations of entities, especially leaving out locations and organisations, all with similar results.

Another use of named entities is for acronym expansion. Organisation names sometimes provide valuable clues to the tender's subject. For example, knowing that the contracting authority is the USDA (US Department of Agriculture) increases the likelihood of a tender to be relevant to agriculture. This information is already taken into account by the classifier if the full name appears in the text. However if the acronym alone appears, only limited inference is possible (unless the acronym systematically appeared in all tenders of its kind).

We have tried to expand acronyms based on information collected from the training collection. Firstly we have built an acronym list from all organisation entities of the form: ''full name (acronym)''. We thus collected 1068 acronyms, excluding two-letters acronyms, which were deemed too ambiguous, especially since our collection includes many two-letter state abbreviations. We then expanded acronyms in the documents, except when they appeared inside brackets, and used the window bigram selection. Unfortunately, this approach yielded a micro-F1 of .5265, a decrease of 4.2% over the baseline. One possible explanation for this poor performance is the high degree of ambiguity in the acronyms. For example, in our collection ISS refers to ''integrated security system'' or ''international space station''. In this case we put both expansions in the document.

Finally, named entities were used as a basis for window selection. The accuracy information from Table 2 was exploited. We built a filter that rejects a passage when ''enough'' negative indicators are found, based on its accuracy in the 1000-sentences.

When named entities were used as the sole criterion, there was a modest increase of 2.6% micro-F1 over the baseline (Table 1, ''window entity''). When combined with bigrams, i.e. using both bigrams and named entities for window selection (''window bigram and entity''), the results are similar to the ''window bigram'' method. The macro-F1 measure shows a 10.9% increase over the baseline, and 2.1% over the window bigram. However upon analysis this increase is not so encouraging because the classifier worked well mostly on marginal classes with few documents.

Another possible use of named entities is to exploit the accuracy information from Table 2. We have built a sentence filter that rejects a sentence if enough negative indicators are found. For indicators with a 100% accuracy, one instance is enough to reject a sentence. For others, we give a weight to each entity equals to its accuracy minus 75.7% (i.e. the accuracy of the always-negative classifier). We sum up the weights, and reject the sentence if it is above a threshold (which we have set to .40 in this experiment). The results obtained were similar to ''window entity'' in Table 1.

6. Results on standard collections

We tried our approach on the following standard collections:

- Reuters 21578.[1] We used the aptemod split, with 8000 features selected by InfoGain.
- Twenty Newsgroups.[2], a collection of approximately 20,000 newsgroup documents partitioned across 20 different newsgroups. This time we obtained the best results by selecting 20,000 index terms by InfoGain.
- The 4-Universities collection.[3] containing 8282 university web pages, manually classified in seven categories. We followed CMU suggestion and selected 2000 index terms by InfoGain. We also used their script to replace some numbers with generic tokens. We have not however implemented the suggested cross-tests,

[1] http://www.daviddlewis.com/resources/testcollections/reuters21578/.
[2] http://people.csail.mit.edu/jrennie/20Newsgroups/.
[3] http://www-2.cs.cmu.edu/afs/cs.cmu.edu/project/theo-20/www/data/.

Table 3
Classification results on standard collections

Collection	Method	Macro-F1	Micro-F1
Reuters 21578	Baseline	.3861	.7910
	Window bigram	.3903 (+1.1%)	.7947 (≈)
20 Newsgroup	Baseline	.8170	.8235
	Window bigram	.8213 (≈)	.8270 (≈)
4-Universities	Baseline	.6141	.6685
	Window bigram	.6297 (+2.5%)	.6957 (+4.1%)

i.e. train on three, and test on one university. Instead, we created a fixed split of 70% training, and 30% testing documents.

For all collections, the same classifier and thresholding techniques were used as above, except that we used 20 iterations for scut to compensate for the smaller testing sets. On the 4-Universities data set, we found it was easier to filter in the passages (rather than filtering out as with the other collections). We selected the 15,000 bigrams with highest InfoGain score.

Table 3 shows the results. Not surprisingly, the technique does not have much impact on Reuters and 20 Newsgroups, except for the Reuters macro-F1 measure, which gets a moderate increase of 1.1%. A closer look shows that relatively few documents actually had some contents filtered out (about 5% for Reuters and 8% for 20 Newsgroups). Furthermore, in the Reuters collection, most of the modified documents were of the class "earn". This is the most frequent class (it accounts for more than 36% of the collection) and accordingly it is "overclassified" by the Naive Bayes classifier. Filtering contents in this class has enabled some false positives to be classified under the correct categories, thus increasing the macro-F1.

We tried to increase the coverage of classes in Reuters by forcing a proportional distribution of selected bigrams over the classes. However this did not yield better results.

The results are more convincing on the 4-Universities data set, with an increase of the micro measure of 4.1%. The reason for this result, however, seems to be different than for FBO. Although it was difficult to pick good candidates for filtering out, there were some obvious candidates for a bigram-based feature selection. For example the top bigrams according to their InfoGain score included: "computer-science" (good discriminant for the department class), "research-interests" (faculty), "phone-digits" (faculty and staff), etc. Indeed, when indexing on bigram features, we found an increase of the micro-F1 measure to 0.7055, while macro-F1 remained stable at 0.6205. Another difference with FBO is that when we tried extracting named entities and indexing them as tokens, we obtained an increase of around 2.6% over the baseline.

Finally we note that our results are consistent with the entropy information found in the collections. The FBO collection had the lowest sentence-based perplexity score (59.9), which gives an indication about its recurrent patterns, while the 4-Universities data set had the highest InfoGain bigrams scores.

7. Conclusion

We investigated the use of bigrams and named entities to perform content filtering. Our domain of application was the classification of call for tenders. Our findings are that filtering a window of terms around selected bigrams works well for this kind of collection: we could obtain an increase of 10.9% of micro-F1 on our collection, and 4.1% on the 4-Universities data set. More tests are needed in the future to verify our claim that this technique works well on web pages.

Using named entities in various ways did not bring a significant improvement to this result. This could be due to the fact that most entities had low probability in the collections, and also in part the poor accuracy of the named entity extractor.

We are currently investigating the use of these techniques not for content filtering, but to set different indexing weights based on the selection of passages. Another idea worth pursuing is taking advantage of the sequence of relevant and non-relevant sentences in the document. This idea is similar to the HMM proposed in Denoyer et al. (2001).

Acknowledgements

This project was financed jointly by NSERC and Nstein Technologies.

References

Biber, D. (1993). Using register-diversified corpora for general language studies. *Computational linguistics, 19*(2).

Cavnar, W. (1993). *N*-gram-based text filtering for trec-2. In *Second text retrieval conference (TREC)*.

Denoyer, L., Zaragoza, H., & Gallinari, P. (2001). Hmm-based passage models for document classification and ranking.

Jason, D. M., Rennie, Lawrence Shih, J. T., & Karger, D. R. (2003). Tackling the poor assumptions of naive bayes text classifiers. In *Proceedings of the twentieth international conference on machine learning*.

Jansche, M. (2002). Named entity extraction with conditional markov models and classifiers. In *The 6th conference on natural language learning*.

Lehrberger, J. (1982). Automatic translation and the concept of sublanguage. In R. Kittredge, J. Lehrberger (Eds.), *Sublanguage: Studies of language in restricted semantic domains*.

Lewis, D. (1992). An evaluation of phrasal and clustered representations on a text categorization task. In *15th ACM international conference on research and development in information retrieval (SIGIR)* (pp. 37–50).

McCallum, A. K. (1996). Bow: A toolkit for statistical language modeling, text retrieval, classification and clustering. <http://www.cs.cmu.edu/~mccallum/bow>.

Nobata, C., Sekine, S., Murata, M., Uchimoto, K., Utiyama, M., & Isahara, H. (2001). Sentence extraction system assembling multiple evidence.

Orasan, C., Pekar, V., & Hasler, L. (2004). A comparison of summarisation methods based on term specificity estimation. In *Proceedings of the fourth international conference on language resources and evaluation (LREC-04)* (pp. 1037–1041).

Paradis, F., & Nie, J.-Y. (2005). Étude sur l'impact du sous-langage dans la classification automatique d'appels d'offres. In CORIA, Grenoble, France.

Paradis, F., Ma, Q., Nie, J.-Y., Vaucher, S., Garneau, J.-F., Gérin-Lajoie, R., & Tajarobi, A. (2004). Mboi: Un outil pour la veille d'opportunités sur l'internet. In *Colloque sur la Veille Strategique Scientifique et Technologique, Toulouse, France*.

Tan, C.-M., Wang, Y.-F., & Lee, C.-D. (2002). The use of bigrams to enhance text categorizatio. *Information Processing and Management: an International Journal, 38*(4), 529–546.

Yang, Y. (1999). An evaluation of statistical approaches to text categorization. *Journal of Information Retrieval, 1*(1/2), 67–88, an excellent reference paper for comparisons of classification algorithms on the Reuters collection.

Yang, Y. (2001). A study on thresholding strategies for text categorization. In *Proceedings of SIGIR-01, 24th ACM international conference on research and development in information retrieval*.

YimingYang, J. O. P. (1997). A comparative study on feature selection in text categorization. In *Proceedings of ICML-97, 14th international conference on machine learning*.

Zhang, L., & Yao, T. (2003). Filtering junk mail with a maximum entropy model. In *Proceeding of 20th international conference on computer processing of oriental languages (ICCPOL03)* (pp. 446–453).