# A Method Based on the Chi-Square Test for Document Classification

Michael Oakes
Dept. of Computer Science
University of Sunderland
michael.oakes@sund.ac.uk

Robert Gaizauskas
Dept. of Computer Science
University of Sheffield
robertg@dcs.shef.ac.uk

Helene Fowkes
Dept. of Information Studies
University of Sheffield
H.F.Fowkes@shef.ac.uk

## ABSTRACT

We introduce a method for document classification based on using the chi-square test to identify characteristic vocabulary of document classes.

## 1. INTRODUCTION

*Scrip* [1] is a daily news bulletin, available electronically, and circulated widely amongst those working in the pharmaceutical industry. Our aim was to split the incoming stream of *Scrip* texts into topic-based substreams, which would become inputs to domain specific information extraction (IE) engines. The topics of interest were defined as product launches, licensing, meetings, people (personnel data), announcements by regulatory authorities, company relations and clinical trials. Together the document routing and IE work are part of the TRESTLE project. [2]

## 2. THE CHI-SQUARE APPROACH

The chi-square test has been used to identify vocabulary characteristic of male versus female speech by Rayson et al. [2]. This work suggested to us that an application to document classification ought to be possible. For, if the characteristic vocabulary of a class $C$ of texts can be successfully distinguished from that of a class $D$, then to classify a new text as $C$ or $D$ it should be sufficient to determine whether its vocabulary is more similar to that of $C$ or $D$.

### 2.1 Identifying Characteristic Vocabulary

In our setting we can identify a *specific corpus* $S$, a set of texts from *Scrip* in a given period about a specific topic, such as clinical trials, a *residue corpus* $R$ which is the set of *Scrip* texts from the same period which are *not* about the given topic and a *general corpus* $G$, which is all of the *Scrip* texts from the given period; i.e., $G = S \cup R$. The chi-square procedure for identifying characteristic vocabulary is as follows. For each unique word (type) $W$ in each corpus, count the following, where $count(C, W)$ is the number of times word type $W$ occurs in corpus $C$: $a$ = number of times $W$ occurs in the specific corpus, i.e. $a = count(S, W)$; $b$ = number of times $W$ occurs in the residue corpus, i.e. $b = count(R, W)$; $c$ = total number of words (tokens) in the specific corpus which are not $W$, i.e. $c = |S| - a$; $d$ = total number of words (tokens) in the residue corpus which are not $W$ , i.e. $d = |R| - b$. The values $a$ - $d$ are called the observed frequencies (O), and may be arranged in a $2 \times 2$ contingency table, i.e.

| | Specific Corpus | Residue Corpus |
|---|---|---|
| Word | $a$ | $b$ |
| ¬ Word | $c$ | $d$ |

We next calculate the expected frequencies (E) for each table cell (counts which would have been obtained given the size of the corpora and the rarity of the word in question, if the word were equally typical of each corpus) using the formula:

$$E_{i,j} = \frac{\text{column}_i \text{ total} \times \text{row}_j \text{ total}}{\text{grand total}}$$

For example, for the observed frequency $a$, the total for the row in the contingency table containing $a$ is $a+b$, the column total for $a$ is $a+c$, and the grand total is the total of all four counts, $a + b + c + d$. Given these observed and expected frequencies, the chi-square value is calculated as the sum of the quantities $(O-E) \times (O-E)/E$ for each position in the contingency table. I.e.

$$\chi^2 = \Sigma_{i,j} \frac{(O_{i,j} - E_{i,j})^2}{E_{i,j}}$$

If chi-square is greater than 3.84, we can be 95% confident that the word does occur more frequently in one of the two text types. If the ratio $a/b$ is greater than the ratio $(a+c)/(a+d)$, then the word is more typical of the specific corpus (a "positive indicator"), otherwise it is more typical of the residue corpus (a "negative indicator"). For the calculation to be reliable we must discard any words where E is less than 5). As an example, the most significant positive indicators found for the specific class of clinical trials texts were *patients, Phase, study, treatment, trial, trials, III, with, cancer, II, results, studies, product, therapy, approval, failure, placebo, days, in, panel.* The most significant negative indicators were *pharmaceutical, PEOPLE, companies, products, makes, 1999, industry, president, on, appointment.*

[1] *Scrip* is the trademark of PJB Publications Ltd. See `http://www.pjbpub.co.uk`.
[2] See http://www.dcs.shef.ac.uk/nlp/projects/trestle.


*SIGIR'01,* September 9-12, New Orleans, Louisiana, USA..

## 2.2 Classifying Documents Using Characteristic Vocabulary

Given a collection of *Scrip* texts which have been classified as belonging or not belonging to a specified category, the chi-square analysis process described above produces as output a ranked list of "keywords" deemed to be the most and least characteristic of the texts in the category. We have used these lists to classify unseen *Scrip* texts into various categories. The classification algorithm takes two inputs: A list $L$ of the keywords with significant chi-square scores (greater than 3.84, significant at the 5% level), and a set $D$ of documents to be classified. Each of the unseen documents is read in turn, and awarded a score of 1 for every positive indicator word (token) it contains and -1 for every negative indicator it contains, giving the highest total scores to those documents containing typical vocabulary for the specific domain, ideally those actually about that domain. The output of this process is a ranked list of document scores for all the documents in $D$. If one assumes that one's test set contains texts divided into categories in the same proportion as one's training data, then one can select the equivalent proportion of documents with the highest document scores from the test set.

# 3. THE SVM APPROACH

To provide a basis for comparison with the chi-square approach described above, we also implemented a support vector machine (SVM) approach [1] to document classification. Our training documents were converted into feature vectors (inputs to the SVM) as follows. First the entire vocabulary of the document set (both training and test) was found, for example [apple banana carrot dill eggplant], then for a document containing the words "apple carrot carrot dill", a word-occurrence-count feature vector would be created: [1 0 2 1 0]. In our experiments, all words were used (even those with a frequency of 1) but no stemming. The SVMs were trained with a linear decision boundary (LDB).

# 4. TRAINING AND TEST DATA

Three months' supply of *Scrip*, 2074 documents varying in length from around 30 to 2000 words, were manually classified by two independent judges. A *Scrip* document could be assigned to any number of categories (or none). A third person acted as an adjudicator if the two judges disagreed whether a document should be placed in a given category. The first 525 documents and their judgements became the training set for both the chi-square method and the SVM method. The last 1579 documents and their judgements became the test set. The degree of consistency between the two judges (reflecting the relative ease with which documents could be classified into each domain) was measured using the Kappa statistic [3]. In five out of the seven domains, meetings, people, regulatory, relations and trials, inter-annotator agreement was significant at the 1% level.

# 5. EVALUATION

The two primary metrics used in our evaluation are the standard metrics of *recall* and *precision*. In this context, recall was defined as the number of matches (where the machine placed a document in a category where the human judges also placed it) divided by the total number of documents placed by the human judges in that category. Precision was defined as the number of matches divided by the total number of documents placed by the machine in that category. Table 1 shows results for the chi-square method. Note that the worst results are those obtained for the launches and licensing categories – the two for which the kappa statistic scores showed the worst results for human agreement. Other results are substantially better. Table 2 shows the results of the SVM method.

| | matches | machine | human | prec | recall |
|---|---|---|---|---|---|
| launches | 10 | 58 | 59 | .172 | .169 |
| licensing | 12 | 72 | 70 | .167 | .171 |
| meetings | 87 | 138 | 103 | .630 | .845 |
| people | 205 | 260 | 222 | .788 | .923 |
| regulatory | 105 | 259 | 154 | .405 | .682 |
| relations | 142 | 406 | 181 | .350 | .785 |
| trials | 288 | 371 | 430 | .776 | .670 |

**Table 1: Results for the Chi-Square Method**

| | matches | machine | human | prec | recall |
|---|---|---|---|---|---|
| launches | 22 | 35 | 59 | .629 | .373 |
| licensing | 15 | 19 | 70 | .789 | .214 |
| meetings | 101 | 102 | 103 | .990 | .981 |
| people | 206 | 209 | 222 | .986 | .928 |
| regulatory | 83 | 138 | 154 | .601 | .539 |
| relations | 89 | 145 | 181 | .614 | .492 |
| trials | 316 | 366 | 430 | .863 | .730 |

**Table 2: Results for SVM Method using word occurrence feature vectors, separated by LDB**

# 6. CONCLUSIONS

It is clear that the SVM classifier performs considerably better than the chi-square classifier on the same data. Despite this, we believe the chi-square approach shows sufficient promise to warrant further experimentation.

# 7. ACKNOWLEDGMENTS

The authors would like the acknowledge the financial support of GlaxoSmithKline which has made this work possible, and the helpful comments of Peter McMeekin, Charlie Hodgman, David Pearson and Derek Black.

# 8. ADDITIONAL AUTHORS

Anna Jonsson (Department of Information Studies, Sheffield University, email: A.Jonsson@shef.ac.uk), Vincent Wan (Department of Computer Science, Sheffield University, email: V.Wan@dcs.shef.ac.uk ), Micheline Beaulieu (Dept. Information Studies, Sheffield University, M.Beaulieu@shef.ac.uk).

# 9. REFERENCES

[1] S. Dumais, J. Platt, D. Heckerman, and M. Sahami. Inductive learning algorithms and representations for text categorisation. In *Proceedings of the 7th Int. Conf. on Information and Knowledge Management*, 1998.

[2] P. Rayson, G. Leech, and M. Hodges. Social Differentiation in the Use of English Vocabulary. *Int. J. of Corpus Linguistics 2(1): 133-152*, 1997.

[3] S. Siegel and N. J. Castellan. *Nonparametric Statistics for the Behavioral Sciences.* McGraw-Hill, 1988.