情　報　学　基　礎 33－9
（1994．5．18）

自動文書分類のための新しい確率モデル

岩山真
(株) 日立製作所基礎研究所
iwayama@harl.hitachi.co.jp

徳永健伸
東京工業大学工学部情報工学科
take@cs.titech.ac.jp

本論文では，自動文書分類のための新しい確率モデルについて報告する．本論文で提案する手法は，従来の確率モデルに比べ以下の利点を持つ．1) term の文書内頻度に基づく，2) 分類対象となる文書の term も重み付けされる，3) 数の少ない不十分な訓練データに大きく左右されない．本確率モデルの優位性は，“Wall Street Journal” の自動文書分類実験を通して確認された．

A Probabilistic Model for Text Categorization: Based on a Single Random Variable with Multiple Values

Iwayama, Makoto
Advanced Research Laboratory
Hitachi, Ltd.
iwayama@harl.hitachi.cp.jp

Tokunaga, Takenobu
Department of Computer Science
Tokyo Institute of Technology
take@cs.titech.ac.jp

Text categorization is the classification of documents with respect to a set of predefined categories. In this paper, we propose a new probabilistic model for text categorization, that is based on a *Single random Variable with Multiple Values* (SVMV). Compared to previous probabilistic models, our model has the following advantages; 1) it considers within-document term frequencies, 2) considers term weighting for target documents, and 3) is not affected by having insufficient training cases. We verify our model's superiority over the others in the task of categorizing news articles from the “Wall Street Journal”.

1 Introduction

Text categorization is the classification of documents with respect to a set of predefined categories. As an example, let us take a look at the following article from the "Wall Street Journal" (1989/11/2).

> McDermott International, Inc. said its Babcock & Wilcox unit completed the sale of its Bailey Controls Operations to Finmeccanica S.p.A for $295 million. Finmeccanica is an Italian state-owned holding company with interests in the mechanical engineering industry. Bailey Controls, based in Wickliffe, Ohio, makes computerized industrial controls systems. It employs 2,700 people and has annual revenue of about $370 million.

Two categories (topics) are manually assigned to this article; "TENDER OFFERS, MERGERS, ACQUISITIONS (TNM)" and "COMPUTERS AND INFORMATION TECHNOLOGY (CPR)." While there may be certain rules or standards for categorization, it is very difficult for human experts to assign categories consistently and efficiently to large numbers of daily incoming documents. The purpose of this paper is to propose a new probabilistic model for automatic text categorization.

While many text categorization models have been proposed so far, in this paper, we concentrate on the probabilistic models [9, 5, 4, 6, 1, 13, 14] because these models have solid formal grounding in probability theory. Section 2 quickly reviews the probabilistic models and lists their individual problems. In section 3, we propose a new probabilistic model based on a *Single random Variable with Multiple Values* (SVMV). Our model is very simple, but still solves all the problems of the previous models. In section 4, we verify our model's superiority over others through experiments in which we categorize "Wall Street Journal" articles.

2 A Brief Survey of Probabilistic Text Categorization

In this section, we will briefly review three major probabilistic models for text categorization. Originally, these models are exploited for text retrieval, but the adaptation to text categorization is straightforward.

In a model of probabilistic text categorization, the probability

$$P(c|d) \quad = \quad \text{the probability that a document } d \text{ is categorized into a category } c \tag{1}$$

is calculated. Usually, a set of categories is defined beforehand. For every document d_i, probability $P(c|d_i)$ is calculated and all the documents are ranked in decreasing order according to their probabilities. The larger $P(c|d_i)$ a document d_i has, the more probably it will be categorized into category c. This is called the *Probabilistic Ranking Principle* (PRP) [8]. Several strategies can be used to assign categories to a document based on PRP [6].

There are several ways to calculate $P(c|d)$. Three representatives are Robertson and Sparck Jones' method [9], Kwok's method [5], and Fuhr's method [4].

2.1 Probabilistic Relevance Weighting (PRW)

Robertson and Sparck Jones [9] make use of the well-known logistic (or log-odds) transformation of the probability $P(c|d)$.

$$g(c|d) = \log \frac{P(c|d)}{P(\overline{c}|d)} \tag{2}$$

where $\overline{c}$ means "not c", that is "a document is not categorized into c." Since this is a monotonic transformation of $P(c|d)$, PRP is still satisfied after transformation.

Using Bayes' theorem, Eq. (2) becomes

$$g(c|d) = \log \frac{P(d|c)}{P(d|\overline{c})} + \log \frac{P(c)}{P(\overline{c})}. \tag{3}$$

Here, $P(c)$ is the prior probability that a document is categorized into c. This is estimated from given training data, i.e., the number of documents assigned to the category c. $P(d|c)$ is calculated as follows. If we assume that a document consists of a set of *terms* (usually nouns are used for the first approximation) and each term appears independently in a document, $P(d|c)$ is decomposed to

$$P(d|c) = \prod_{t_i \in d} P(T_i = 1|c) \prod_{t_j \in c-d} P(T_j = 0|c) \tag{4}$$

where "$c-d$" is a set of terms that do not appear in d but appear in the training cases assigned to c. "t_i" represents the name of a term and "$T_i = 1, 0$" represents whether or not the corresponding term "t_i" appears in a document. Therefore, $P(T_i = 1, 0|c)$ is the probability that a document does or does not contain the term t_i, given that the document is categorized into c. This probability is estimated from the training data; the number of documents that are categorized into c and have the term t_i. Substituting Eq. (4) into Eq. (3) yields

$$g(c|d) = \sum_{t_i \in d} \log \frac{P(T_i = 1|c)}{P(T_i = 1|\overline{c})} \sum_{t_j \in c-d} \log \frac{P(T_j = 0|c)}{P(T_j = 0|\overline{c})} + \log \frac{P(c)}{P(\overline{c})}. \tag{5}$$

We will refer to Robertson and Sparck Jones' formulation as *Probabilistic Relevance Weighting* (PRW).

While PRW is the first attempt to formalize well-known relevance weighting [12, 10] by probability theory, there are several drawbacks in PRW.

[Problem 1: no within-document term freq.]
PRW does not make use of within-document term

frequencies. $P(T = 1, 0|c)$ in Eq. (5) takes into account only the existence/absence of the term t in a document. In general, frequently appearing terms in a document play an important role in text retrieval [10]. Salton experimentally verified the importance of within-document term frequencies in his vector model [11].

[Problem 2: no term weighting for target doc.] PRW uses term weighting only for categories (i.e., $P(T = 1, 0|c)$). Term weighting for target documents (i.e., $P(T = 1, 0|d)$) would be necessary for sophisticated text retrieval/categorization [4, 6].

[Problem 3: affected by insuff. training cases] In practical situations, the estimation of $P(T = 1, 0|c)$ is not always straightforward. Let us consider the following case. In the training data, we are given R documents that are assigned to c. Among them, r documents have the term t. In this example, the straightforward estimate of $P(T = 1|c)$ is "r/R." If "$r = 0$" (i.e., none of the documents in c has t) and the target document d contains the term t, $g(c|d)$ becomes $-\infty$, which means that d is never categorized into c. Robertson and Sparck Jones mentioned that, in PRW, there will be other special cases like the above example [9]. A well-known remedy for this problem is to use "$(r + 0.5)/(R + 1)$" as the estimate of $P(T = 1|c)$ [9].

2.2 Component Theory (CT)

To solve problems 1 and 2 of PRW, Kwok stresses the assumption that a document consists of terms [5]. This theory is called the *Component Theory* (CT).

To introduce within-document term frequencies (i.e., to solve problem 1), CT assumes that a document is completely decomposed into its constituting terms. Therefore, rather than counting the number of documents, as in PRW, CT counts the number of terms in a document for probability estimation. This leads to within-document term frequencies. Moreover, to incorporate term weighting for target documents (i.e., to solve problem 2), CT defines $g(c|d)$ as the geometric mean probabilities over components of the target document d;

$$\frac{P(d|c)}{P(d|\overline{c})} = [\prod_{t \in d} \frac{P(T|c)}{P(T|\overline{c})}]^{\frac{1}{|d|}}. \tag{6}$$

Following Kwok's derivation, $g(c|d)$ becomes

$$\begin{aligned} g(c|d) &= \sum_t P(T = t|d)(\log \frac{P(T = t|c)}{P(T \neq t|c)} \\ &\quad + \log \frac{P(T \neq t|\overline{c})}{P(T = t|\overline{c})}) + \log \frac{P(c)}{P(\overline{c})}. \end{aligned} \tag{7}$$

For precise derivation, please refer to Kwok's paper [5].

Here, note that $P(T = t|d)$ and $P(T = t|c)$ represent the within-document term frequencies for the target document d and the category c respectively. Therefore, CT is not subject to problems 1 and 2. However, problem 3 still affects CT. Furthermore, Fuhr pointed out that transformation, as in Eq. (6), is not monotonic of $P(c|d)$. It follows then, that CT does not satisfy the probabilistic ranking principle (PRP) any more.

2.3 Retrieval with Probabilistic Indexing (RPI)

Fuhr solves problem 2 by assuming that a document is *probabilistically* indexed to its term vectors [4]. This model is called *Retrieval with Probabilistic Indexing* (RPI).

In RPI, a document d has a binary vector $\boldsymbol{x} = (T_1, \ldots, T_n)$ where each component corresponds to a term. $T_i = 1$ means that the document d contains the term t_i. X is defined as the set of all possible indexings, where $|X| = 2^n$. Conditioning $P(c|d)$ for each possible indexing gives

$$P(c|d) = \sum_{\boldsymbol{x} \in X} P(c|d, \boldsymbol{x}) P(\boldsymbol{x}|d). \tag{8}$$

Applying Bayes' theorem leads to

$$P(c|d) = P(c) \sum_{\boldsymbol{x} \in X} \frac{P(\boldsymbol{x}|c) P(\boldsymbol{x}|d)}{P(\boldsymbol{x})}. \tag{9}$$

By assuming that each term appears independently in a target document d and in a document assigned to c, Eq. (9) is rewritten as

$$\begin{aligned} P(c|d) &= P(c) \prod_i (\frac{P(T_i = 1|c) P(T_i = 1|d)}{P(T_i = 1)} \\ &\quad + \frac{P(T_i = 0|c) P(T_i = 0|d)}{P(T_i = 0)}). \end{aligned} \tag{10}$$

Here, all the probabilities are estimated from the training data using the same method described in Section 2.1.

Since Eq. (10) includes the factor $P(T = 1, 0|d)$ as well as $P(T = 1, 0|c)$, RPI takes into account term weighting for target documents, essentially, solving problem 2. However, problems 1 and 3 still exist in RPI. In particular, because of problem 3, $P(c|d)$ would become an illegitimate value. In our experiments, as well as in Lewis' experiments [6], $P(c|d)$ ranges from 0 to more than 10^{10}.

3 A Probabilistic Model Based on a Single Random Variable with Multiple Values (SVMV)

In this section, we propose a new probabilistic model for text categorization, and compare it to the previous three models from several viewpoints. Our model is very simple, but still solves all the problems of the previous models.

We modify Fuhr's derivation of $P(c|d)$ by replacing the index vector $\boldsymbol{x}$ with a single random variable T whose value is one of several possible terms. For example, $T = t_i$ means that a randomly extracted term from a document is t_i. Conditioning $P(c|d)$ for each possible event gives

$$P(c|d) = \sum_{t_i} P(c|d, T = t_i)P(T = t_i|d). \quad (11)$$

If we assume conditional independence between c and $T = t_i$, given d, that is $P(c|d, T = t_i) = P(c|T = t_i)$, we obtain

$$P(c|d) = \sum_{t_i} P(c|T = t_i)P(T = t_i|d). \quad (12)$$

Using Bayes' theorem, this becomes

$$P(c|d) = P(c) \sum_{t_i} \frac{P(T = t_i|c)P(T = t_i|d)}{P(T = t_i)}. \quad (13)$$

All the probabilities in Eq. (13) can be estimated from given training data based on the following instructions.

- $P(T = t_i|c)$ is the probability that a randomly extracted term in a document is t_i, given that the document is assigned to c. This is estimated from the within-document term frequency of t_i in the training cases that are categorized into c.
- $P(T = t_i|d)$ is the probability that a randomly extracted term in a target document d is t_i. This is estimated from the within-document term frequency of t_i in d.
- $P(T = t_i)$ is the prior probability that a randomly extracted term in a randomly selected document is t_i. This is estimated from the within-document term frequency of t_i in the training data.
- $P(c)$ is the prior probability that a randomly selected document is categorized into c. This is estimated from the number of documents that are categorized into c in the training data.

Here, let us recall the three problems of PRW. Since SVMV's primitive probabilities are based on within-document term frequencies, SVMV does not have problem 1. Furthermore, SVMV does not have problem 2 either because Eq. (13) includes a factor $P(T = t|d)$, which accomplishes term weighting for the target document d.

For an example of problem 3, let us re-consider the previous example; R documents in the training data are categorized into a category c, none of the R documents has term t_i, but a target document d does. If the straightforward estimate of $P(T_i = 1|c) = 0$ or $P(T = t_i|c) = 0$ is adopted, the document d would never be categorized into c in any of the previous models (PRW, CT, and RPI). In SVMV, the probability of $P(c|d)$ is not affected by such estimates. This is because $P(c|d)$ in Eq. (13) takes the sum of each term's weight. In this example, the weight for t_i is estimated to be 0, as in the other models, but this does not affect the total value of $P(c|d)$. A similar argument applies to all other problems in [9] that are caused by having insufficient training cases. SVMV is *formally* proven not to suffer from the serious effects (like never being assigned to a category or always being assigned to a category) of having insufficient training cases. In other words, SVMV can directly use the straightforward estimates. In addition, we experimentally verified that the value of $P(d|c)$ in SVMV is always a legitimate value (i.e., 0 to 1) unlike in RPI.

Table 1 summarizes the characteristics of the four probabilistic models.

Table 1 Summary of the four probabilistic models

	PRW	CT	RPI	SVMV
solve problem 1	×	○	×	○
solve problem 2	×	○	○	○
solve problem 3	×	×	×	○
satisfy PRP	○	×	○	○

As illustrated in the table, SVMV has the best characteristics for text categorization. In the table, we can also see the ordering of the four probabilistic models,

$$[A1] \qquad \text{PRW} < \text{CT}, \text{RPI} < \text{SVMV}.$$

If we ignore the factor for the probabilistic ranking principle, we have total ordering such that

$$[A2] \qquad \text{PRW} < \text{RPI} < \text{CT} < \text{SVMV}.$$

In the next section, we will experimentally verify SVMV's superiority and the ordering assumptions in [A1] and [A2].

4 Experiments

This section describes experiments conducted to evaluate the performance of our model (SVMV) compared to the other three (PRW, CT, and RPI).

4.1 Data and Preprocessing

A collection of Wall Street Journal (WSJ) full-text news stories [7] was used in the experiments. We extracted 12,380 articles from 1989/7/25 to 1989/11/2[1].

The WSJ articles from 1989 are indexed into 78 categories (topics). Articles having no category were excluded. 8,907 articles remained; each having 1.94 categories on average. The largest category is "TENDER OFFERS, MERGERS, ACQUISITIONS (TNM)" which encompassed 2,475 articles; the smallest one is "RUBBER (RUB)", assigned to only 2 articles. On average, one category is assigned to 443 articles.

All 8,907 articles were tagged by the Xerox Part-of-Speech Tagger [2]. From the tagged articles, we extracted

[1]These are all 1989's articles contained in [7].

the root words of nouns using the "ispell" program[2]. As a result, each article has a set of root words representing it, and each element in the set (i.e. root word of a noun) corresponds to a term.

Before the experiments, we divided 8,907 articles into two sets; one for training (i.e., for probability estimation), and the other for testing. The division was made according to chronology. All articles that appeared from 1989/7/25 to 1989/9/29 went into a training set of 5,820 documents, and all articles from 1989/10/2 to 1989/11/2 went into a test set of 3,087 documents.

4.2 Category Assignment Strategies

In the experiments, the probabilities, $P(c)$, $P(T_i = 1|c)$, $P(T = t_i|c)$, and so forth, were estimated from the 5,820 training documents, as described in the previous sections. Using these estimates, we calculated the posterior probability ($P(c|d)$) for each document (d) of the 3,087 test documents and each of the 78 categories (c). The four probabilistic models are compared in this calculation.

There are several strategies for assigning categories to a document based on the probability $P(c|d)$. The simplest one is the *k-per-doc* strategy [3] that assigns the top k categories to each document. A more sophisticated one is the *probability threshold* strategy, in which all the categories above a user-defined threshold are assigned to a document.

Lewis proposed the *proportional assignment* strategy based on the probabilistic ranking principle [6]. Each category is assigned to its top scoring documents in proportion to the number of times the category was assigned in the training data. For example, a category assigned to 2% of the training documents would be assigned to the top scoring 0.2% of the test documents if the proportionality constant was 0.1, or to 10% of the test documents if the proportionality constant was 5.0.

4.3 Results and Discussions

By using a category assignment strategy, several categories are assigned to each test document. The best known measures for evaluating text categorization/retrieval models are *recall* and *precision*, calculated by the following equations [6]:

$$\text{Recall} = \frac{\text{the number of categories that are correctly assigned to documents}}{\text{the number of categories that should be assigned to documents}},$$

$$\text{Precision} = \frac{\text{the number of categories that are correctly assigned to documents}}{\text{the number of categories that are assigned to documents}}.$$

[2]Ispell is a program for correcting English spelling. We used the "ispell version 3.0" which is available for anonymous ftp from `ftp.cs.ucla.edu` in the `pub/ispell` directory.

Note that recall and precision have somewhat mutually exclusive characteristics. To raise the recall value, one can simply assign many categories to each document. However, this leads to a degradation in precision; i.e., almost all the assigned categories are false. A *breakeven* point might be used to summarize the balance between recall and precision, the point at which they are equal.

We first evaluated the three category assignment strategies. Table 2 lists the optimal breakeven points identified for the three strategies.

Table 2 Optimal breakeven points for three category assignment strategies

	Breakeven Pts.
Prop. assignment	0.63 (by SVMV)
Prob. thresholding	0.47 (by SVMV)
k-per-doc	0.43 (by SVMV)

From Table 2, we find that SVMV with proportional assignment gives the best result (0.63). The superiority of proportional assignment over the other strategies has already been reported by Lewis [6]. Our experiment verified Lewis' assumption. Note that in any of the three strategies, SVMV (our model) gives the highest breakeven point.

Figure 1 shows the recall/precision trade off for the four probabilistic models with proportional assignment strategy. As a reference, the recall/precision curve of a well-known vector model [11] ("TF·IDF") is also presented. Table 3 lists the breakeven point for each model.

Fig. 1 Recall/precision with proportional assignment strategy

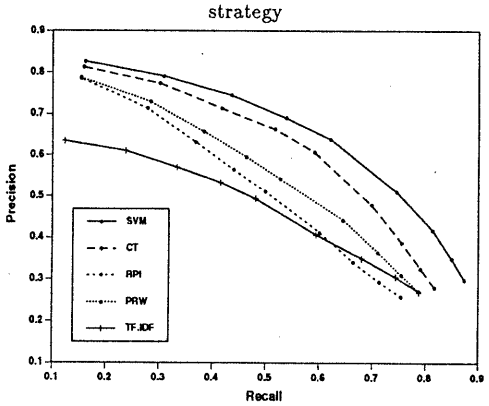

Table 3 Breakeven points with proportional assignment strategy

	Breakeven Pts.
SVMV	0.63
CT	0.60
RPI	0.51
PRW	0.53
TF·IDF	0.48

From Figure 1 and Table 3, we can see the ordering

$$\text{TF·IDF} < \text{RPI} < \text{PRW} < \text{CT} < \text{SVMV}.$$

This verifies that our model (SVMV) is the best model for text categorization. Also, note that this order is almost the same as our prediction [A2]. The only difference is the order of RPI with respect to PRW. We predicted "PRW < RPI" because RPI takes into account term weighting for target documents (i.e., $P(T = 1, 0|d)$). However, since $P(T = 1, 0|d)$ reduces to 1 or 0 in actual estimation, PRW and RPI come to have the same characteristics in Table 1. Thus, it is not surprising that our prediction of "PRW < RPI" was not verified in these experiments.

In summary, we can conclude that:

- SVMV (our model) with proportional assignment strategy is the best probabilistic model for text categorization.
- The models that consider within-document term frequencies (SVMV, CT) are better than those that do not (PRW, RPI).
- The models that consider term weighting for target documents (SVMV, CT) are better than those that do not (PRW, (RPI)).
- The models that are not affected by having insufficient training cases (SVMV) are better than those that are (CT, RPI, PRW).
- All the probabilistic models are superior to the traditional vector model (TF·IDF).

5 Conclusion

We have proposed a new probabilistic model for text categorization. Compared to previous models, our model has the following advantages; 1) it considers within document term frequencies, 2) considers term weighting for target documents, and 3) is not affected by having insufficient training cases. We have also provided empirical results verifying our model's superiority over the others in the task of categorizing news articles from the Wall Street Journal.

References

[1] W. B. Croft. Document representation in probabilistic models of information retrieval. *Journal of the American Society for Information Science*, Vol. 32, No. 6, pp. 451–457, 1981.

[2] D. Cutting and Pedersen J. The xerox part-of-speech tagger version 1.0. annonymous ftp from `parcftp.xerox.com`, 1993.

[3] B. Field. Towards automatic indexing: Automatic assignment of controlled language indexing and classification from free indexing. *Journal of Documentation*, Vol. 31, No. 4, pp. 246–265, 1975.

[4] N. Fuhr. Models for retrieval with probabilistic indexing. *Information Processing & Retrieval*, Vol. 25, No. 1, pp. 55–72, 1989.

[5] K. L. Kwok. Experiments with a component theory of probabilistic information retrieval based on single terms as document components. *ACM Transactions on Information Systems*, Vol. 8, No. 4, pp. 363–386, 1990.

[6] D. D. Lewis. An evaluation of phrasal and clustered representation on a text categorization task. In *Proceedins of the Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 37–50, 1992.

[7] M. Liberman, editor. *CD-ROM I.* Association for Computational Linguistics Data Collection Initiative, University of Pennsylvania, 1991.

[8] S. E. Robertson. The probability ranking principle in IR. *Journal of Documentation*, Vol. 33, pp. 294–304, 1977.

[9] S. E. Robertson and K. Sparck Jones. Relevance weighting of search terms. *Journal of the American Society for Information Science*, Vol. 27, pp. 129–146, 1976.

[10] G. Salton and McGill M. J. *Introduction to Modern Information Retrieval.* McGraw-Hill Publishing Company, 1983.

[11] G. Salton and C. S. Yang. On the specification of term values in automatic indexing. *Journal of Documentation*, Vol. 29, No. 4, pp. 351–372, 1973.

[12] K. Sparck Jones. A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, Vol. 28, No. 1, pp. 11–21, 1972.

[13] S. K. M. Wong and Y. Y. Yao. A probability distribution model for information retrieval. *Information Processing & Management*, Vol. 25, No. 1, pp. 39–53, 1989.

[14] C. T. Yu, W. Meng, and S. Park. A framework for effective retrieval. *ACM Transactions on Database Systems*, Vol. 14, No. 2, pp. 147–167, 1989.