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

Makoto IWAYAMA
Advanced Research Laboratory
Hitachi Ltd.
HATOYAMA
SAITAMA 350-03, JAPAN
iwayama@harl.hitachi.co.jp

Takenobu TOKUNAGA
Department of Computer Science
Tokyo Institute of Technology
2-12-1, ÔOKAYAMA, MEGURO-KU
TOKYO 152, JAPAN
take@cs.titech.ac.jp

Abstract

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 less 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 (Robertson and Sparck Jones, 1976; Kwok, 1990; Fuhr, 1989; Lewis, 1992; Croft, 1981; Wong and Yao, 1989; Yu et al., 1989) 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 solves some problems of the previous models. In section 4, we verify our model's superiority over the 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 have been exploited for information retrieval, but the adaptation to text categorization is straightforward.

In a model of probabilistic text categorization,

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

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) (Robertson, 1977). Several strategies can be used to assign categories to a document based on PRP (Lewis, 1992).

There are several ways to calculate $P(c|d)$. Three representatives are (Robertson and Sparck Jones, 1976), (Kwok, 1990), and (Fuhr, 1989).

2.1 Probabilistic Relevance Weighting (PRW)

Robertson and Sparck Jones (1976) make use of the well-known logistic (or log-odds) transformation of the

probability $P(c|d)$.

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

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

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

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

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

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

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

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

While PRW is the first attempt to formalize well-known relevance weighting (Sparck Jones, 1972; Salton and McGill, 1983) by probability theory, there are several drawbacks in PRW.

[Problem 1] no within-document term frequencies

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 information retrieval (Salton and McGill, 1983). Salton and Yang experimentally verified the importance of within-document term frequencies in their vector model (Salton and Yang, 1973).

[Problem 2] no term weighting for target documents

In the PRW formulation, there is no factor of term weighting for target documents (i.e., $P(\cdot|d)$). According to Eq. (5), even if a term exists in a target document, only the importance of the term in a category (i.e., $P(T = 1|c)$) is considered for overall probability. Term weighting for target documents would also be necessary for sophisticated information retrieval (Fuhr, 1989; Kwok, 1990).

[Problem 3] affected by having insufficient 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 other special cases like the above example (Robertson and Sparck Jones, 1976). A well-known remedy for this problem is to use "$(r + 0.5)/(R + 1)$" as the estimate of $P(T = 1|c)$ (Robertson and Sparck Jones, 1976). While various smoothing methods (Church and Gale, 1991; Jelinek, 1990) are also applicable to these situations and would be expected to work better, we used the simple "add one" remedy in the following experiments.

2.2 Component Theory (CT)

To solve problems 1 and 2 of PRW, Kwok (1990) stresses the assumption that a document consists of terms. 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;

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

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

$$g(c|d) = \sum_{t \in d} P(T = t|d)(\log \frac{P(T = t|c)}{P(T \neq t|c)} + \log \frac{P(T \neq t|\overline{c})}{P(T = t|\overline{c})}) + \log \frac{P(c)}{P(\overline{c})}. \quad (7)$$

For precise derivation, refer to (Kwok, 1990).

Here, note that $P(T = t|d)$ and $P(T = t|c)$ represent the frequency of a term t within a target document d and that within a category c respectively. Therefore, CT is not subject to problems 1 and 2. However, problem 3 still affects CT. Furthermore, Fuhr (1989) 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 (1989) solves problem 2 by assuming that a document is *probabilistically* indexed by its term vectors. 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

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

By assuming conditional independence between c and d given $\boldsymbol{x}$[1], and using Bayes' theorem, Eq. (8) becomes,

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

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

$$P(c|d) = P(c) \prod_i \left(\frac{P(T_i = 1|c) P(T_i = 1|d)}{P(T_i = 1)} + \frac{P(T_i = 0|c) P(T_i = 0|d)}{P(T_i = 0)} \right). \quad (10)$$

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. While this in principle solves problem 2, if we use a simple estimation method counting the number of documents which have a term, $P(T = 1,0|d)$ reduces to 1 or 0 (i.e, binary, not weighted). For example, when a target document d has a term t, $P(t = 1|d) = 1$ and when not, $P(T = 1|d) = 0$. In the following experiments we used this binary estimation method, but non binary estimates could be used as in (Fuhr, 1989).

[1]More precisely, $P(c|d, x) = P(c|x)$ which assumes that if we know x, information for c is independent of that for d. This assumption sounds valid because x is a kind of representation of d.

As far as other problems are concerned, RPI still problematic. In particular, because of problem 3, $P(c|d)$ would become an illegitimate value. In our experiments, as well as in Lewis' experiments (1992), $P(c|d)$ 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 yet solves problems 1, 2, and 3 in PRW.

Document representation of our model is basically the same as CT, that is a document is a set of its constituting terms. The major difference between our model and others is the way of document characterization through probabilities. While almost all previous models assume that an event space for a document is whether the document is indexed or not by a term[2], our model characterizes a document as random sampling of a term from the term set that represents the document. For example, an event "$T = t_i$" means that a randomly selected term from a document is t_i. If we want to emphasis indexing process like other models, it is possible to interpret "$T = t_i$" as a randomly selected element from a document being indexed by the term t_i.

Formally, our model can be seen as modifying Fuhr's derivation of $P(c|d)$ by replacing an index vector with a single random variable whose value is one of possible terms. 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 d, given $T = t_i$, 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 definitions.

- $P(T = t_i|c)$ is the probability that a randomly selected term in a document is t_i, given that the document is assigned to c. We used $\frac{NC_i}{NC}$ as the estimator. NC_i is the frequency of the term t_i in the category c, and NC is the total frequency of terms in c.

[2]In section 2 explaining previous models, we simplified "a document is indexed by a term" as "a document contains a term" for ease of explanation.

- $P(T = t_i|d)$ is the probability that a randomly selected term in a target document d is t_i. We used $\frac{ND_i}{ND}$ as the estimator. ND_i is the frequency of the term t_i in the document d, and ND is the total frequency of terms in d.
- $P(T = t_i)$ is the prior probability that a randomly selected term in a randomly selected document is t_i. We used $\frac{N_i}{N}$ as the estimator. N_i is the frequency of the term t_i in the given training documents, and N is the total frequency of terms in the training documents.
- $P(c)$ is the prior probability that a randomly selected document is categorized into c. We used $\frac{D_c}{D}$ as the estimator. D_c is the frequency of documents that is categorized to c in the given training documents, and D is the frequency of documents in the training documents.

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 a target document d.

For problem 3, let us reconsider 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 the previous models (PRW, CT, and RPI). In SVMV, the probability $P(c|d)$ is much less 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 little affect the total value of $P(c|d)$. A similar argument applies to all other problems in (Robertson and Sparck Jones, 1976) 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) by 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
Problem 1 considered	no	yes	no	yes
Problem 2 considered	no	yes	(yes)	yes
Problem 3 considered	no	no	no	yes
PRP satisfied	yes	no	yes	yes

As illustrated in the table, SVMV has better characteristics for text categorization compared to the previous models. In the next section, we will experimentally verify SVMV's superiority.

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 (Liberman, 1991)[3] was used in the experiments. We extracted all 12,380 articles from 1989/7/25 to 1989/11/2.

The WSJ articles from 1989 are indexed with 78 categories (topics). Articles having no category were excluded. 8,907 articles remained; each having 1.94 categories on the 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 the average, one category is assigned to 443 articles.

All 8,907 articles were tagged by the Xerox Part-of-Speech Tagger (Cutting et al., 1992)[4]. From the tagged articles, we extracted the root words of nouns using the "ispell" program[5]. 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. We did not reduce the number of terms by using stop words list or feature selection method, etc. The number of terms amounts to 32,975.

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

[3] We used "ACL/DCI (September 1991)" CD-ROM which is distributed from the Linguistic Data Consortium (LDC). For more details, please contact Mark Liberman (myl@unagi.cis.upenn.edu).

[4] The xerox part-of-speech tagger version 1.0 is available via anonymous FTP from the host parcftp.xerox.com in the directory pub/tagger.

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

(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 (Field, 1975) 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 (Lewis, 1992). 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 models are *recall* and *precision*, calculated by the following equations (Lewis, 1992):

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

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

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.

For each strategy, we calculated breakeven points by using the four probabilistic models. Table 2 shows the best breakeven points identified for the three strategies along with the used models.

Table 2 Best 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 (1992). Our experiment verified Lewis' assumption. In addition, for any of the three strategies, SVMV gives the highest breakeven point among the four probabilistic models.

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 (Salton and Yang, 1973) ("TF·IDF")[6] is also presented. Table 3 lists the breakeven point for each model. All the breakeven points were obtained when proportionality constant was about 1.0.

Fig. 1 Recall/precision with proportional assignment strategy

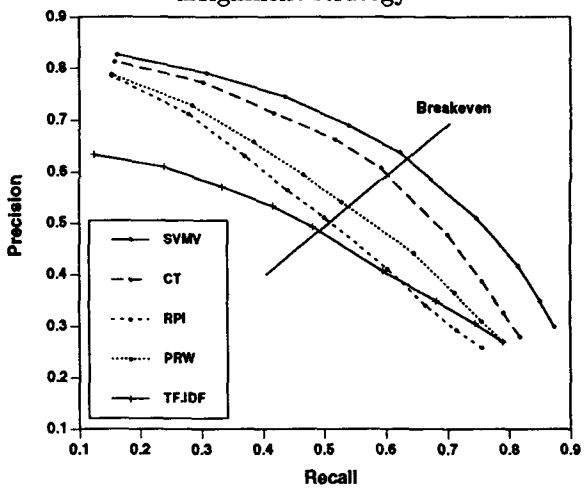

Table 3 Breakeven points with proportional assignment strategy

	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 that:

- as far as this dataset is concerned, SVMV with proportional assignment strategy gives the best result among the four probabilistic models,
- the models that consider within-document term frequencies (SVMV, CT) are better than those that do not (PRW, RPI),

[6] In the model we used, each element of document vector is the "term frequency" multiplied by the "inverted document frequency." Similarity between every pair of vectors is measured by cosine. Note that this is the simplest version of TF·IDF model, and there has been many improvements which we did not consider in the experiments.

- the models that consider term weighting for target documents (SVMV, CT) are better than those that do not (PRW, (RPI)), and
- the models that are less affected by having insufficient training cases (SVMV) are better than those that are (CT, RPI, PRW).

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 less 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."

There are several directions along which this work could be extended.

- We have to compare our probabilistic model to other non probabilistic models like decision tree/rule based models, one of which has recently been reported to be promising (Apté et al., 1994).
- While we used simple document representation in which a document is defined as a set of nouns, there could be considered several improvements, such as using phrasal information (Lewis, 1992), clustering terms (Sparck Jones, 1973), reducing the number of features by using local dictionary (Apté et al., 1994), etc.
- We are incorporating our probabilistic model into cluster-based text categorization that offers an efficient and effective search strategy.

Acknowledgments

The authors are grateful to Hiroshi Motoda for beneficial discussions, and would like to thank the anonymous reviewers for their useful comments.

References

C. Apté, F. Damerau, and S. M. Weiss. 1994. Automated learning of decision rules for text categorization. *ACM Transactions on Office Information Systems*. (to appear).

K. W. Church and W. A. Gale. 1991. A comparison of the enhanced Good-Turing and deleted estimation methods for estimating probabilities of English bigrams. *Computer Speech and Language*, 5:19–54.

W. B. Croft. 1981. Document representation in probabilistic models of information retrieval. *Journal of the American Society for Information Science*, 32(6):451–457.

D. Cutting, J. Kupiec, J. Pedersen, and P. Sibun. 1992. A practical part-of-speech tagger. In *In Proc. of the Third Conference on Applied Natural Language Processing*.

B. Field. 1975. Towards automatic indexing: Automatic assignment of controlled language indexing and classification from free indexing. *Journal of Documentation*, 31(4):246–265.

N. Fuhr. 1989. Models for retrieval with probabilistic indexing. *Information Processing & Retrieval*, 25(1):55–72.

F. Jelinek. 1990. Self-organized language modeling for speech recognition. In A. Waibel and K. Lee, editors, *Readings in Speech Recognition*, pages 450–506. Morgan Kaufmann.

K. L. Kwok. 1990. Experiments with a component theory of probabilistic information retrieval based on single terms as document components. *ACM Transactions on Information Systems*, 8(4):363–386.

D. D. Lewis. 1992. 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*, pages 37–50.

M. Liberman, editor. 1991. *ACL/DCI (CD-ROM)*. Association for Computational Linguistics Data Collection Initiative, University of Pennsylvania, September.

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

S. E. Robertson. 1977. The probability ranking principle in IR. *Journal of Documentation*, 33:294–304.

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

G. Salton and C. S. Yang. 1973. On the specification of term values in automatic indexing. *Journal of Documentation*, 29(4):351–372.

K. Sparck Jones. 1972. A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28(1):11–21.

K. Sparck Jones. 1973. Collection properties influencing automatic term classification performance. *Information Storage and Retrieval*, 9:499–513.

S. K. M. Wong and Y. Y. Yao. 1989. A probability distribution model for information retrieval. *Information Processing & Management*, 25(1):39–53.

C. T. Yu, W. Meng, and S. Park. 1989. A framework for effective retrieval. *ACM Transactions on Database Systems*, 14(2):147–167.