Proceedings of Research on Computational Linguistics Conference XII,
Hsinchu, 1999, pp. 171-185

Automatically Controlled-Vocabulary Indexing for Text Retrieval

Kuang-hua Chen and Chien-tin Wu

Department of Library and Information Science

National Taiwan University

1, SEC. 4, Roosevelt RD.

Taipei, TAIWAN, 10617, R.O.C.

E-mail: khchen@ccms.ntu.edu.tw; jtwu@steelman.lis.ntu.edu.tw

Abstract

The IR society has made efforts in free-term indexing for a long time. By contrast, few efforts are made in controlled-vocabulary indexing. A new model for controlled-vocabulary indexing is proposed in this paper. This proposed model, TF×OSDF×CSIDF, distinguishes subject-specific words from common words and domain-specific words in documents. 60,400 MEDLINE records are used as training data and testing data and 100 MeSH subject headings are used as the testing controlled vocabularies. The preliminary experiments show good results. The precision and the recall concurrently exceed 90% using abstracts as training materials. The precision reaches 90% and the recall still keeps at 70% using title only. The problem of indexer's consistency could be alleviated using the proposed model to automatically generate index terms.

1. Introduction

The quality of indexing not only depends on professional knowledge and experience of librarians or subject specialists, but also is restricted by time and cost. Lack of indexing

experts and subject specialists, the information exploration has confronted libraries with manpower problem. In addition, the issue of indexer consistency still cannot be resolved effectively. In 1950, researchers started to employ machine to enhance indexing process. In recent years, the Internet has made the subject access become the mainstream of information seeking behavior and prompted researches of automatic indexing, classification and abstracting. The researchers of automatic indexing always take complete substitution of human indexing as the ultimate goal. Although there is a long way to go, many researchers claim that the performance of automatic indexing is the same as that of manual indexing at least. (Cleverdon and Mills, 1963; Cleverdon, 1976)

Most researches of automatic indexing focus on the free-term indexing. (Salton, 1988; Ponte and Bruce Croft, 1998) By contrast, the researchers do not pay much attention to the automatic indexing for controlled vocabularies. The free-term indexing is to identify keywords or key phrases, which represent subjects of document and use them as index terms directly. Basically, these keywords and key phrases couldn't represent true "concept" of user's information need. As to controlled-vocabulary indexing, indexer has to translate subject concepts into controlled vocabularies. From this viewpoint, controlled-vocabulary indexing may be regarded as concept indexing. Besides, the free-term indexing usually increases recall rather than precision. This has pushes researchers to study controlled-vocabulary indexing for information retrieval again.

This paper proposes a new model, uses titles and abstracts of documents which have been indexed manually as the training materials, and makes controlled-vocabulary indexing automatic easily. Section 2 discusses the idea and proposes the new model. Section 3 describes the design of experiments and carries out a series of experiments. Section 4 discusses the experimental results in detail. Section 5 is the short conclusions.

2. The Idea and the Proposed Model

The idea behind the proposed model is based on the content-bearing words. It is assumed that there should be some kind of relationships among controlled vocabularies and content-bearing words. If some content-bearing words are found in a document, the related controlled vocabulary should be assigned to the document.

The training process will construct a function between document and subject headings (a set of controlled vocabularies). After this function is determined, documents could be transferred into correspondent feature values, and then calculate indexing scores of documents for certain subject headings. Indexing score implies the possibility that documents are indexed in some subject headings.

The previous researches on automatic indexing have been associated with the exploitation of statistical techniques. Luhn (1997) considered that the justification of measuring word significance by use-frequency is based on the fact that a writer normally repeats certain words as he advances or varies his arguments and he elaborates on an aspect of a subject. Salton (1989) suggested that the general aim of statistical measures be to reject both very high and very low frequency words from the texts being indexed.

In tradition, the considered types of frequency are term frequency (TF) and document frequency (DF). DF is often transferred into inverse document frequency (IDF) while adopted. TF indicates occurrence of word in a document, while DF refers to distinct occurrence of word in a document collection. If there are N documents in a document collection, IDF is represented as $log(N/DF)$. (Sparck Jones, 1972) Low IDF will decrease weight of word and make word be rejected from the candidate list of index terms. Although IDF has adopted widely in various indexing models and has been verified as an effective measure for weighting words, it also filters out some kind of words with great benefit in subject

identification. Take Figure 1 as an example.

A+B+C = Words with high DF and low IDF

A = Common Words

B = Domain-specific Words

C = Subject-specific Words

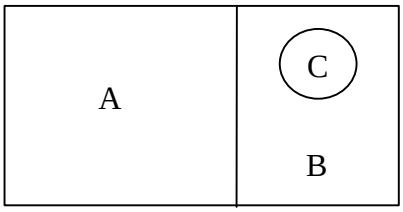

Figure 1. Words with Low IDF

In Figure 1, the largest rectangle indicates the words with low IDF. A, B and C areas within rectangle refer to common words, domain-specific words and subject-specific words, respectively. The words of A and B cannot offer useful information in subject identification actually. For example, in documents discussing education, there are high-frequency words, such as “education”, “school”, “teacher”, and “student”, which are not discriminative enough. Unlike words of A and B, the words of C benefit subject identification greatly. Take documents concerning about AIDS as instance. Occurrence of AIDS in such kind of documents should be very high. If IDF is adopted as unique measurement, then it will assign low weight to AIDS which reflects subject of these documents, and then AIDS will be rejected finally.

In general, there should exist some subject-specific words in the documents with the same subject, which are highly close to the subject and with high occurrence. As mentioned above, although subject-specific words are very useful, the IDF measurement cannot distinguish them from common words and domain-specific words in a document collection with the same subject. Therefore, we could not use IDF directly. That is to say, the indexing model has to be enhanced with a capacity to separate subject-specific words from high DF words and increase weight of subject-specific words while evaluating significance of words. Increasing no additional training documents, a new model is proposed for solving problem mentioned above. In this new model, the training documents, which originally belong to

distinct subject headings, will be combined into a document. Please refer to Figure 2. For convenience, the following description takes original training documents as “original set”, and the combined documents as “combined set”. The new model will weight words using the multiplication of TF in documents, DF in original set (OSDF) and IDF in combined set (CSIDF).

The distributional tendency of common words, domain-specific words and subject-specific words are shown in Table 1. Common words and domain-specific words will be of lower CSIDF. CSIDF of a word will not be changed for different subject headings. Therefore, common words and domain-specific words will be of low weight when they found in documents of different subjects. Unlike CSIDF, OSDF of a word varies by different subject headings, and some subject-specific words with high weight will be figured out for certain subject headings.

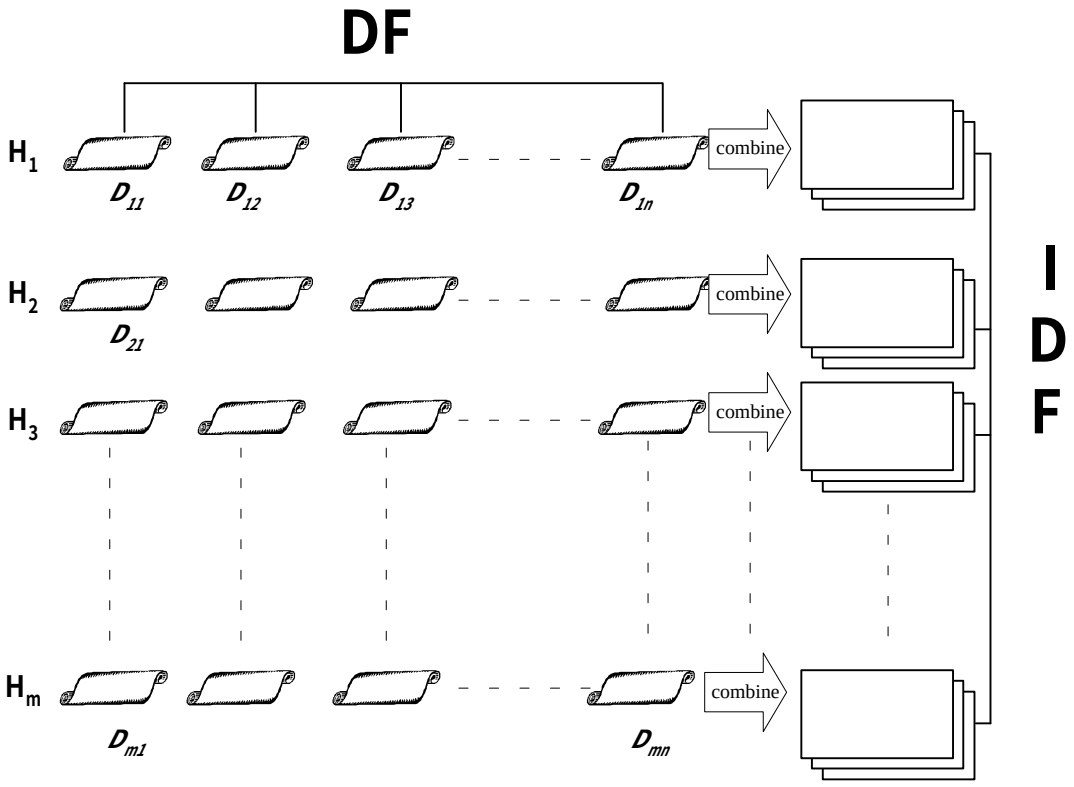

Figure 2. Diagram for Calculation of Term Weight

Table 1. Distributional Tendency of Words

	OSDF	CSIDF
Common Words	High	Low
Domain-specific Words	High	Low
Subject-specific Words	High	High

3. Experiments

3.1 Learning Process

We choose MEDLINE (MEDlars onLINE) as the source of training documents (MEDLINE, 1998), and MeSH (Medical Subject Headings) as the controlled vocabularies. (Medical Subject Headings, 1998) Since the documents were collected from late of 1997, subject headings for training are extracted from 1997 MeSH Tree, not 1998 MeSH Tree. A sample training text is shown in Figure 3.

Title
A method to test blood flow limitation of peritoneal-blood solute transport.
Local Messages
Undefined
Abstract
Current transperitoneal transport models assume that effective blood flow to the microcirculation does not limit solute exchange with dialysate in the cavity. Despite evidence that gas transfer across the peritoneum (assumed to equal the effective blood flow) occurs at rates that exceed maximum urea transfer rates by a fact or of two to three, the assumption has been strongly challenged. To address this problem at the tissue level, a technique to determine the effect of local blood flow on small-solute transport was developed in this study. Diffusion chambers were affixed to the serosal side of the anterior abdominal wall of rats, and solutions containing radiolabeled urea or mannitol were placed in the chambers. During each experiment, the local blood flow beneath the chamber was monitored with laser Doppler flowmetry and the disappearance of the tracer versus time was simultaneously measured under three conditions of blood flow: control, 30% of control, and zero blood flow. The results demonstrated no significant differences for either solute between control and the condition in which blood flow was reduced by 70%. However, there was a significant reduction in the rate of mass transfer with no blood flow. It was concluded that blood flow at > or = 30% of control values does not limit solute transfer across the abdominal wall peritoneum during dialysis.

Figure 3. Sample Text of MEDLINE Record

One hundred of subject headings of 1997 MeSH Tree were selected for training and testing. The selection of subject headings is a crucial step. In order to test the model in terms of average performance, the extracted subject headings should equally distribute in the MeSH Tree. This could avoid subject headings concentrating on certain fields.

Besides distribution, both depth and width of the subject headings in MeSH Tree have to be taken into account. On the one hand, the amount of records in database could be used to judge the width of a subject heading. Therefore, the subject headings associated with 1,000 to 2,000 records from 1991 to 1997 in MEDLINE were chosen. On the other hand, the distance from root or leaf node to a subject heading could reflect relative depth. The subject headings with the following criteria are chosen.

- Distance to root equal or more than 1 layer and less than 4 layers
- Distance to leaf are less than 6 layers

The average distances to root and to leaf are 1.58 and 1.62, respectively.

600 records for each subject heading under consideration are collected. 400 out of 600 records are for positive training; the others are for positive testing. There are totally 400 records collected for negative testing, which are not indexed by any subject headings under consideration. Table 2 lists detailed statistics of records in the three sets. The collected records only contain titles and abstracts. The volume of whole experimental records is 84 MB: positive training set is 56 MB; positive testing set is 27.8 MB; negative testing set is 654 KB. The collected records are English documents and the average number of words in abstract and title are 107.6 and 11.3, respectively, after filtering out the stop words.

Assume there are m subject headings, H_1, H_2, H_3, …, H_m and l distinct words, W_1, W_2, W_3,…, W_l. We will take H_j as instance to illustrate the following experimental processes.

Table 2. Amount of Experimental Records

	Training Set	Testing Set	Total
Positive	40,000	20,000	60,000
Negative	--	400	400
Total	40,000	20,400	60,400

The OSDF and CSIDF of words in the training set are calculated. Formula of CSIDF is shown below,

$$CSIDF(W) = \log_2\left(\frac{P - O(W)}{O(W)}\right)$$

whereas P represents the amount of documents in combined set, $O(W)$ is the number of documents which contain word W. CSIDF is negative when W appears more than half of documents in combined set. After this step, relationship between H_j and a set of words R_j is constructed. Words with high weight in R_j have the high possibility to be subject-specific words. R_j can be regarded as the weighted vector shown as follows,

Subject Heading	Weighted Vector
H_1	$R_1=\{w_{11}, w_{12}, \ldots, w_{1k}, \ldots w_{1l}\}$
H_2	$R_2=\{w_{21}, w_{22}, \ldots, w_{2k}, \ldots w_{2l}\}$
…	…
H_j	$R_j=\{w_{j1}, w_{j2}, \ldots, w_{jk}, \ldots w_{jl}\}$
…	…
H_m	$R_m=\{w_{m1}, w_{m2}, \ldots, w_{mk}, \ldots w_{ml}\}$

whereas w_{jk} is the weight (OSDF×CSIDF) in R_j for the word W_{jk}.

3.2 Evaluating Process

After R_j has been constructed, we will calculate indexing score (*IS*) of a document for each subject heading according to the weighted vectors. The *IS* is shown as follows.

$$IS = \frac{\sum (OSDF \times CSIDF) \times (TF)}{\text{number of words in the document}}$$

The normalization is used to avoid favoring the lengthy documents. When an unseen document appears, the process of automatic indexing will preprocess it first, and then compute the indexing score based on each subject heading. The lower *IS* indicates that this document should not be indexed by H_j; the higher *IS* indicates that it is likely to assign H_j as one index term for this document.

An indexing threshold T_j, is determined to distinguish documents with high *IS* from those with low *IS*. If *IS* is larger than T_j, the subject heading H_j will be assigned to the document. Otherwise, it will not. Because, it is not easy to determine a threshold, ten thresholds from 0.1 to 1.0 were used in our experiments. Finally, one best threshold will be chosen.

The precision and recall are used for performance evaluation. Note that the precision and recall are from viewpoint of subject headings rather than from documents. The precision and recall of individual subject heading (H_1-H_{100}) will be merged to compute the average precision and average recall for the final evaluation.

4. Experimental Results

There are four sets of data in experiments: data for abstract part both in training set and testing set, data for title part both in training set and testing set. The detailed results of abstract part could be referred to Table 3; those of title part can be referred to Table 4.

4.1 The Precision vs the Recall

Let's consider the performance for the abstract part of training set. When the threshold is between 0.3 and 0.6, both the precision and recall are higher than 90%. When the threshold equals to 0.43, both the precision and recall are higher than 94% according to the interpolation. Consider the performance for the abstract part of testing set. While the threshold equals to 0.43, both the precision and recall are higher than 90%.

Consider the performance for the title part of training set. When the threshold equals to 0.4, the precision is higher than 90% and recall is higher than 70%. When the threshold equals to 0.1, both the precision and recall are higher than 76%. Consider the performance for the title part of testing set. When the threshold is close to 0.3, the precision reaches 90% and recall reaches 70%. When the threshold is 0.1, both the precision and recall are higher than 78%.

Table 3. Precision & Recall of Abstract Part (TF×OSDF×CSIDF)

Threshold	Training Set			Testing Set		
	Precision	Recall	F-score	Precision	Recall	F-score
0.1	77.31%	99.62%	87.06%	76.78%	96.63%	85.57%
0.2	87.17%	98.83%	92.63%	86.47%	92.90%	89.57%
0.3	91.68%	97.36%	94.43%	91.01%	89.49%	90.24%
0.4	93.92%	95.32%	94.61%	93.32%	86.19%	89.61%
0.5	95.49%	92.88%	94.17%	95.00%	83.39%	88.82%
0.6	96.33%	90.24%	93.19%	95.91%	80.62%	87.60%
0.7	96.92%	87.31%	91.86%	96.56%	77.87%	86.21%
0.8	97.32%	84.51%	90.46%	97.01%	75.51%	84.92%
0.9	97.62%	81.69%	88.95%	97.35%	73.15%	83.53%
1.0	97.83%	79.09%	87.47%	97.59%	71.09%	82.26%

Table 4. Precision & Recall of Title Part (TF×OSDF×CSIDF)

Threshold	Training Set			Testing Set		
	Precision	Recall	F-score	Precision	Recall	F-score
0.1	80.52%	76.48%	78.45%	80.87%	78.24%	79.53%
0.2	86.85%	74.86%	80.41%	86.78%	74.38%	80.10%
0.3	89.94%	72.86%	80.50%	89.67%	70.76%	79.10%
0.4	91.92%	70.73%	79.94%	91.54%	67.34%	77.60%
0.5	92.98%	68.50%	78.88%	92.57%	64.48%	76.01%
0.6	93.60%	66.09%	77.48%	93.18%	61.76%	74.28%
0.7	94.10%	63.78%	76.03%	93.71%	59.58%	72.85%
0.8	94.51%	61.47%	74.49%	94.14%	57.67%	71.52%
0.9	94.88%	59.30%	72.98%	94.56%	55.67%	70.08%
1.0	95.19%	57.12%	71.40%	94.92%	53.92%	68.77%

4.2 The Abstract vs the Title

Basically, words in titles are much fewer than those in abstracts. Therefore, the performance of title part is supposed to be unstable. The experimental results show the consistency with this prediction. In comparison with abstract part, the recall of title part is 80% of that of abstract part. Although the recall agrees with the original supposition, the precision is much better than the predicted one. In fact, the precision of title part could reach 95%. Generally speaking, the title contains lots of useful information, which is very effective in subject identification, and worthy of using in the construction of indexing models.

Take testing subject heading 001 as instance. Table 5 is the statistics of average indexing score of testing documents in abstract and title parts. In 200 documents indexed by subject heading 001, the indexing scores of title part are divergent. The higher standard deviation and range reveal unbalanced distribution of indexing scores and imply the higher

possibility of error. Although the recall in title part decreases, the precision does not drop too much. The stable precision indicates that once the title provides information, it will be useful.

Table 5. Indexing Score of Documents in Testing Set

Heading 001		Range	Min.	Max.	Mean	Std. Dev.
Positive Document	Abstract	16.70	0.18	16.72	3.53	3.45
	Title	38.96	0.00	38.96	8.27	8.35

4.3 The Proposed Model vs the Traditional Model

We carry out the same experiments for the traditional model as a baseline model. These experiments are divided into two parts: one is for TF×OSIDF; the other is for TF×CSIDF.

The experimental results of TF×OSIDF show that the precision and the recall are zero when threshold is 0.1. Obviously, subject-specific words play the crucial role in subject identification. As to TF×CSIDF, although the training documents belong to different subjects, the weight of subject-specific word is not enhanced without the aid of OSDF. Despite of the better performance than TF×OSIDF, the performance of indexing model using TF×CSIDF is still inferior to our model. Table 6 and Table 7 show the performance of TF×CSIDF in details.

In comparison with the traditional model in terms of recall (please refer to Table 3 and 4), our model not only shows less diversity in the training set and the testing set, but also performs stably. Because our model identifies the importance of subject-specific words, we shorten the gap between training set and testing set. By contrast, the recall of traditional model in testing set drops quickly. The largest difference between traditional model and our model in recall is higher than 94%; the least difference is also higher than 60%. In terms of precision, some results of the traditional model are better than those of our model. However, it sacrifices the recall to the precision.

Table 6. Recall and Precision of Abstract Part (Traditional Model TF×CSIDF)

Threshold	Training Set			Testing Set		
	Precision	Recall	F-score	Precision	Recall	F-score
0.1	86.67%	97.80%	91.90%	84.86%	84.30%	84.58%
0.2	93.33%	89.01%	91.12%	90.51%	60.65%	72.63%
0.3	95.08%	73.26%	82.76%	91.18%	39.20%	54.83%
0.4	96.10%	54.45%	69.51%	91.54%	23.92%	37.93%
0.5	97.09%	38.04%	54.66%	92.62%	14.30%	24.77%
0.6	98.53%	24.84%	39.68%	95.55%	7.94%	14.66%
0.7	99.81%	15.46%	26.77%	99.34%	4.52%	8.65%
0.8	99.89%	9.47%	17.30%	99.60%	2.46%	4.80%
0.9	100.00%	5.70%	10.79%	100.00%	1.36%	2.68%
1.0	100.00%	3.45%	6.67%	100.00%	0.71%	1.41%

Table 7. Recall and Precision of Title Part (Traditional Model TF×CSIDF)

Threshold	Training Set			Testing Set		
	Precision	Recall	F-score	Precision	Recall	F-score
0.1	87.81%	90.38%	89.08%	84.94%	70.80%	77.23%
0.2	92.34%	84.16%	88.06%	89.35%	58.54%	70.74%
0.3	94.43%	76.79%	84.70%	91.31%	47.57%	62.55%
0.4	96.46%	68.30%	79.97%	93.71%	37.38%	53.44%
0.5	98.00%	59.40%	73.97%	95.99%	28.93%	44.46%
0.6	99.24%	50.77%	67.17%	98.27%	22.18%	36.19%
0.7	99.69%	42.44%	59.53%	99.24%	17.08%	29.14%
0.8	99.91%	34.94%	51.77%	99.77%	12.90%	22.85%
0.9	99.96%	28.33%	44.15%	99.90%	9.84%	17.92%
1.0	100.00%	22.64%	36.92%	100.00%	7.36%	13.71%

5. Conclusions

A new indexing model is proposed for controlled-vocabulary indexing in this paper. Increasing no additional training documents, the new model uses various frequencies through combination and separation of the same training documents, and distinguishes subject-specific words from common words and domain-specific words. The preliminary experiments show good results using 100 MeSH subject headings and 60,400 abstracts and titles. The precision and recall concurrently exceed 90% using abstracts as training materials. As to title, the precision reaches 90% and the recall still keeps at 70%.

The future works should consider phrase terms, enhance the indexing procedure, and test the performance for full texts. Firstly, phrases bear more semantic information than single words. Therefore, the performance of indexing model will be improved using phrase terms. Secondly, it's not efficient for a system to compute index features of all controlled vocabularies in the present design. Clustering could be employed to deal with the problem. Thirdly, there are more and more online full-text databases in recent years. We could use full texts as training materials rather than abstracts and titles.

References

Cleverdon, C. W. (1976), "The Cranfield Tests on Indexing Language Devices," *Aslib Proceedings*, vol. 19, no. 6, pp. 173-194.

Cleverdon, C. W. and J. Mills (1963), "The Testing of Index Language Devices," *Aslib Proceedings*, vol.15, no. 4, pp.106-130.

Luhn, H. P. (1997), "The Automatic Derivation of Information Retrieval Encodements from Machine-Readable Texts," *Readings in Information Retrieval*, Morgan Kaufmann Publishers, Inc., San Francisco, pp. 21-24.

Ponte, J.M. and W. Bruce Croft (1998), "A Language Modelling Approach to Information Retrieval," *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 275-281.

Salton, Gerard (1988), "Syntactic Approaches to Automatic Book Indexing," *Proceeding of the 26th Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, NJ, pp.120-138.

Salton, Gerard (1989), *Automatic Text Processing: the Transformation, Analysis and Retrieval of Information by Computer*, Addison-Wesley Publishing Company, Inc., New York, p. 275.

Sparck Jones, K. (1972), "A Statistical Interpretation of Term Specificity and Its Application in Retrieval," *Journal of Documentation*, vol. 28, no. 1, pp.11-21.