Regression Model and Query Expansion for NTCIR-2 Ad Hoc Retrieval Task

Kazuaki KISHIDA
Faculty of Cultural Information Resources, Surugadai University
698 Azu, Hanno, Saitama 357-8555 Japan
kishida@surugadai.ac.jp

Abstract

This paper describes procedures and results in a monolingual retrieval experiment using NTCIR-2 test collection. First, we discuss a simplified logistic regression model, which enable us to adjust the regression model for working well in each of various document databases. To do automatically the adjustment, a method for estimating parameters in the regression model, is developed based on a kind of classical discriminant analysis. Second, a query expansion techniques is introduced to enhance effectiveness of search for short queries. In order to expand the original short query, we try to select some terms from an automatically constructed global thesaurus, and add them to a set of original search terms. The thesaurus is made from terms in the title field and the author keyword filed in each document records included in the test collection. The term-term relationship is measured by co-occurrence frequency of each pair of terms in both of two fields, and the degree of relationship is used for deciding newly added term and its weight in the query.

Keyword: *Information Retrieval, Regression Model, Automatic Query Expansion, Automatic Thesaurus Construction*

1 Introduction

Regression model is widely used for data analysis or prediction in various scientific fields. There is no enough reason that we reject usefulness of regression model for predicting relevant documents in the context of information retrieval (IR). For example, Fuhr and Buckley[1] have discussed usage of regression for retrieval model, and also, a logistic regression model developed at UC Berkeley, often shows good performance at retrieval experiments such as TREC[2] or NTCIR[3].

The main purpose of this paper is to describe procedures and results in our monolingual retrieval experiment for investigating a regression model by using the NTCIR-2 test collection. Our concern is with a method for adapting the regression model into each of various document databases, i.e., estimating reasonably regression coefficients in the context of IR. This paper attempts to develop a simple regression model in which the parameters are able to be estimated easily, and to apply a kind of discriminant analysis to the parameter estimation.

Another focus of this paper is on query expansion technique for short queries. In many cases, it is difficult for actual users to find out good search terms that work well for retrieval. One of the solutions is to expand the original query by adding some new terms automatically or semi-automatically. This paper explores an automatic method for query expansion through an automatically constructed global thesaurus, which holds information on term-term relationships measured by co-occurrence.

This paper is organized as follows. First, in Section 2, a simplified logistic regression model and a method for parameter estimation will be introduced. Second, in Section 3, a query expansion technique using automatically constructed thesaurus will be described. In Section 4, we will discuss some empirical findings from a monolingual retrieval experiment using the Japanese test collection of NTCIR-2. Finally, some results from an experiment using the English collection of NICIR-2, will be briefly shown in Section 5.

2 Simplified Logistic Regression Model and Parameter Estimation

2.1 Simplified logistic regression model

As already mentioned, it has been shown that retrieval effectiveness of the UC Berkeley's logistic regression model indicates good effectiveness at NTCIR-1 ad hoc task[3]. The model is

$$v = a_0 + \frac{1}{\sqrt{N+1}}\Phi + a_6 N \, , \qquad (2.1)$$

$$\Phi = a_1 \sum\nolimits_{i=1}^{N} \frac{qtf_i}{ql + a_2} + a_3 \sum\nolimits_{i=1}^{N} \log \frac{dtf_i}{dl + a_4} + a_5 \sum\nolimits_{i=1}^{N} \log \frac{ctf_i}{cl}, \quad (2.2)$$

where v is a relevance probability of a given document, N is the distinct number of terms appearing in both of the query and the document, qtf_i is frequency of term t_i within the query, ql is the query length, dtf_i is frequency of term t_i within the document, dl is the document length, ctf_i is frequency of term t_i within the whole collection, cl is total frequency of all terms within the collection, and $a_j \, (j = 0, \ldots, 6)$ are constants.

Unfortunately, it is difficult to estimate reasonably the values of $a_j \quad (j = 0, \ldots, 6)$ because the Berkeley's formula is a non-linear function on parameters. This may prevent us to adjust the formula to work optimally for various kinds of document database. A method of estimating efficiently the parameters is needed for enabling to adapt the logistic regression model widely into various situations.

A straightforward approach for solving this problem is to convert the non-linear formula (2.1) into a linear one by removing some constants and variables at the cost of its power for inferring relevance probability. If $a_0 = a_2 = a_4 = 0$ and $\left(\sqrt{N} + 1\right)^{-1}$ is deleted, we obtain a simplified formula,

$$v = b_1 \sum\nolimits_{i=1}^{N} \frac{qtf_i}{ql} + b_2 \sum\nolimits_{i=1}^{N} \log \frac{dtf_i}{dl} + b_3 \sum\nolimits_{i=1}^{N} \log \frac{ctf_i}{cl} + b_4 N . \quad (2.3)$$

If, for each document, we set that

$$x_1 = \sum\nolimits_{i=1}^{N} (qtf_i / ql), \quad x_2 = \sum\nolimits_{i=1}^{N} \log(dtf_i / dl), \quad x_3 = \sum\nolimits_{i=1}^{N} \log(ctf_i / cl) \text{ and } x_4 = N,$$

(2.3) reduces to a simple linear form,

$$v = b_1 x_1 + b_2 x_2 + b_3 x_3 + b_4 x_4 . \quad (2.4)$$

In order to give a theoretical basis to (2.3) or (2.4), we can refer to a framework of "combining evidence", which would be a fundamental assumption underlying various regression models for IR. In this framework, each of x_1, x_2, x_3 and x_4 in (2.4) is to be considered as a piece of evidence for relevance of each document given a query. We can find more theoretically sophisticated explanation about "combining evidence" in a recent monograph[4]. In our approach at NTCIR-2 ad hoc task, these pieces of evidence $x_j \, (j = 1, \ldots, 4)$ are simply added with each weight of $b_j \quad (j = 1, \ldots, 4)$. This means that the formula (2.4) accumulates each degree of evidence linearly, and the degree of relevance is to be inferred from the amount of weighted summation.

The advantage of this approach is its flexibility, i.e., various factors can be incorporated into the formula, and it is possible to combine them in very flexible manner. Inevitably, we have to continue theoretically and empirically exploring variables to be included in the formula and the way of combining them. From this perspective, our experiment reported in this paper would be a starting point for obtaining a final optimum formula.

2.2 Parameter estimation

The NTCIR-1, which consists of about 330,000 document records, query topics and relevance judgment information on each query, is available as a training data set for estimating parameters in (2.3) or (2.4). We can immediately calculate the value of x_j $(j = 1, \ldots, 4)$ in (2.4) if a query and a document are given. However, the value of independent variable v is unable to be known because the relevance judgment is dichotomous, i.e., relevant or non-relevant. Then, we are enforced to replace the continuous variable v with a categorical variable at the stage of training.

With this replacement, we need to apply a technique of discriminant analysis to the parameter estimation instead of classical least square method. Let D_n be a document collection, which consists of n documents, for training on a particular query. Also, $\mathbf{x}_j \, (j = 1, \ldots, 4)$ are defined as n-dimensional column vectors of which elements are the values of x_j in each document within D_n. Using a matrix notation, we can write

$$\mathbf{v} = \mathbf{X}\mathbf{b}, \quad (2.5)$$

where $\mathbf{X}$ is a $n \times 4$ matrix of which j-th column is $\mathbf{x}_j$, i.e., $\mathbf{X} = (\mathbf{x_1}, \mathbf{x_2}, \mathbf{x_3}, \mathbf{x_4})$, $\mathbf{b}$ is a column vector of coefficients in (2.4), i.e., $\mathbf{b} = (b_1, b_2, b_3, b_4)^T$, and $\mathbf{v}$ is a n-dimensional column vector of which element v_k represents a numerical value of k-th document in D_n $(k = 1, \ldots, n)$. It should be noted that we can partition the rows of $\mathbf{X}$ into two groups based on relevancy of the corresponding document (i.e., relevant group and non-relevant group) if information on the relevance judgment is available.

A technique of discriminant analysis with the classical Fisher's criterion can be employed for calculating an estimate of the coefficient vector $\mathbf{b}$. The Fisher's criterion means that we are to look for the linear function $\mathbf{Xb}$ which maximizes the ratio of the between-groups sum of square and the within-groups sum of square, given a $\mathbf{X}$ [5]. This paper makes use of a unique special algorithm based on singular value decomposition (SVD) [6] for calculating $\mathbf{b}$ according to an approximation of Fisher's criterion (see Appendix for details). As a result, we can obtain a set of $\mathbf{b}_q$ ($q = 1,\ldots,L$), where $\mathbf{b}_q$ is an estimate of vector $\mathbf{b}$ by the algorithm given the q-th query, and L is the number of queries in the training data set.

Finally, we calculate an estimate $\hat{\mathbf{b}}$ to be used through the experiment by a very simple formula,

$$\hat{\mathbf{b}} = L^{-1}\sum_{q=1}^{L}\mathbf{b}_q \text{ .} \qquad (2.6)$$

3 Query Expansion by Automatically Constructed Global Thesaurus

Query expansion technique would be very useful for searching by short queries. In general, it is difficult for actual users to translate their information needs into a set of appropriate search terms. Inevitably, the query tends to be short, i.e., it contains only a few words, and in most of the cases, the resulting performance of the search would be poor. One of the solutions is to expand the original query, i.e., to add automatically or semi-automatically some new search terms to the original query by using a thesaurus or other devices.

Many researchers have already proposed various methods for query expansion based on their own research assumptions or conditions. In this paper, we try to construct automatically a global thesaurus from document records in the whole set of the collection, and to add new terms automatically according to a set of heuristic rules. This means that our approach does not requires any external thesauri nor any human efforts for deciding added terms.

A classical approach to automatic thesaurus construction is to identify associations among terms appearing in the title or abstract of each document by analyzing frequencies of co-occurrence (see Doyle[7], Lesk[8] and so on). For example, the degree of association between term t_j and t_k can be measured by Dice coefficient,

$$r_{jk} = 2n_{jk} \Big/ \left(n_j + n_k\right), \qquad (3.1)$$

where n_j and n_k are the numbers of documents including term t_j and t_k respectively, and n_{jk} is the number of documents including both of term t_j and t_k. With this statistical information, we can detect automatically term-term relationships in a large corpus. The relationships allow us to select new search terms that are closely related to each original search terms.

Unfortunately, it is well known that query expansion using the global thesaurus constructed from simple co-occurrence information, often shows poor search performance[9]. Then, more sophisticated and complicated approaches have already been proposed (see [10]). However, it may be worthwhile confirming empirically performance of the classical co-occurrence approach in the context of the NTCIR experiment before attempting to develop an alternative method.

It should be noted that, in this experiment, we confine added terms to ones extracted from the field of author-assigned keyword in each record in order to enhance retrieval effectiveness. Unlike descriptors, the author keywords are not controlled, but these are able to be considered as 'pseudo-descriptors' because each author would attempt to assign the keyword for representing appropriately topics discussed in his/her article. Then, we can assume that the author keywords are useful for discerning topics of documents rather than terms used freely in the abstract.

Similarly, terms in the title are also likely to indicate more exactly topics of each document. Under these considerations, the terms included in our thesaurus were limited to ones in the titles and author keywords.

Our procedure for query expansion is as follows.

(1) Obtaining statistical information: co-occurrence frequencies of each pair of author keywords and terms in the title are computed using the test collection NTCIR-2, and the degree of each relationship is calculated by Dice coefficient (3.1).

(2) Constructing a thesaurus: terms in the title are organized as entries in a dictionary, which holds information on Dice coefficients of author keywords against each entry.

(3) Expanding the original query: for a term in the original text of each query, author keywords of which coefficient is over a threshold value are added to the query.

4 Experiment Using Japanese Collection of NTCIR-2

4.1 Indexing of Japanese text

A simple dictionary-based method, a kind of longest matching, was used for segmentation of Japanese text in this experiment. We use a machine-readable dictionary of ChaSen[11], which is a famous morphological analyzer for Japanese text. We extracted a string of characters from text of each document and query, which matches with longest entry in the machine-readable dictionary, and considered the string as an indexing term or a search term.

If a portion of the text remains to be not matched with any entries, it was tokenized into sub-strings which consist of the same kind of character. For example, we assume that the target text is '情報検索システムの研究' (a study of information retrieval system), and the dictionary has only an entry '情報' (information). First of all, '情報' is extracted from the text by a matching operation. Next, '検索' (retrieval) is extracted as a sequence of *Kanji* characters. Similarly, 'システム' (system) is extracted as a sequence of *Katakana* characters, 'の' (of) as a *Hiragana* character, and '研究' (a study) as a sequence of *Kanji* characters. Since words consisting of only *Hiragana* characters are not considered to be content-bearing words in most of the cases, we used a heuristic rule that the *Hiragana* sequence is always removed. As a result, four terms, '情報,' '検索,' 'システム,' and '研究' are automatically identified from the above text.

Furthermore, we tried to combine automatically the terms identified by the above algorithm into compound terms. The heuristic rule was that two adjacent terms in the text are combined into a compound term unless numerical or functional characters (including *Hiragana*) are placed between the two terms. According to the rule, '情報検索' (information retrieval) and '検索システム' (retrieval system) are added as identified terms in the case of above example.

4.2 Results in experiment

An inverted file of indexing terms identified automatically from the Japanese collection of NTCIR-2, was created. The indexing terms were extracted from texts within three fields in each record, i.e., title, abstract and author keyword fields. It should be noted that the terms appearing in over 70,000 documents (about 10% of the whole database) were removed from the inverted file.

The collection includes 736,166 records, and total frequency of all indexing terms in the collection amounts to 129,008,980 and the average document length is about 175.2 terms. This means that $cl = 129{,}008{,}980$ in our model (2.3).

In order to estimate regression coefficients in (2.3), we need a sample including relevant and non-relevant documents for each query in the training data set. The samples were obtained by the original Berkeley's logistic regression model (2.1), in which the parameters at the NTCIR-1 experiment were used, i.e., $a_0 = -3.51$, $a_1 = 37.4$, $a_2 = 35$, $a_3 = 0.33$, $a_4 = 80$, $a_5 = -0.1937$ and $a_6 = 0.0929$. On the other parameter in (2.1) depending the indexing method, we set as $cl = 38{,}648{,}949$.

For each query from 031 to 083 in the training data set NTCIR-1, the top ranked 200 documents were used as a sample, and 52 values of each $\mathbf{b}_q$ were obtained by the method described above (one query was excluded due to a computational problem). Finally, using (2.6), we obtained that $\hat{b}_1 = 12.8$, $\hat{b}_2 = 0.021$, $\hat{b}_3 = -0.078$ and $\hat{b}_4 = -0.33$.

We tried three retrieval runs in this experiment as indicated Table 1. In all runs, the simplified logistic regression model (2.3) was used. The SRGDU1 and SRGDU3 are runs for short queries, and SRGDU2 is a run for long queries.

Table 1. Monolingual Retrieval Runs for Japanese Collection

ID	Query Fields Used by Run	Query Expansion
SRGDU1	Description	no
SRGDU2	Title, Description, Narrative, Concept	no
SRGDU3	Description	yes

In SRGDU3, our query expansion method was applied according to some heuristic rules, i.e.,

(a) Terms of which Dice coefficient (3.1) is over 0.1, are added as new ones to the set of original query terms.

(b) If a new term has relationships with more than one original term, the maximum of Dice coefficients is adopted.

(c) If there are more than 20 new terms of which Dice coefficient is over 0.1 for each original term, only top ranked 20 terms are added.

The automatically constructed thesaurus by using documents in the NTCIR-2 test collection, has 749,238 entries and holds information on Dice coefficients between total 46,818,808 pairs of terms in the title field and terms in the author keyword field.

The NTCIR-2 test collection includes 49 query topics (from 101 to 149). According to the above rules, 18.47 terms were added on average per query (the maximum is 55 and the minimum is 0). Each weight qtf_i of the new terms was calculated by multiplying the value of Dice coefficient with qtf_i of original term.

The standard recall-precision curve with interpolation is shown as Figure1 and Table 2. The SRGDU2 for long query without query expansion achieves best retrieval effectiveness among the three runs. It is reasonable that run for long query (SRGDU2) outperforms one for short query (SRGDU1). The mean average precision of SRGDU2 is 0.315.

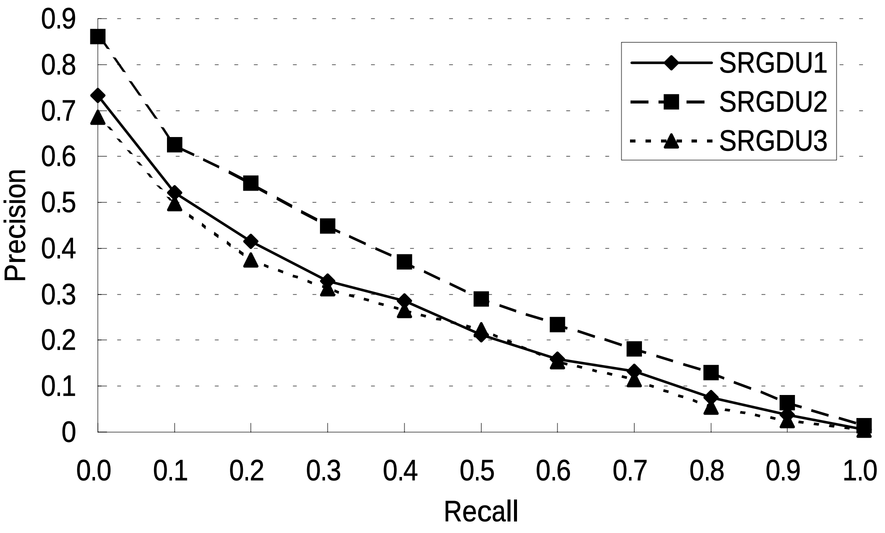

Figure 1 Recall-Precision Curve (interpolated)

Query expansion was unable to improve retrieval effectiveness, i.e., Figure 1 indicates that SRGDU3 for short query with expansion deteriorates retrieval performance. The mean average precision of SRGDU3 is 0.2264 in comparison with 0.2397 of SRGDU1 (see Table 2).

Table 2. Recall-Precision Curve

	SRGDU1	SRGDU2	SRGDU3
0.0	.7328	.8604	.6858
0.1	.5217	.6249	.4967
0.2	.4147	.5426	.3752
0.3	.3289	.4482	.3127
0.4	.2855	.3708	.2642
0.5	.2119	.2904	.2226
0.6	.1584	.2345	.1536
0.7	.1321	.1814	.1144
0.8	.0757	.1296	.0542
0.9	.0374	.0646	.0247
1.0	.0060	.0143	.0044
M.A.P.	.2397	.3150	.2264

5 Experiment Using English Collection of NTCIR-2

We also attempt a monolingual retrieval experiment for NTCIR-2 English test collection using our simplified logistic regression model. However, this experiment is only tentative because no training data set for this task is available, i.e., we can not conduct any parameter estimations for the English collection. Then, it is obliged to use the same parameters with those for the Japanese collection.

A simple stopword list and a well-known stemming algorithm, Porter's method[12], were used for indexing three fields of each record in the collection, i.e., title, abstract and author keyword, with no effort for identifying compound terms. Similarly, the search terms in queries written by English were identified automatically.

The collection contains 322,058 records, and total frequency of all word was 33,033,670 and the average length of documents was 102.57. We tried two runs for the English collection, i.e., SRGDU4 for long query (using Title, Description, Narrative, Concept) and SRGDU5 for short query (using Title and Description). The resulting mean average precision was 0.2569 for SRGDU4 and 0.1658 for SRGDU5, which indicate clearly poor performance.

6 Concluding Remarks

This paper discussed methods for ad hoc monolingual retrieval task at NTCIR-2 retrieval experiment. A simplified logistic regression model was proposed for enabling us to estimate easily the regression coefficients, and a method for the parameter estimation in such situation that only dichotomous relevance judgment is available, is developed by applying a kind of discriminant analysis.

Also, a query expansion technique using automatically constructed global thesaurus were discussed to improve performance for short queries. The thesaurus holds values of Dice coefficients for each pair of terms in the title field and terms in the author keyword field of each document within the test collection. Using the thesaurus, some new terms derived from the author keyword field were added to the original query according to some heuristic rules.

The experiment indicates that we need further improvement of our methods in many points. In particular, our query expansion method had no effect, rather degraded retrieval performance slightly. We may have to substantially improve our method, e.g., to introduce a local analysis of a set of documents instead of global thesaurus.

Similarly, for the simplified regression model and the parameter estimation, further investigations are clearly needed. For example, a very important issue is to evaluate the effect of discarding the power of the nonlinear function (2.1) for inferring relevance probability.

Acknowledgment

The author appreciates Dr. F. Gey and Dr. A. Chen at the UC Berkeley for teaching the Berkeley's formula of logistic regression when the author was a visiting scholar at the UC Berkeley.

Reference

[1] Fuhr, N.; Buckley, C. A probabilistic learning approach for document indexing. *ACM Transactions on Information Systems*, 9(3), 223-248 (1991).

[2] Cooper, W. S.; Chen, A.; Gey, F. Full text retrieval based on probabilistic equations with coefficients fitted by logistic regression. *The Second Text REtrieval Conference (TREC-2)*, D. K. Harman ed. National Institute of Standards and Technology, 1994. p.57-66.

[3] Chen, A.; Gey, F.C.; Kishida, K.; Jiang, H.; Liang, Q. Comparing multiple methods for Japanese and Japanese-English text retrieval. *Proceedings of the First NTCIR Workshop on Research in Japanese Text retrieval and Term Recognition*. 1999. p.49-58.

[4] Croft, W. B. ed. *Advances in Information Retrieval: Recent Research from the Center for Intelligent Information Retrieval*. Kluwer Academic Publishes, 2000.

[5] Mardia, K. V.; Kent, J. T.; Bibby, J. M. *Multivariate Analysis*. London, Academic Press, 1979.

[6] Press, W. H.; Teukolsky, S. A.; Vetterling, W. T.; Flannery, B. P. *Numerical Recipes in C*. 2nd ed. New York, Cambridge Univ. Press, 1992

[7] Doyle, L. Indexing and abstracting by association. *American Documentation*. Vol.13, No.4, p.379-390 (1962).

[8] Lesk, M. E. Word-word associations in document retrieval system. *Journal of Documentation*, 20(1), p.8-36 (1969).

[9] Peat, H. J.; Willett, P. The limitation of term co-occurrence data for query expansion in document retrieval system. *Journal of the American Society for Information Science*, 42(5), p.378-383 (1991).

[10] Baeza-Yates, R.; Ribeiro-Neto, B. *Modern Information Retrieval*. Harlow, England, Addison-Wesley, 1999. 513p.

[11] Matsumoto, Y.; Kitauchi, A.; Yamashita, T.; Hirano, Y.; Imaichi, O; Imamura T. Japanese *Morphological Analysis System ChaSen Manual*, NAIST Technical Report, 1997.
http://cactus.aist-nara.ac.jp/lab/nlt/chasen.html

[12] Porter, M. F. An algorithm for suffix stripping. *Program*, 14(3), p.130-137 (1980).

Appendix: *Algorithm for discriminant analysis with an approximation of Fisher's criterion*

The objective is to find out a m-dimensional vector $\mathbf{b}$ that maximize the ratio of the between-groups sum of square and the within-groups sum of square of $\mathbf{Xb}$, where $\mathbf{X}$ is a $n \times m$ matrix. It should be noted that $\mathbf{X}$ is a data matrix on m variables in a sample consisting of n documents ($d_1,\ldots d_n$), which is partitioned into two groups, i.e., the relevant group and the non-relevant one. Let l be the number of the relevant documents.

First, $\mathbf{X}$ can be decomposed such as $\mathbf{X} = \mathbf{U}\Lambda\mathbf{V}^T$, where $\mathbf{U}$ is a $n \times m$ orthogonal matrix, $\mathbf{V}$ is a $m \times m$ orthogonal matrix, and Λ is a $m \times m$ diagonal matrix. This is known as singular value decomposition or SVD. We can obtain a useful formula

$$\mathbf{U} = \mathbf{XV}\Lambda^{-1} \tag{A.1}$$

from SVD.

Second, we define as $\mathbf{U} = (\mathbf{u}_1,\ldots\mathbf{u}_n)^T$ where $\mathbf{u}_i$ is a m -dimensional column vector, and

$$\mathbf{m}_R \equiv l^{-1} \sum_{i:d_i \text{ is relevant}} \mathbf{u}_i \ ,$$

$$\mathbf{m}_N \equiv (n-l)^{-1} \sum_{i:d_i \text{ is non-relevant}} \mathbf{u}_i \text{ , and}$$

$$\tilde{\mathbf{b}} = \mathbf{m}_R - \mathbf{m}_N \, .$$

Then, it is easy to show that

$$\mathbf{u}_i^T\tilde{\mathbf{b}} = l^{-1} \qquad \text{if } d_i \text{ is relevant and}$$

$$\mathbf{u}_i^T\tilde{\mathbf{b}} = (n-l)^{-1} \qquad \text{if } d_i \text{ is non-relevant,}$$

using $\mathbf{u}_i^T \mathbf{u}_j = 1$ if $i = j$, and $\mathbf{u}_i^T \mathbf{u}_j = 0$ if $i \neq j$. As a result, $\mathbf{U}\tilde{\mathbf{b}}$ proves to be satisfying the Fisher's criterion if $n \neq 2l$, because the values of all relevant documents and all non-relevant documents are equal respectively, and the two kind of values are different (l^{-1} and $(n-l)^{-1}$). This means that the within-groups sum of square is 0 and the between-groups sum of square is a constant.

From (A.1), $\mathbf{U}\tilde{\mathbf{b}} = \mathbf{X}\mathbf{V}\Lambda^{-1}\tilde{\mathbf{b}}$. If we set

$$\mathbf{b}_q = \mathbf{V}\Lambda^{-1}\tilde{\mathbf{b}}, \tag{A.2}$$

we finally obtain $\mathbf{U}\tilde{\mathbf{b}} = \mathbf{X}\mathbf{b}_q$. Since the $\mathbf{X}\mathbf{b}_q$ satisfies a kind of Fisher's criterion, we can use (A.2) as a estimate of $\mathbf{b}$ given a query and the relevance judgment.