Seventh International Conference on Computer and Information Technology

Statistical Evaluation of Measure and Distance on Document Classification Problems in Text Mining

Masayuki Goto
Faculty of Environmental and Information Studies
Musashi Institute of Technology
Tsuzuki-ku, Yokohama, 224-0015, JAPAN
Email: goto@yc.musashi-tech.ac.jp

Takashi Ishida Shigeichi Hirasawa
School of Creative Science and Engineering
Waseda University
Shinjyuku-ku, Tokyo, 169-8555 JAPAN
Email: ishida@hirasa.mgmt.waseda.ac.jp

Abstract

This paper discusses the document classification problems in text mining from the viewpoint of asymptotic statistical analysis. By formulation of statistical hypotheses test which is specified as a problem of text mining, some interesting properties can be visualized. In the problem of text mining, the several heuristics are applied to practical analysis because of its experimental effectiveness in many case studies. The theoretical explanation about the performance of text mining techniques is required and this approach will give us very clear idea. The distance measure in word vector space is used to classify the documents. In this paper, the performance of distance measure is also analized from the new viewpoint of asymptotic analysis.

1 Introduction

Recently, text mining techniques have been more important by development of information technology to realize saving the huge amount of digital data [1],[2]. In this paper, the text classification problems [3] of document data are focused as an important application of text mining. For text classification, many models and algorithms have been proposed. One of them is a technique using a vector space model. In these methods, digital documents are automatically divided into terms by morphological analysis at first. A term is sometimes called a word. Next, the meaningful and important words are selected by using the mutual information or other metrics. The vector space is constructed by regarding the number of words as the dimension of Euclidean space. A document is represented as a point in the vector space by counting the number of each word appearing in it.

Basically, documents are characterized by the frequency of each word in the word set which is given by the morphological analysis. However, it is useful to introduce word selection step or measures based on term weighting schemes in practice. Without an appropriate word selection, the performance of information retrieval and document classification is deteriorated by unimportant words [4],[5]. Measure for term weighting is used for measuring the relevance of a word (term). A term weighting formula that provides appropriate weights can be introduced instead of an appropriate step of term selection. Many term weighting schemes have been proposed in the information retrieval field and its effevtiveness has been clarified by experiments. For example, "Term frequency-inverse document frequency"(TF-IDF) is one of the most commonly used term weighting schemes [6].

That is, it is important to remove the influence of unnecessary words in the text mining field. In other words, it is not easy to remove the unnecessary words completely. Therefore, we evaluate the performance in the document classification in the case where the unnecessary words are including in the word set. The existence of the unnecessary words is one of the characteristics of the text mining problems. The evaluation under this setting is not only interesting but also useful. From the results of formulation by hypothesis testing, the importance to exclude and reduce the influence of the unnecessary words is shown. Moreover, the performance of distance measure between documents in a large dimensional word vector space is analyzed. In the problems of text mining, the dimension of the word vector space is usually huge in comparison of the number of words appearing in a document. That is, it is essential to consider the problem of *sparseness* in the text mining. Although the frequencies of words appearing in a document could be small in many cases, many kinds of such word with small frequency can usually be used to classify the documents. Then, we analyze the performance of distance measure under the condition that the number of words diverges to infinity even if the frequencies of the words are still small. This

0-7695-2983-6/07 $25.00 © 2007 IEEE
DOI 10.1109/CIT.2007.171

IEEE computer society

asymptotic evaluation is different from that of usual statistics and information theory. Usual statistics focus on the convergence and the performance of statistics and estimators with a parametric model when the number of samples diverge to infinity. On the other hand, our asymptotic evaluation focus on the combination of huge dimensions and small sample in each dimension. We show in this paper that the estimator of distance measure calculated by the frequencies of words converges to the true distance valuse. From the asymptotic results about the distance measure, we can give the explanation of the fact given in many experiments that the classification by using the empirical distance between documents of the cosine measure is not so bad. It is also suggested that the KL-divergence is not useful for text mining problems.

2 Document Model

In this section, a vector space model is defined for our analysis. For construction of the vector space model for text classification, morphological analysis can be used to separate document data into words. By selecting important key words and counting the number of each word in a document, the document can be expressed as a point in the vector space using the frequencies of words. Then the similarity between documents can be calculated by a distance between the points expressing the documents.

2.1 A Vector Space and Document-word Matrix

Let the set of documents be $\Delta = \{d_1, d_2, \cdots, d_D\}$. All documents in Δ are separated into morphologies by morphological analysis and important words are selected from all of the morphologies.

Let the word set which is constructed by selecting the important words from all documents in Δ be $\Sigma = \{w_1, w_2, \cdots, w_W\}$. Then each document can be expressed by a W-dimensional vector as follows:

$$d_i = (v_{i1}, v_{i2}, \cdots, v_{iW})^T. \qquad (1)$$

That is, the document space can be constructed by regarding each component of vectors as the information about frequency of each word. Here, T is meaning transposition of a vector.

The matrix

$$A = (d_1, d_2, \cdots, d_D)^T, \qquad (2)$$

is called document word matrix. The number of words is usually very fuge. For example, it may be 3000 words and more. There is a possibility where some unefficient words are included in the word set because the word set is usually constructed by automatic calculation using some algorithm.

2.2 A Document Vectors

After the construction of the vector space to represent the documents d_i as the point in the space, it is very important how to determine the each value of the component of the word vector. The easiest way is to let the values be symple frequencies of words. Let f_{ij} be the number of the word w_j appearing in the document d_i. Then the document vector is given by $v_{ij} = f_{ij}$, that is

$$d_i = (f_{i1}, f_{i2}, \cdots, f_{iW})^T. \qquad (3)$$

However, this simplest way is sometimes not effective for the purposes of information retrieval and document classification.

Let f_{w_i} be a frequency of w_i in all documents Δ, F be a total frequency which is the summation of frequencies of all words in Σ. That is, F is given by

$$F = \sum_{w_i} \sum_{d_j} f_{ij} = \sum_{d_j} f_{d_j} = \sum_{w_i} f_{w_i}. \qquad (4)$$

Though we can use the frequencies $v_{ij} = \frac{f_{ij}}{F}$ and $v_{ij} = \frac{f_{ij}}{f_{d_j}}$, it may not be effective for practical problems. Though TF-IDF measure [7] is an effective way in information ritrieval, we consider the above measure to analyze the essential performance of the text classification problems.

2.3 Similarity between documents

By representation of document vector, a similarity between documents d_i and d_k can be calculated by using a distance between these vectors. Though it is possible to use the Euclid distance and the inner product, these method are usually not effective to classify the documents. This is because two documents near from origin may not similar each other but the distance is small.

One of usual way is to let the distance be a cosine between document vectors d_i and d_k.

$$sim(d_i, d_k) = \frac{d_i^T d_k}{|d_i||d_k|}. \qquad (5)$$

In the information retrieval problems, retrieval key words are expressed by a query vector and the documents which have high similarity with the query vector are listed up as a result. The similarity gives a way of ranking of documents. Another way to measure similarity is to use the probability measure based on Bayesian theory. Though the naive Bayes models are simple formulation, its performance is well. In this paper, we discuss about the performance of distance measures as basic study. The analysis of probability measure will be future work.

3 Consideration of document classification by statistical hypothesis test

In this paper, the asymptotic performance is investigated by assuming the simplest probabilistic model in order to clarify the basic characteristics of document classification problem.

We assume the case where there are two classes C_1 and C_2 and the documents are emitted from either C_1 or C_2. As a general case, the appearance probabilities of documents and words from C_1 and C_2 are different each other. Moreover, the document is independent of words in the meaning of probability theory. Let p_j^t be an appearance probability of j-th word w_j from the class C_t ($t \in \{1, 2\}$).

Without loss of generality, we can assume $p_j^1 > p_j^2$ for $j = 1, 2, \cdots, p$, $p_j^1 < p_j^2$ for $j = p+1, p+2, \cdots, p+q$, and $p_j^1 = p_j^2$ for $j = p+q+1, p+q+2, \cdots, W$. That is, the initial p words have high probability in the documents from class C_1, the next q words tends to occur in the documents from class C_2, and the last $W - p - q$ words have the same probability in these two classes and don't have any information to classify the documents.

Here, the KL-divergence $L(p^t; p^u)$ of $\mathbf{p}^u$ based on $\mathbf{p}^t$ is defined by

$$L(p^t; p^u) = \sum_{j=1}^{W} p_j^t \log \frac{p_j^t}{p_j^u}. \tag{6}$$

For the testing hypotheses, the Neyman-Pearson theorem specifies optimum decision regions. Let the two optimum decision regions be denoted by

$$\mathcal{U}_K = \left\{ \hat{q} : \sum_{j=1}^{W} \hat{p}_j \log \frac{p_j^1}{p_j^2} \geq K \right\}, \tag{7}$$

$$\mathcal{U}_K^C = \left\{ \hat{q} : \sum_{j=1}^{W} \hat{p}_j \log \frac{p_j^1}{p_j^2} < K \right\}. \tag{8}$$

If the empirical distribution of words in a document, $\hat{q}$, satisfies $\hat{q} \in \mathcal{U}_K$, then the document is classified to class C_1. If the empirical distribution, $\hat{q}$, satisfies $\hat{q} \in \mathcal{U}_K^C$, then the document is classified to class C_2. Let α be an error probability of the event (type I error) where the document from the class C_1 is mis-classified to the class C_2, β be an error probability of the event (type II error) where the document from the class C_2 is mis-classified to the class C_1.

To evaluate the optimum performance of the document classification, we define the probability models of these two classes where the the last $W - p - q$ words can be removed. Letting

$$S^t = \sum_{j=1}^{p+q} p_j^t,$$

and $\tilde{p}_j^t = p_j^t / S^t$, the KL divergence between the probability distributions without meaningless words for classification as follows:

$$\tilde{L}(p^t; p^u) = \sum_{j=1}^{p+q} \tilde{p}_j^t \log \frac{\tilde{p}_j^t}{\tilde{p}_j^u}. \tag{9}$$

Then, the following theorem is obtained as an analogy of the Stein's Lemma and the Sanov's theorem [8].

Theorem 1 *Let $\beta \in (0, 1)$ be given. Let α^* be the smallest probability of type I error over all decision rules such that the probability of type II error does not exceed β. Then, for all β in $(0, 1)$ and $f_d \to \infty$,*

$$(\alpha^*)^{1/f_d} \to \exp\left\{ -S\tilde{L}(p^t; p^u) \right\}, \tag{10}$$

where $S = S^1 = S^2$.

Theorem 2 *When the document is emitted from the class C_1, the inequation*

$$Pr\{\hat{q} \in \mathcal{U}_K\} \leq \left\{ -f_d S \tilde{L}(q^*; p^u) \right\}, \tag{11}$$

is satisfied, where q^* is given by

$$\tilde{L}(q^*; p^1) = \min_{\hat{q} \in \mathcal{U}_K} (\hat{q}; p^1). \tag{12}$$

(Proofs) The similar process in [8] leads to get the above theorems.

Let discuss the meaning of the above theorems. From the definition, $S < 1$ is satisfied and S is meaning the probability mass of meaningful words to classification. Usually, many words and terms are automatically extracted from the document set and a document-word matrix is constructed. Therefore, many meaningless words to classification may be included in the word set. Because the number of words may be over several thousand, the words which occur in different classes may be included. When the probability of meaningless words is $(1 - S)$, S appears in the exponent of the asymptotic error probability and the classification performance is deteriorated. If the meaningless words can be completely removed previously, we can expect the optimal performance in the meaning of the statistical hypotheses testing and $\tilde{L}(p^t; p^u)$ is the exponent of the asymptotic error probability.

The TF-IDF measure is a method to try giving the large weights and the small weights for the meaningful and the meaningless words, respectively. The reason of effectiveness of the TF-IDF measure can be suggested by the above results about the deterioration of error probability.

4 Consideration of similarity measure for document classification

In this section, we discuss the performance of the similarity measure given by Eq.(5). Usually, we can expect that the performance of divergence between the empirical distributions may be good. However, the divergence is not used so much in the field of information retrieval. This is because the similarity measure by using divergence is not effective for many cases of document classification. Then, we should clarify the mechanism of the cosine type measure in the document classification and information retrieval problems.

In order to evaluate easily, let consider the case of two classes again. Let p_j^t be the appearance probability of j-th word w_j in the document from the class C_t $(t \in \{1, 2\})$.

If the probability model is known, we can calculate

$$sim_d^*(p^t, p^u) = \sum_{j=1}^{W} p_j^t \log \frac{p_j^t}{p_j^u}, \tag{13}$$

and

$$sim_c^*(p^t, p^u) = \frac{\sum_{j=1}^{W} p_j^t p_j^u}{\sqrt{\sum_{j=1}^{W} \left(p_j^t\right)^2} \sqrt{\sum_{j=1}^{W} \left(p_j^u\right)^2}}. \tag{14}$$

However, it is realistic in the document classification problems to assume that only the distance between two empirical distributions representing the documents can be used.

Let consider the statistics $\hat{q}^t = \frac{f_{tj}}{f_{d_t}}$ and $\hat{q}^u = \frac{f_{uj}}{f_{d_u}}$. Here we assume that the numbers of words appearing in two documents are the same and is N.

$$sim_d(\hat{q}^t, \hat{q}^u) = \sum_{j=1}^{W} \hat{q}_j^t \log \frac{\hat{q}_j^t}{\hat{q}_j^u} \tag{15}$$

$$sim_c(\hat{q}^t, \hat{q}^u) = \frac{\sum_{j=1}^{W} \hat{q}_j^t \hat{q}_j^u}{\sqrt{\sum_{j=1}^{W} (\hat{q}_j^t)^2} \sqrt{\sum_{j=1}^{W} (\hat{q}_j^u)^2}}, \tag{16}$$

That is, the estimations of the distances given by the above equations are used for document classification instead of the true distances Eqs. (13) and (14).

Then, from the asymptotic properties of the maximum likelihood estimatiors, we have

$$sim_d(\hat{q}^t, \hat{q}^u) = sim_d^*(\hat{q}^t, \hat{q}^u) + O\left(\frac{1}{\sqrt{N}}\right), \quad a.s. \tag{17}$$

$$sim_c(\hat{q}^t, \hat{q}^u) = sim_c^*(\hat{q}^t, \hat{q}^u) + O\left(\frac{1}{\sqrt{N}}\right), \quad a.s. \tag{18}$$

when $N \longrightarrow \infty$. However, this type of asymptotic evaluation is not useful for the document classification problems. This is because the number of dimension of word vector is sometimes huge rather than the data number N. We cannot assume that the dimension of word vector is fixed and $N \longrightarrow \infty$ in the practical cases.

Then, we discuss about the performance of the estimations of the distances when the number of statistics (the number of words) is large. That is, it is important to evaluate the convergence of the estimated distances to the true distance when the number of words tends to ∞.

Let us assume the probabilistic independence of words and documents. p_j^t is the probability of j-th words in the document from the class C_t $(t \in \{1, 2\})$.

Moreover, let assume that $p_j^1 = r_1/p$, $p_j^2 = s_1/p$ for $j = 1, 2, \cdots, p$, $p_j^1 = r_2/q$, $p_j^2 = s_2/q$ for $j = p+1, p+2, \cdots, p+q$, and $p_j^1 = p_j^2 = r/(W-p-q)$ for $j = p+q+1, p+q+2, \cdots, W$. Here, $r = 1 - r_1 - r_2 = 1 - s_1 - s_2$. In this case, the true distance can be calculated as

$$sim_d^*(p^t, p^u) = r_1 \log \frac{r_1}{s_1} + r_2 \log \frac{r_2}{s_2}, \tag{19}$$

and

$$sim_c^*(p^t, p^u) = \frac{\frac{r_1 s_1}{p} + \frac{r_2 s_2}{q} + \frac{r^2}{W-p-q}}{\sqrt{\frac{r_1^2}{p} + \frac{r_2^2}{q} + \frac{r^2}{W-p-q}} \sqrt{\frac{s_1^2}{p} + \frac{s_2^2}{q} + \frac{r^2}{W-p-q}}}. \tag{20}$$

To express the situation in the document classification, we derive the asymptotics by the operation $W \longrightarrow \infty$ under the condition such that the ratio $p : q : n : W$ is fixed. Letting $R_1 = p/W$, $R_2 = q/W$, if the operation $W \longrightarrow \infty$ is considered under the condition that R_1 and R_2 are fixed, it is representing the sisuation with the high dimensional space of word vector and small data number of each word appering in the document.

The true distances are given by

$$sim_d^*(p^t, p^u) = r_1 \log \frac{r_1}{s_1} + r_2 \log \frac{r_2}{s_2}, \tag{21}$$

and

$$sim_c^*(p^t, p^u) = \frac{\frac{r_1 s_1}{R_1} + \frac{r_2 s_2}{R_2} + \frac{r^2}{1-R_1-R_2}}{\sqrt{\frac{r_1^2}{R_1} + \frac{r_2^2}{R_2} + \frac{r^2}{1-R_1-R_2}} \sqrt{\frac{s_1^2}{R_1} + \frac{s_2^2}{R_2} + \frac{r^2}{1-R_1-R_2}}}. \tag{22}$$

for all p, q, W satisfying $R_1 = p/W$ and $R_2 = q/W$ for fixed R_1 and R_2. Even if the frequency of each word is not large, the more the number of words is large the more the number of the frequency of all words is also large. Let consider the case the total number of the frequency of all words is proportional to W. Let the total number of the frequency of all words be $N = nW$ where n is a fixed positive integer. Then, the operation such that the number of dimension tends to ∞ gives the following convergence theorem.

Theorem 3 *Fixing n, $R_1 = p/W$ and $S_2 = q/W$, and letting $W \longrightarrow \infty$, the following convergence is satisfied:*

$$sim_c(\hat{q}^t, \hat{q}^u) \longrightarrow sim_c^*(p^t, p^u), \quad a.s. \tag{23}$$

However, $sim_d(\hat{q}^t, \hat{q}^u)$ doesn't converge to the true distance $sim_d^*(p^t, p^u)$.

Theorem 4 *Fixing n, $R_1 = p/W$ and $S_2 = q/W$, and letting $W \longrightarrow \infty$, the following convergence for $sim_d(\hat{q}^t, \hat{q}^u)$ is satisfied:*

$$sim_d(\hat{q}^t, \hat{q}^u) \longrightarrow R_1\mu_1 + R_2\mu_2 + (1 - R_1 - R_2)\mu_3, \quad a.s. \tag{24}$$

where μ_1, μ_2, and μ_3 are given by

$$\mu_1 = E\left[\frac{k_1}{n} \log \frac{k_1/n}{k_1'/n}\right], \tag{25}$$

$$\mu_2 = E\left[\frac{k_2}{n} \log \frac{k_2/n}{k_2'/n}\right], \tag{26}$$

and

$$\mu_3 = E\left[\frac{k_3}{n} \log \frac{k_3/n}{k_3'/n}\right]. \tag{27}$$

Here, k_1, k_1', k_2, k_2', k_3, k_3' are the random variables which means the frequencies of n Bernoulli trials with the following averages:

$$E\left[\frac{k_1}{n}\right] = \frac{r_1}{R_1}, \quad E\left[\frac{k_1'}{n}\right] = \frac{s_1}{R_1},$$

$$E\left[\frac{k_2}{n}\right] = \frac{r_2}{R_2}, \quad E\left[\frac{k_2'}{n}\right] = \frac{s_2}{R_2},$$

and

$$E\left[\frac{k_3}{n}\right] = E\left[\frac{k_3'}{n}\right] = \frac{r}{1 - R_1 - R_2}.$$

(Proofs) The outline of the proofs of Theorem 3 and 4 are in Appendix A and B.

In Theorem 4, if the equations

$$E\left[\frac{k_1}{n} \log \frac{k_1/n}{k_1'/n}\right] = \frac{r_1}{R_1} \log \frac{r_1}{s_1},$$

$$E\left[\frac{k_2}{n} \log \frac{k_2/n}{k_2'/n}\right] = \frac{r_2}{R_2} \log \frac{r_2}{s_2},$$

and

$$E\left[\frac{k_3}{n} \log \frac{k_3/n}{k_3'/n}\right] = \frac{r}{1 - R_1 - R_2} \log \frac{r}{r} = 0,$$

are satisfied, then $sim_d(\hat{q}^t, \hat{q}^u) \longrightarrow sim_d^*(p^t, p^u)$. However, the above equations are not generally satisfied.

Theorems 3 and 4 show the important fact that the cosine distance measure $sim_c(\hat{q}^t, \hat{q}^u)$ is superior to the KL-divergence measure $sim_d(\hat{q}^t, \hat{q}^u)$ under the condition which can be usually assumed to the text classification problems. This characteristic has been sometimes pointed out by using the results of simulation experiments, although the KL-divergence measure is essential and useful in the statistics and information theory.

Theorems 3 and 4 are derived by modeling the situation of sparseness. The effective measures and techniques for the problems with sparseness would be different from those for other usual problems. The results shown in this paper are explaining this fact from the theoretical viewpoint.

In the problem of document classification, there are two word qroups, i.e., a group of words appearing with high probability in the documents of a category and a group of other words not appearing in the category. For the statistical analysis and learning problems, it is important to evaluate the asymptotic performance when the number of dimension is fixed and number of data diverges to ∞. However, situation in the field of text mining is usually different. Though the number of words and the size of text data are large, the number of dimension of vector space model expressing the characteristics of documents is also huge. This leads to the sparseness problem.

The above asymptotic equations in Theorem 3 are the results under the condition representing the sparseness. Even if the number of data in each statistic is small, the true distance between documents can be estimated by collecting many statistics with small sample size and using the cosine measure. By this result, we can give the explanation of the fact given in many experiments that the classification by using the empirical distance between documents is not so bad.

5 Conclusion

This paper discussed the document classification problems in text mining from the viewpoint of statistical asymptotics. By formulation of statistical hypotheses testing which is specified as the problem of text mining, some interesting properties can be visualized. Moreover, we showed the asymptotic equations of distance measure under the condition of sparseness. From these results, the performance of document classification problems in text mining can be discussed from the theoretical viewpoints. In some settings, the cosine mesure is effective for text mining problems, rather than the KL-divergence measure.

Analysis for more general case and probability measure will be future work.

References

[1] M. Hearst, "Untangling text data mining," *ACL'99 Proceedings*, pp.3-10, (1999)

[2] K. Kita, *Probabilistic language models*, The university of Tokyo press, Tokyo, (1999)

[3] M. Suzuki, "Text classification based on the bias of word frequency over categories," *Proceedings of the International Conference on Artificial Intelligence and Its Applications (AIA)*, pp.400-405, (2006)

[4] M.Nagao, M.Mizutani, H.Iketa, "An Automated Method for the Extraction of Important Words from Japanese Scientific Documents," *Transactions of Information Processing Society of Japan*, Vol.17, No.2, pp.110-117, (1976)

[5] K. W. Church, P.Hanks, "Word Association Norms, Mutual Information and Lexicography," *Computational Linguistics*, Vol.6, No.1, pp.22-29, (1990)

[6] G. Salton, C. Buckley, "Weighting approaches in automatic text retrieval," Information Processing and Management, Vol.24, No.5, pp.513-523, (1988)

[7] A.Aizawa: "An Information-theoretic perspective of tf-idf measures," *Information Processing and Management*, Vol.39, pp.45-65, (2003)

[8] Richarde E. Blahut: *Principles and Practice of Information Theory*, Addison-Wesley Publishing Co., (1987)

Appendix A: The proof of Theorem 3

The numerator of $sim_c(\hat{q}^t, \hat{q}^u)$ is given by

$$\sum_{j=1}^{W} \hat{q}_j^t \hat{q}_j^u = \sum_{j=1}^{p} \hat{q}_j^t \hat{q}_j^u + \sum_{j=p+1}^{p+q} \hat{q}_j^t \hat{q}_j^u + \sum_{j=p+q+1}^{W} \hat{q}_j^t \hat{q}_j^u$$

For $j = 1, 2, \cdots, p$, because

$$E\left[\hat{q}_j^t\right] = \frac{r_1}{p}, \quad E\left[\hat{q}_j^u\right] = \frac{s_1}{p},$$

we have

$$E\left[W\hat{q}_j^t\right] = \frac{r_1}{R_1}, \quad E\left[W\hat{q}_j^u\right] = \frac{s_1}{R_1}.$$

Similarly, we have

$$E\left[W\hat{q}_j^t\right] = \frac{r_2}{R_2}, \quad E\left[W\hat{q}_j^u\right] = \frac{s_2}{R_2},$$

for $j = p+1, p+2, \cdots, p+q$ and

$$E\left[W\hat{q}_j^t\right] = E\left[W\hat{q}_j^u\right] = \frac{r}{1 - R_1 - R_2},$$

for $j = p+q, p+q+1, \cdots, W$. We have, therefore,

$$\frac{1}{p}\sum_{j=1}^{p} W^2 \hat{q}_j^t \hat{q}_j^u \to \frac{r_1 s_1}{R_1^2}, \quad a.s.$$

when $p \to \infty$. Because $p = WR_1$,

$$\frac{1}{W}\sum_{j=1}^{p} W^2 \hat{q}_j^t \hat{q}_j^u \to \frac{r_1 s_1}{R_1}, \quad a.s.$$

is satisfied when $W \to \infty$ for $j = 1, 2, \cdots, p$. From the similar discussion, we have

$$\frac{1}{W}\sum_{j=p+1}^{p+q} W^2 \hat{q}_j^t \hat{q}_j^u \to \frac{r_2 s_2}{R_2}, \quad a.s.$$

$$\frac{1}{W}\sum_{j=p+q+1}^{W} W^2 \hat{q}_j^t \hat{q}_j^u \to \frac{r^2}{1 - R_1 - R_2}, \quad a.s.$$

The convergence of the denominator of $sim_c(\hat{q}^t, \hat{q}^u)$ is also given as

$$\sqrt{W\sum_{j=1}^{W}(\hat{q}_j^t)^2}\sqrt{W\sum_{j=1}^{W}(\hat{q}_j^u)^2} \to$$

$$\sqrt{\frac{r_1^2}{p} + \frac{r_2^2}{q} + \frac{r^2}{W-p-q}}\sqrt{\frac{s_1^2}{p} + \frac{s_2^2}{q} + \frac{r^2}{W-p-q}}, \quad a.s.$$

by the similar discussion.

Therefore, we have $sim_c(\hat{q}^t, \hat{q}^u) \to sim_c^*(p^t, p^u)$, a.s.

Appendix B: The proof of Theorem 4

With a similar discussion with Appendix A, the convergence of $sim_c(\hat{q}^t, \hat{q}^u)$ can be proved.

$$sim_d(\hat{q}^t, \hat{q}^u) =$$

$$\sum_{j=1}^{p} \hat{q}_j^t \log\frac{\hat{q}_j^t}{\hat{q}_j^u} + \sum_{j=p+1}^{p+q} \hat{q}_j^t \log\frac{\hat{q}_j^t}{\hat{q}_j^u} + \sum_{j=p+q+1}^{W} \hat{q}_j^t \log\frac{\hat{q}_j^t}{\hat{q}_j^u}$$

For example, the term

$$\sum_{j=1}^{p} \hat{q}_j^t \log\frac{\hat{q}_j^t}{\hat{q}_j^u} = \frac{R_1}{p}\sum_{j=1}^{p} W\hat{q}_j^t \log\frac{W\hat{q}_j^t}{W\hat{q}_j^u}$$

does not converge to $\frac{r_1}{R_1}\log\frac{r_1}{s_1}$, but to $R_1\mu_1$ almost surely. This is because

$$E[g(X)] \neq g(E(X))$$

is not generally satisfied, where X is a random variable and $g(x)$ is a function.

By evaluation of all terms by the same discussion, we have the following convergence and the proof is complete:

$$sim_d(\hat{q}^t, \hat{q}^u) \to R_1\mu_1 + R_2\mu_2 + (1 - R_1 - R_2)\mu_3, \quad a.s.$$