Differences between the Quantitative Features of Articles from Biomedical Engineering and the Medical Sciences Aimed at Effective Retrieval of Literature

Motoki Sakai

Abstract – The field of biomedical engineering (BM) is concerned with solving medical problems in diverse ways, for example, by using information technology (information technology-based BM is the focus of this study.) BM research is linked to both information technology and medical (ME) research, and both BM and ME research papers contain common terms such as disease names. However, novice BM researchers have difficulty in searching for and retrieving BM articles because of this. This study attempted to overcome this difficulty by adopting a text mining-based method for classifying BM and ME articles to investigate the quantitative features of these articles. The classification of BM and ME articles was examined by collecting five types of articles containing the common terms obstructive sleep apnea syndrome (OSAS), T-wave alternans (TWA), late potential (LP), epilepsy (EPY), or event-related potential (ERP). Gathered BM and ME articles were quantitatively converted into document-term matrices (D-T matrices), following which the term-term matrices and the two feature values, the norm of the term-term matrix and the first principal component of the D-T matrix, were determined for the classification. The classification was conducted through the use of the fact that two of the feature values of the BM articles were smaller than those of ME. As a consequence, BM and ME articles could be distinguished with 96% averaged sensitivity and with 82% averaged specificity, thereby verifying the effectiveness of the proposed method.

Keywords – Biomedical Engineering, Medical, Classification, Text Mining.

I. INTRODUCTION

The raison d'être of biomedical engineering (BM) research is to address medical problems by employing techniques from information technology, mechanical engineering, or electrical and electronic engineering. Among these techniques related to BM research, this study focuses on the use of information technology (henceforth, the term “BM research” means “BM research related only to information technology.”).

For example, researchers in the BM field conduct their studies using information-processing techniques to automatize a medical diagnosis, or to extract information about an imperceptible phenomenon. Achieving these aims requires BM researchers to not only have knowledge of information technology, but also of medical (ME) research, because they cannot create appropriate solutions unless they understand what kinds of problems currently exist in the ME research field. In summary, the BM research field is positioned in a common research area between information technology and ME.

Usually, a BM researcher surveys related research articles in the early stage of the study, after which they examine the collected measurable articles, and then identify a problem that should be addressed. However, conducting this survey can pose problems for novice BM researchers such as undergraduate students.

As discussed, the BM research field is related to both information technology and ME research, and specific research areas of BM and ME often have common terms. For example, let us assume this common term is “arrhythmia.” One ME article may present a case report of patients suffering from severe arrhythmia. On the other hand, a BM article might propose a new signal processing technology to automatically detect an arrhythmia episode from the electrocardiogram (ECG) signal. In this way, a common term is used in different contexts in both BM and ME articles. This is problematic for novice BM researchers who need to study BM articles exclusively. For example, if a BM researcher queries “arrhythmia” on a search engine such as Google Scholar, all the articles appearing in the search results may not necessarily be BM articles, because both BM and ME articles could be included in the search result. Experienced BM researchers would easily be able to cull only the BM articles, but novice BM researchers, not being overly familiar with academic terms in information technology and ME research, would not necessarily be able to select them with ease, as it may take time to determine whether a retrieved journal paper is a BM article. Therefore, in this paper, a text mining-based supporting technology to automatically cull only BM articles from search results is proposed for novice BM researchers.

Text mining is a technique to quantitatively analyze documents, and is applied to solve issues in varied research fields. For instance, researchers ([11] and [2]) analyzed literary works quantitatively, and attempted to effect new solutions for old problems in the history of literature. In [3] and [4], researchers analyzed documents written on a micro-blog or Facebook, and estimated authors’ political opinion. In the field of business science, text-mining techniques have been developed for application in marketing [5] and [6].

In a previous study [7], an author suggested that it might be possible to distinguish BM and ME articles by using the features of term-term matrices generated from those of BM and ME. However, the quantitative difference between BM and ME articles has not yet been clarified or generalized sufficiently, because the previous study only examined articles with the common term “OSAS.” Therefore, we aim to evaluate the difference between the quantitative features of BM and ME articles with a greater variety of literature.

Copyright © 2015 IJECE, All right reserved
664
II. BM AND ME RESEARCH ARTICLES USED IN THIS STUDY

The quantitative features of BM and ME articles were investigated with the view of developing a method to discriminate between articles from these two fields of research. Five common terms were selected for this purpose: obstructive sleep apnea syndrome, T-wave alternans, late potential, epilepsy, and event-related potential. These terms are often found in both BM and ME research. In this research, 20 BM and 20 ME articles were collected for each of these common terms for evaluation. In this section, some examples of these articles are shown below each common term.

- **Obstructive sleep apnea syndrome**
  Obstructive sleep apnea syndrome (OSAS) is the most common type of sleep apnea and is caused by repetitive occlusions of the upper airways. OSAS causes hypertension, arrhythmia, cardiac arrest, diabetes, or dyslipidemia, which poses further risks of brain infarct or cardiac infarct. Therefore, many biomedical engineering and clinical approaches have been presented recently for the early detection and treatment of OSAS.

  A few of these studies, including [8], [9], [10], were selected as BM articles, which mainly present signal processing techniques to automatically detect sleep apnea episodes from long-duration electrocardiogram (ECG) recordings. On the other hand, papers [11], [12], [13] were selected as ME articles. These papers present case reports of OSAS patients or the mortality risk due to sleep apnea.

- **T-wave alternans**
  T-wave alternans (TWA) is one of the most promising predictors of sudden cardiac death, which is defined as a beat-to-beat change in the amplitude of the T-wave which is one of the component waves of ECG. TWA cannot always be observed with the naked eye, and many research efforts to detect TWA with signal processing techniques have been presented in the BM research field (e.g., [14], [15]). On the other hand, mortality after myocardial infarction [16] or the importance of a Multicenter Automatic Defibrillator Implantation Trial [17] is discussed in the literature pertaining to ME research.

- **Late potential**
  As with TWA, the late potential (LP) is also a promising predictor of sudden cardiac death. This is a high-frequency component that occasionally appears in the ECG signal recorded for post-myocardial infarction patients, and has extremely small amplitude. Therefore, LP cannot be detected without signal processing techniques, and various approaches have been proposed by BM researchers (e.g., [18], [19]). In ME research articles, for example, the relationship between LP and ventricular tachycardia, or atrial fibrillation, has been reported [20], [21].

- **Epilepsy**
  Epilepsy (EPY) is a chronic brain disorder, which causes repetitive seizures, and results from various factors. Generally, EPY is diagnosed by the electroencephalographic (EEG) signal. In BM research articles, various methods such as a wavelet transformation-based EPY detection method [22] and a bivariate analysis-based EPY prediction method using the EEG signal [23], have been proposed. By contrast, ME papers have reported a prognostic effect of EPY surgery [24], [25] and so on.

- **Event-related potential**
  Event-related potential (ERP) is an electrical response in the brain, which is evoked by an internal (emotional) or external stimulus. ERPs are often applied at the brain-computer interface as a technique to control electrical devices, such as interacting with a computer without hands or feet, because ERPs reflect human emotion and decision making. BM researchers have proposed an effective method to detect ERP in the EEG signal using methods such as wavelet transformation and support vector machines [26], [27]. Meanwhile, for example, medical professions have reported differences between the ERP of healthy people and those of patients with certain disorders such as Alzheimer’s disease [28], [29].

III. METHOD

We quantitatively discriminated between articles from the BM and ME fields of research by first transforming text documents into a numerical dataset, then selecting their effective features, and then classifying them using their features. The processing diagram is shown in Fig. 1.

A. Collection of articles related to “Common terms,” Creation of dictionary from BM and ME articles, Creation of document-term matrices

In this research, it is assumed that the 40 selected research papers that include one of the five common terms (e.g., OSAS, TWA, or ERP) have already been corrected without distinction between BM and ME.

Discrimination between BM and ME articles requires the quantification of text documents. First, a dictionary containing $N$ terms $t = [t_1, t_2, \ldots, t_N]$ is created from the collected BM and ME articles. Next, the document-term (D-T) matrix created from collected BM and ME articles [30] is adopted as shown in Eq. (1).

$$D = \begin{bmatrix} tf(t_1, d_1) & \cdots & tf(t_N, d_1) \\ \vdots & \ddots & \vdots \\ tf(t_1, d_j) & \cdots & tf(t_N, d_j) \end{bmatrix},$$

where $tf(t_i, d_j)$ is the frequency of term $t_i$ in document $d_j$. The number of columns $N$ corresponds to the total number of terms contained in a dictionary created from the collected BM and ME articles. In this research, $(40 \times N)$ D-T matrices are regarded as quantitative representations for BM and ME articles.

Generally, $tf(t_i, d_j)$ is normalized by the total number of terms contained in document $d_j$, because the number of terms differs from one document to another. In most situations, each term frequency $tf(t_i, d_j)$ is weighted because not all terms are important to feature in a document. The weighting function is defined as follows:


\[
\text{w}(t) = \ln \left( \frac{j}{df(t)} \right),
\]

where \( j \) is the number of documents, and \( df(t) \) is the number of documents containing the term \( t \). In this research, \( j = 40 \). Using the weighting function described by Eq. (2), the tf representation \( tf(t, d_k) \) shown in Eq. (1) is converted into a tf-idf representation \( tfidf(t, d_k) \) using Eq. (3) as follows:

\[
tfidf(t, d_k) = tf(t, d_k) \circ w(t), \quad t = \{t_1, t_2, \cdots, t_N\}.
\]

where \( \circ \) denotes an element-wise multiplication.

Practically, \( 40 \times (1 \times N) \) D-T matrix \( D_k \) is created for each document \( d_k \), \( k = \{1, 2, \cdots, 40\} \) in this research, and each element of each D-T matrix is normalized by Eq. (3); then, 40 tf-idf weighted D-T matrices \( (1 \times N) \) are generated. In the following step, each D-T matrix \( D_k \) is classified whether it represents a BM or ME article.

**B. Creation of term-term matrices**

Feature extraction from each document \( d_k \) was conducted by computing a term-term matrix \( T_k \) as follows:

\[
T_k = wD_k^T \cdot wD_k,
\]

where \( wD_k \) is a tf-idf weighted document-term matrix for the \( k \)-th document, and \( wD_k^T \) is the transposed matrix of \( wD_k \).

**C. Calculation of principal components of D-T matrices**

Generally, a document-term matrix is high dimensional. Therefore, dimension reduction is achieved by a decomposition algorithm such as singular value decomposition or principal component analysis (PCA). In this research, PCA is adopted to perform dimension reduction for the document-term matrix. The use of PCA enables \( N \) principal components to be obtained. In the following feature extraction step, only the effective principal components are selected as features for classification.

**D. Feature extraction**

In this research, features to classify BM and ME articles are extracted from the principal components of the D-T and term-term matrices. The first principal component of the D-T matrix is adopted as one of the features. In addition, the norm of the term-term matrix is calculated as a feature of the term-term matrix. The premise that the two features described above are selected to classify BM and ME articles is indicated in Section 6.

**E. Classification**

The rules of classification between BM and ME articles are empirically defined based on the resultant distribution of these two features, namely the first principal component of the D-T matrix and the norm of the term-term matrix. In this research, the number of classification rules is two (referred to as Rule1 and Rule2, respectively.) If either Rule1 or Rule2 is true, an article can be regarded as relating to BM. Rule1 states that an article can be classified as BM when the value of the first principal component of its D-T matrix is less than the 50-th percentile of the value of the 40 articles of the first principal component (20 BM + 20 ME articles). Rule2 states that an article can be classified as BM when the norm of its term-term matrix is less than a threshold value. The definition of the threshold value is as follows:

\[
\text{Threshold} = 0.4 \times \sum_{k=1}^{40} \| F_k \|, \quad \{k = 1, \cdots, 40\},
\]

where \( \| \cdot \| \) denotes the norm of the term-term matrix of the \( k \)-th document. The constant 0.4 shown in Eq. (5) was empirically determined.

These Classification Rules were determined under the assumption that the number of collected BM journal papers was equivalent to that of ME journal papers.

**IV. EVALUATION**

The classification precision between BM and ME research papers was examined by using the sensitivity and specificity. These terms are defined as follows:

\[
\text{Sensitivity} = \frac{TP}{TP + FN},
\]

\[
\text{Specificity} = \frac{TN}{TN + FP}.
\]

In Eqs. (6) and (7), TP, TN, FP, and FN represent the quantities of true positives, true negatives, false positives, and false negatives, respectively. In this study, TP indicates the number of correct classifications of BM articles, and TN denotes the number of correct classifications of ME articles.
V. Results

Figs. 2 through 6 show the visualized term-term matrices of BM and ME articles. Concretely, these figures illustrate the averaged term-term matrices of research papers including the common terms “OSAS,” “TWA,” “LP,” “EPY,” and “ERP,” respectively. Averaging was performed with 20 term-term matrices of BM articles (or 20 term-term matrices of ME articles) for every common term. The colors shown in Figs. 2 through 6 represent the magnitude of the elements, with brown and blue indicating larger and smaller magnitudes, respectively.

Fig. 2. Visualized features of articles related to OSAS: average term-term matrix of (a) BM article and (b) ME article

Fig. 3. Visualized features of articles related to TWA: average term-term matrix of (a) BM article and (b) ME article

The distribution of the principal components of the D-T matrices of BM and ME articles are visualized in Figs. 7, 8, 9, 10, and 11, which show the first and second principal components of the OSAS, TWA, LP, EPY, and ERP articles, respectively. In these figures, each “+” (or “o”) indicates the position of a BM (or ME) research paper in a two-dimensional plane described by the first and second principal components of the D-T matrices.

Fig. 4 Visualized features of articles related to LP: average term-term matrix of (a) BM article and (b) ME article

Fig. 5. Visualized features of articles related to EPY: average term-term matrix of (a) BM article and (b) ME article

Fig. 6. Visualized features of articles related to ERP: average term-term matrix of (a) BM article and (b) ME article
Fig. 7. Two-dimensional mapping with first and second principal components of tf–idf weighted document-term matrices of OSAS articles.

Fig. 8. Two-dimensional mapping with first and second principal components of tf–idf weighted D-T matrices of TWA articles.

Fig. 9. Two-dimensional mapping with first and second principal components of tf–idf weighted D-T matrices of LP articles.

Fig. 10. Two-dimensional mapping with first and second principal components of tf–idf weighted D-T matrices of EPY articles.

Fig. 11. Two-dimensional mapping with first and second principal components of tf–idf weighted D-T matrices of ERP articles.

Table 1: Results of classification

<table>
<thead>
<tr>
<th>Common Term</th>
<th>Sensitivity [%]</th>
<th>Specificity [%]</th>
</tr>
</thead>
<tbody>
<tr>
<td>OSAS</td>
<td>100</td>
<td>85</td>
</tr>
<tr>
<td>TWA</td>
<td>90</td>
<td>85</td>
</tr>
<tr>
<td>LP</td>
<td>95</td>
<td>75</td>
</tr>
<tr>
<td>EPY</td>
<td>95</td>
<td>70</td>
</tr>
<tr>
<td>ERP</td>
<td>100</td>
<td>95</td>
</tr>
<tr>
<td>Average</td>
<td>96</td>
<td>82</td>
</tr>
</tbody>
</table>

Table 1 lists the accuracy, sensitivity, and specificity of classification between BM and ME articles using the proposed algorithm.

VI. DISCUSSION

The goal of this research is to investigate and identify quantitative features of BM and ME research articles, and to automatically discriminate between those from BM and
those from ME. In this research, D-T matrices were firstly created from BM and ME articles containing a common term, such as “OSAS,” “TWA,” or “ERP,” and effective features for the classification between BM and ME articles were investigated by using D-T matrices (or term-term matrices created from D-T matrices). In particular, the first principal component of the D-T matrices and the norm of term-term matrices were extracted as features for classification.

As shown in Figs. 2 through 6, the term-term matrices of BM research papers were sparser than those of ME articles. In addition, in previous research [7], the differences between the features of the term-term matrices of BM and ME articles were confirmed using a small dataset. (Only research articles related to “OSAS” were examined.) In contrast, in this research, five research themes were selected to distinguish BM articles from ME articles. Our results confirmed that the sparseness of the term-term matrix of a BM article could be adapted as an effective feature to discriminate between BM and ME articles.

On the other hand, Figs. 7 through 11 show the effectiveness of using the first principal component of the D-T matrix. In these figures, the values of the first principal component of the BM articles are lower than those of the ME articles, which indicates that the value of the first principal component could be used as one of the features to achieve an efficient classification.

Rule1 and Rule2 were employed to perform the classification with the above-mentioned two features. As a result, the average sensitivity and specificity was determined to be 96% and 82%, respectively. This result indicates that most BM articles were correctly distinguished from ME articles. Although some of the themes were somewhat difficult to use for discrimination purposes (e.g., articles related to EPY), the two features and two rules were to some extent found to be effective in discriminating BM articles from ME articles for the five selected research themes. Therefore, this result implies that the proposed method is effective to classify general BM and ME articles.

Moreover, the fact that for ME articles the values of the norm of the term-term matrix and first principal component of the D-T matrix are larger than those of BM articles indicates that ME articles contain richer vocabularies as compared to BM articles. In fact, even if novice BM researchers do not have a pressing need to read ME articles, which contain many medical terms, they might well have to read those ME papers to determine whether they are BM articles. In this context, the proposed algorithm was shown to be effective to discriminate between BM and ME articles, because it can save novice BM researchers time and effort when they have to survey and examine the literature.

VII. CONCLUSION
The purpose of this study was to quantitatively discriminate between BM and ME articles. The proposed method was evaluated by collecting articles based on five research topics, which included the common terms abbreviated as “OSAS,” “TWA,” “LP,” “EPY,” or “ERP.” The collected BM and ME articles were quantified by using the D-T matrices and the term-term matrices created from the D-T matrices, and the norm calculated from the term-term matrix and first principal component of the D-T matrix were used as features for the classification. The classifications were then performed drawing on the quality that two of the feature values of BM articles were found to be smaller than those of ME articles. As a result, BM and ME articles could be discriminated with 96% averaged sensitivity and with 82% averaged specificity.

REFERENCES
Motoki Sakai

Motoki Sakai received his B.E., M.E., and Ph.D. degrees from the University of Aizu, Aizuwakamatsu, in 2003, 2005, and 2009. He joined QRS Corporation in 2009, and worked as an assistant professor at the Fukushima University Graduate School of Symbiotic Systems Science and Technology, and Tokyo Denki University, the School of Information Environment from 2010 to 2015. Currently, he works as an assistant professor at Chiba University of Commerce, Faculty of Policy Informatics. His current research interests include data science.