Text Mining Technique to Distinguish between Clinical Medicine and Biomedical Engineering Research Articles

Motoki Sakai

Abstract—Biomedical engineering (BM) is used to solve medical problems in diverse ways such as by using information technology. The BM research field is related to both information technology and clinical medicine (CL) research, and both BM and CL articles include common terms such as disease names. Young researchers have difficulty in effectively searching BM articles due to this. To solve this problem, this paper proposes a text mining-based method to distinguish between BM and CL articles. In this research, 20 BM and 20 CL articles, which included “obstructive sleep apnea syndrome (OSAS)” as the common term, were collected. First, a quantitative representation for each of the BM and CL articles was defined as the document-term matrix; a feasibility of discrimination was verified in the feature space obtained by principal component analysis of the document-term matrix. Furthermore, the document-term matrices of BM and CL articles were converted into term-term matrices, and their norms were used as features for the discrimination. As a result, the BM and CL articles could be discriminated with 83% sensitivity and 80% specificity. In future work, the general validity of the proposed method will be evaluated with a sufficient number of common terms besides OSAS.

Index Terms—Biomedical engineering, clinical medicine, research support, text mining

I. INTRODUCTION

A. Research field of Biomedical Engineering

The role of biomedical engineering (BM) research is to solve a medical problem using information technology. BM is studied not only in the research field of information technology, but also in other research fields such as mechanical engineering. However, the BM research presented in this paper is limited to research field of information technology. For example, researchers in the BM field conduct their studies using information-processing techniques to automatize a medical diagnosis in order to extract information about imperceptible phenomenon. To that end, BM researchers have to gain not only the knowledge of information technology, but also that of clinical medicine (CL). In other words, the BM research field is positioned in a common research area between information technology and clinical medicine as shown in Fig. 1.

B. Motivation

Generally, researchers investigate related research papers in the early stage of the study, examine the collected articles, and then decide the problem that should be addressed. However, there is difficulty in this investigation for young BM researchers such as junior or senior students.

As described above, the BM research field is related to both information technology and CL research, and specific BM and CL research can have common terms. For example, let us assume this common term is “arrhythmia.” A CL article might present a case report of arrhythmia patients. 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 CL articles. This causes young BM researchers to have difficulty in investigating BM articles exclusively. For example, if a BM researcher queries “arrhythmia” on a search engine, all the articles in the search results may not necessarily be BM articles; both BM and CL articles can be included in the search result. Experienced BM researchers can easily cull only the BM articles, but young BM researchers, not being overly familiar with academic terms in information technology and clinical medicine, cannot select them with ease, as it may take time to determine if a searched article is a BM article or not. Therefore, in this paper, a text mining-based technique for young BM researchers, to cull only BM articles from search results, is proposed.
Text mining is a technique used to quantitatively analyze documents, and is used to solve issues in varied research fields. For example, researches [1], [2] analyzed literary works quantitatively, and attempted to effect new solutions for old problems in the history of literature. In literatures [3], [4], researchers analyzed documents written on a micro-blog or Facebook, and evaluated authors’ political thought. In addition, text mining techniques are applied in marketing [5], [6].

In this research, aimed at supporting young BM researchers, BM and CL articles are quantified with a text mining technique, and features to effectively discriminate between BM and CL articles are evaluated.

II. BM AND CL ARTICLES USED IN THIS RESEARCH

In this research, obstructive sleep apnea syndrome (OSAS) is used as the common term to evaluate the feasibility of our proposed method. OSAS is the most common type of sleep apnea and is caused by repetitive occlusions of the upper airways. The OSAS causes hypertension, arrhythmia, cardiac arrest, diabetes, or dyslipidemia, which causes further risks of brain infarct or cardiac infarct. Therefore, for early detection and treatment of OSAS, many biomedical engineering and clinical approaches have been presented recently.

In this research, 20 BM and 20 CL articles were collected from many publications related to OSAS. To name a few, literature including [7], [8], [9] were selected as BM articles, which mainly present signal processing techniques to automatically detect sleep apnea episodes from long-duration ECG recordings. On the other hand, papers [10], [11], [12] were selected for CL articles. These papers present case reports of OSAS patients or a mortality risk due to sleep apnea.

III. FEASIBILITY EVALUATION FOR DISCRIMINATION BETWEEN BM AND CL ARTICLES

A. Creation of weighted document-term matrices

To discriminate between BM and CL articles, text documents must be quantified. In this research, the document-term matrix created from collected BM and CL articles [13] is adopted as shown in Eq. (1).

\[
D = \begin{pmatrix}
    tf(t_1, d_1) & \cdots & tf(t_n, d_1) \\
    \vdots & \ddots & \vdots \\
    tf(t_1, d_j) & \cdots & tf(t_n, d_j)
\end{pmatrix}
\]  \hspace{1cm} (1)
In Eq. (1), the entry \( tf(t_i, d_k) \) gives a frequency of term \( t_i \) in a document \( d_k \). The number of columns \( N \) corresponds to the total number of terms contained in a dictionary created by collected BM and CL articles. In this research, \((40 \times N)\) document-term matrices are regarded as quantitative representations for BM and CL articles.

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

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

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), \( tf \) representation \( tf(t_i, d_k) \) shown in Eq. (1) is converted into a \( tf-idf \) representation \( tfidf(t_i, d_k) \) with a following Eq. (3):

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

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

**B. Discriminant analysis for BM and CL articles**

Generally, document-term matrix is high dimensional. Therefore, dimension reduction is generated by a decomposition algorithm such as singular value decomposition or principal component analysis (PCA). In this research, PCA is adopted to generate dimension reduction for the document-term matrix. To visualize the distribution of the principal components of BM and CL articles’ document-term matrices, we focus on first and second principal components. In Fig. 2, each “+” (or “o”) illustrates a position of a BM (or CL) article in a two-dimensional plane described with first and second principal components of the document-term matrices. As shown in the figure, BM and CL articles can be discriminated in the feature space obtained by PCA. In addition, Table I shows a confusion matrix to indicate the accuracy of the classification result of the linear discriminant analysis (LDA) for principal components of document-term matrices, and the possibility that BM and CL articles are quantitatively discriminable is indicated.

**Table I: Result of LDA of principal components of \( tf-idf \) weighted document-term matrices**

<table>
<thead>
<tr>
<th>BM articles</th>
<th>CL articles</th>
</tr>
</thead>
<tbody>
<tr>
<td>BM articles</td>
<td>18</td>
</tr>
<tr>
<td>CL articles</td>
<td>1</td>
</tr>
</tbody>
</table>

Fig 2. Two-dimensional mapping with first and second principal components of \( tf-idf \) weighted document-term matrices of BM and CL articles
However, the LDA method is not applicable for this research because LDA requires a learning dataset to obtain a classifier (it cannot be assumed that researchers previously have articles containing the common term for machine learning when searching related articles for their new research). Therefore, an effective discrimination algorithm without learning data is proposed in Section 4.

IV. METHOD

Here, a method, which does not require learning data to discriminate between BM and CL articles, is presented. The process flow is shown in Fig. 3.

- **Collection of articles related to common term**
  In this research, it is assumed that a sufficient number of articles containing the common term are manually collected by a researchers. The following procedure is used to distinguish between BM and CL articles.

- **Creation of document-term matrices**
  First, a dictionary containing $N$ terms $t = \{t_1,t_2, \cdots ,t_N\}$ is created from the collected BM and CL articles. Next, a $(1 \times N)$ document-term matrix $D_k$ is created for each document $d_k$, $k = \{1,2, \cdots ,j\}$ ($j = 40$ in this study), and each element of each document-term matrix is normalized by Eq. (3); the $N \times 1$ tf–idf weighted document-term matrices are generated.

- **Creation of term-term matrices**
  To extract a feature from each document $d_k$, a term-term matrix $T_k$ is computed as follows:

$$T_k = wD_k^T wD_k , \quad (4)$$

where $wD_k$ is a tf–idf weighted document-term matrix for $k$-th document, and $wD_k^T$ is a transposed matrix of $wD_k$.

As described above, BM and CL articles are distinguished with a feature extracted from their term-term matrices. The examples for visualized term-term matrices are illustrated in Fig. 4. Fig. 4 (a) shows an average term-term matrix calculated from 20 term-term matrices of BM articles, and Fig. 4 (b) illustrates the average term-term matrix of CL articles. The colors shown in Fig. 4 represent magnitudes of the element; brown indicates larger magnitudes, whereas blue indicates smaller ones. These figures indicate that the term-term matrix of BM articles is sparser than that of CL articles. Hence, the norm of a term-term matrix is proposed as a feature to distinguish between BM and
CL articles, because it can be assumed that the norm of the term-term matrix of BM articles is smaller than that of CL articles.

Fig. 5 (a) and Fig. 5 (b) show the histograms of the norms calculated for the term-term matrices of 20 BM articles and 20 CL ones, respectively. As shown in Fig. 5, although a distribution of norms of term-term matrix of BM articles overlaps to some extent with that of CL articles, it can be seen that the norms of the term-term matrix of BM articles are substantially smaller than those of CL ones. Thus, it is proposed that a threshold level of the norm, to distinguish between BM and CL articles, needs to be defined first. If the norm calculated from the term-term matrix of a certain article exceeds the threshold level, the article is classified as CL. Similarly, an article is classified as BM when the norm of the term-term matrix does not exceed the threshold level. In this research, the empirical threshold of the norm level is set to 120.

V. RESULTS AND DISCUSSION

To evaluate the accuracy of the discrimination between BM and CL articles, the sensitivity and specificity are calculated using the following equations (5) and (6).

\[
Sensitivity = \frac{TP}{TP + FN}. \quad (5)
\]

\[
Specificity = \frac{TN}{TN + FP}. \quad (6)
\]

In equations (5) and (6), TP, TN, FP, and FN represent the quantities of true positives, true negatives, false positives, and false negatives, respectively. In this paper, TP indicates the number of correct discrimination for BM articles. Table II shows the discrimination accuracy.

<table>
<thead>
<tr>
<th></th>
<th>Sensitivity (%)</th>
<th>Specificity (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>83.2</td>
<td>80.1</td>
</tr>
</tbody>
</table>

Fig 4. Visualized features of BM and CL articles: (a) average term-term matrix of BM article, (b) average term-term matrix of CL article
In Section 3, the document-term matrix was defined as the quantitative representation for BM and CL articles, and its effectiveness was shown. In accordance with this result, the method for discrimination between BM and CL articles was presented in Section 4.

The feature for discrimination was defined as the norm of the term-term matrix converted from the document-term matrix, and the discrimination was performed by determining whether a calculated norm exceeds the empirical threshold level. As a result, the sensitivity was approximately 83%, and the specificity was 80%; this indicates that most articles were correctly discriminated.

As shown in Fig. 4, the term-term matrix of BM articles was sparser than that of CL articles. In other words, this fact indicates that CL articles include richer vocabularies as compared to BM articles. In fact, even though young BM researchers do not have a pressing need to read CL articles, which probably include profound medical terms, they still have to read those articles to determine if the articles are BM articles or not. In this context, the proposed method for discrimination between BM and CL articles is useful because it can save young BM researchers time and effort when reading many articles.

However, the common term focused in this paper was “OSAS” only, and a general validity of the proposed method has not been guaranteed yet. Therefore, it is not clear that the term-term matrix of a BM article is consistently sparser than that of a CL article with the same common term. Additionally, a method for determining the threshold level must be generalized. In future work, the general validity of the proposed method will be evaluated with other common terms such as “T-wave alternans”, “atrial fibrillation”, “evoked potential” or “epilepsy”.

VI. CONCLUSION

The goal of this research was to automatically distinguish between BM and CL articles based on a text mining technique. To assess the feasibility of the proposed method, 20 BM and 20 CL articles related to the OSAS were collected. These articles were quantified with the document-term matrix, and then the norms of the term-term matrices calculated from document-term matrices were defined as a feature for the discrimination. The results indicated that the BM and CL articles could be discriminated with 83% sensitivity and with 80% specificity.

ACKNOWLEDGMENT

The author would like to thank Dr. Yuichi Okuyama of the University of Aizu for his valuable comments.
REFERENCES


AUTHOR BIOGRAPHY

Motoki Sakai received his B.E., M.E., and Ph.D. degrees from the University of Aizu, Aizu-wakamatsu, in 2003, 2005, and 2009. He joined QRS Corporation in 2009, and worked as an extra-ordinary assistant professor at the Fukushima University Graduate School of Symbiotic Systems Science and Technology from 2010 to 2013. Currently, he works as an assistant professor at Tokyo Denki University in the School of Information Environment. His current research interests include data science.