| | | | |
|---|---|---|---|
| 氏　　名 | 松　尾　亮　輔 | | |
| 学　位　の　種　類 | 博士(知識科学) | | |
| 学　位　記　番　号 | 博知第 205 号 | | |
| 学位授与年月日 | 平成 29 年 6 月 23 日 | | |
| 論　文　題　目 | 医学文書における意味を考慮した単語重み付け手法の開発 | | |
| 論文審査委員 | 主査　Ho Bao Tu | 北陸先端科学技術大学院大学 | 教授 |
| | 小坂　満隆 | 同 | 教授 |
| | 池田　　満 | 同 | 教授 |
| | DAM HIEU CHI | 同 | 准教授 |
| | 佐藤　賢二 | 金沢大学大学院 | 教授 |

## 論文の内容の要旨

Term weighting where a term is given a numerical weight regarding its importance, is fundamental to analyze text data. As term weighting can transform documents into the computable forms in a vector space, it enables to execute various document analysis such as text classification, clustering, information retrieval and so on. The traditional measure used in term weighting is TFIDF, derived from the term frequency and the inverse document frequency. TFIDF is simple and effective, and it forms a popular base for advanced algorithms in spite of its age. However, the term importance captured by term weighting methods using TFIDF and its variants does not relate to the term meanings but only to the frequencies. These methods are not suitable for applications that require considering the term meanings.

In medical domain, therefore, semantic term weighting (STW) methods have been developed aiming at assigning weights to document terms based on their meanings by exploiting ontologies and class information of terms. However, these methods are developed for a certain task in medicine. There is no framework of STW that can correspond to any task. Moreover, there is no framework to put effectively term weighting into practice. In order to exploit the computable forms of documents by term weighting for secondary use, we need to consider the nature of documents, the target of analysis and the adequate terms' meanings. To this end, we apply the idea of frame semantics to term weighting. The frame semantics is a research program in empirical semantics that emphasizes the continuities between language and experience. It assigns term meanings based on the frame that characterize a situation. The benefit of the exploitation of the frame semantics is to keep an assumption that the determination of term importance is not unique but diverse depending on the nature of documents and the target of analysis. Moreover, it considers the encyclopedic semantics of terms that is adequate to put semantic term weighting into practice.

The objective of this thesis is to develop STW methods for medical document analysis considering the idea of the frame semantics. We especially focus on two important targets: information retrieval on medical documents and prediction of patients' conditions on EMRs such as mortality prediction. To this end, we propose two

frameworks: a framework of STW using a proposed procedure to apply the idea of the frame semantics to term weighting and a common framework of STW based on a proposed medical knowledge representation. The key idea is to hierarchically divide the terms into categories by exploiting ontologies and class information of terms based on medical knowledge and machine learning techniques. The terms in each category are reasonably considered to have the same medical importance (except the case that a term's weight is the continuous value) regarding a certain aspect of terms' meanings. The categories containing terms are exploited to represent various aspects of terms' meanings on computer. Based on the proposed frameworks, we developed STW methods for the two targets. As the proposed STW methods were verified by the experimental evaluations, we attained the objectives by using the proposed frameworks.

As term weighting transforms documents into computable forms, the proposed STW method considering the idea of the frame semantics can apply to various applications as secondary use when keeping various aspects of terms' meanings in medicine.

Key words: Semantic term weighting, Medical documents, Ontology, Class information, Frame semantics

## 論文審査の結果の要旨

Term weighting is to give a numerical weight to a word in a document regarding its importance to the document. The traditional measure used in term weighting is TFIDF, derived from the term frequency and the inverse document frequency. TFIDF is simple and effective, and it forms a popular base for advanced algorithms in spite of its age. However, the term importance captured by term weighting methods using TFIDF and its variants does not relate to the term meanings but only to the frequencies, and not appropriate for clinical text analysis.

The objective of this thesis is to develop semantic term weighting methods for medical document analysis. A framework where a weight w is derived from wTFIDF and wsemantics was proposed. The wsemantics is represented by two types of semantics as wsemantics1 and wsemantics2 where wsemantics1 is based on the medical importance of the term (patient independence) and wsemantics2 is based on the severity status of the patient (patient dependence). The resources UMLS and ICD-10 are employed to classify terms in EMRs into basic categories. A ranking of causes of death is used to identify the medical importance of terms and classify them into deep categories (wsemantics1), and the Charlson comorbidity index is used to identify the patient disease status with its severity (wsemantics2).

The experimental evaluation demonstrated that the proposed semantic term weighting methods outperformed statistical TFIDF-based methods. The proposed semantic term

weighting methods can apply to various applications in medicine.

The study has shown the candidate's ability of independently conducting the scientific research, and this is an very well done dissertation and we approve awarding a doctoral degree to Mr. Matsuo Ryosuke.