Koroner Arter Hastalığıyla İlgili Özetlerden Anahtar Kelimelerin Çıkartılması: Bir Bilgi Çıkarım Uygulaması

Başak OĞUZ[a], Mehmet Kemal SAMUR[a]

[a] *Biyoistatistik ve Tıp Bilişimi AD, Akdeniz Üniversitesi, Antalya*

Extracting Keywords from Coronary Artery Disease Related Abstracts: an Information Extraction Application

Abstract: *Coronary artery disease refers to the failure of coronary circulation to supply adequate circulation to cardiac muscle and surrounding tissue. It is already the most common form of disease affecting the heart and an important cause of death in the world. Information Extraction is one of the effective approaches to text mining that extracts useful information from natural language texts without a "deep understanding" of the text. The purpose of this study is to extract the word patterns and determine the key/content word or word groups (words synthesizing the content of the article) by using the abstracts. Since the abstracts are the short summaries of the entire article and contain a higher ratio of keywords to the total number of words, they were chosen for the text collection of this current study. The study contains a total of 1478 abstracts that are related with coronary heart disease and published within the last year. The abstracts were retrieved from MedLine by using the MeSH (Medical Subject Headings) terms ("Coronary Artery Disease") and the outcomes were stored following an XML schema conversion. In order to derive keywords from the abstract of an article, we first explored the associations between the words. We take the sentence as the unit of text to look for associations, that is, two words are associated if they co-occur repeatedly in sentences within abstracts. By analyzing the patterns of words, keywords were extracted from the abstracts. This should help researchers in developing guidelines and decision support systems.*

Key Words: Information Extraction; Coronary Artery Disease

Özet: *Koroner arter hastalığı, koroner arterlerin duvarlarında oluşan plaklardan ötürü ortaya çıkan bir hastalıktır. Dünyada şu an en fazla ölüme neden olan hastalık koroner arter hastalığıdır. Bilgi çıkarım, doğal dille yazılmış metinlerden kullanışlı bilgilerin çıkartılması ile ilgilenen bir metin madenciliği yaklaşımıdır. Bu çalışmanın amacı, tıbbi makale özetlerini kullanarak kelime örüntülerini ortaya çıkarmak ve anahtar/önemli kelime ve kelime gruplarını belirlemektir. Özetler, bütün makaleyi kısaltarak anlattığı ve tam metine kıyasla toplam kelime sayısına göre daha fazla anahtar kelime içerdiği için bu çalışmada tercih edilmiştir. Çalışmada, MedLine'dan elde edilen koroner arter hastalığı ile ilgili son 1 yılda yayınlanmış makalelere ait 1478 özet kullanılmıştır. Tıbbi özetlere, MeSH (Medical Subject Headings) terimleri ("Coronary Artery Disease") kullanılarak erişilmiş ve XML formatına dönüştürülerek kaydedilmiştir. Anahtar kelimeleri özetlerden çıkartmak için öncelikle kelimeler arasındaki ilişkilere bakılmıştır. Kelimelerin birlikte bulunma sıklıkları incelenmiş, eğer iki kelime tekrarlı olarak birlikte bulunuyorsa iki kelimenin birbiriyle ilişkili olduğu varsayılmıştır. Bu örüntülerden yola çıkarak önemli/anahtar kelimeler belirlenmiştir. Bu yazılımla birlikte literatürde bulunan metinler içerisindeki önemli kavramların belirlenerek klinik rehberlerin veya karar destek sistemlerinin geliştirilmesinde araştırmacılara fayda sağlanması planlanmaktadır.*

Anahtar Kelimeler: Bilgi Çıkarım; Koroner Arter Hastalığı

Sorumlu Yazarın Adresi

Başak Oğuz, Akdeniz Üniversitesi Biyoistatistik ve Tıp Bilişimi AD.
E-posta: basakoguz@akdeniz.edu.tr