IDMC'07
20-21 Nov.2007
Amir Kabir University

رهیافت کاربردی تکنیکهای پردازش زبان طبیعی و خوشه بندی اطلاعات در داده کاوی پایگاه داده MEDLINE به منظور آنالیز طولی مقالات زیست – پزشکی

فرشید مجیدفر(1)، فرزان مجیدفر(2) ، محمد تفضلی شادپور(3)

چکیده

پیشرفت تکنولوژی در شاخه‌های مختلف زیست‌شناسی و پزشکی ، موجب توسعه تعداد بسیار زیادی از الگوریتم‌ها برای پردازش و تحلیل داده‌های زیستی شده است . استفاده از منابع غیر عددی اطلاعات برای کشف لایه‌های پنهان زیست‌شناسی یکی از این رهیافت ها است. در این مقاله طراحی ، گسترش و اجرای رهیافتی برای یکپارچه‌سازی روش‌های مختلف متن‌کاوی در مقالات پایگاه داده مدلاین به منظور تجزیه و تحلیل طولی مقالات زیست – پزشکی را شرح داده‌ایم. رهیافت پیشنهادی و تحلیل‌های صورت گرفته در این تحقیق برای توسعه و میزان سازی دقیق متدولوژی‌های داده کاوی متنی مقاله های علوم زیست – پزشکی به منظور طبقه بندی طبیعی مقالات ، ارزیابی و پیش بینی روند توسعه تکنولوژی و اکتشاف دانش کاربرد خواهند داشت .

روش پیشنهادی مبتنی بر استفاده از دسترسی به پایگاه داده MEDLINE برای شناسایی آخرین تحقیقات، و جمع آوری مقالات زیست – پزشکی در یک حوزه خاص تکنولوژیک است این مقالات در طی شش مرحله که هر یک به صورت مستقل در محیطهای مختلف برنامه نویسی و به صورت نرم افزار رایانه ای پیاده سازی شده اند ، پردازش می شوند. ابتدا مقالات جمع آوری شده برای تولید فهرستی از واژگان کلیدی پیش پردازش متنی می گردند. سپس تکنیک‌های پردازش زبان طبیعی(NLP) مانند فیلترهای stop-word و part of speech برای پاکسازی به فهرست اعمال می شوند. براساس فهرست واژگان اختصاصی تولید شده ، هر یک از مقالات تبدیل به برداری از واژگان کلیدی می شوند. بردارهای به دست آمده به یک الگوریتم خوشه‌بندی سلسله مراتبی وارد شده تا مقالات بر اساس فهرست واژگان کلیدی به صورت طبیعی گروه‌بندی شوند. سپس گروه های ایجاد شده بر اساس فاکتورهایی مانند زمان انتشار مقاله به صورت طولی تجزیه و تحلیل می شوند و در مرحله نهایی یا تجسم سازی (visualization) ، نتایج آنالیز به تصویر در می آیند. برای نمایش و ارزیابی رهیافت شرح داده شده ، مقالات مدلاین در زمینه تله کاردیولوژی (telecardiology) به عنوان ورودی مورد استفاده قرار گرفتند و بر اساس مراحل گفته شده مورد تجزیه وتحلیل طولی قرار گرفته ، نتایج به تصویر درآمدند..

کلمات کلیدی

داده کاوی ، متن کاوی ، خوشه بندی سلسله مراتبی ،مدلاین ،مقالات زیست-پزشکی ، پیش پردازش ، پاکسازی ،تجسم سازی ، تله کاردیولوژی

---

(1) دانشگاه صنعتی مالک اشتر – مجتمع برق و الکترونیک- مهندس الکترونیک - farshid.majidfar@gmail.com
(2) دانشگاه صنعتی امیر کبیر – دانشکده مهندسی پزشکی –دکترای پزشکی/کارشناس ارشد مدیریت فنآوری اطلاعات پزشکی – majidfar@telemed.ir
(3) دانشگاه صنعتی امیر کبیر – دانشکده مهندسی پزشکی – عضو هیئت علمی (استادیار) – tafazoli@aut.ac.ir