نخستین کنفرانس بین المللی پردازش خط و زبان فارسی
۱۵ و ۱۶ شهریور ۱۳۹۱
دانشگاه سمنان - دانشکده مهندسی برق و کامپیوتر

ایجاد خودکار نمایه (Index) برای تصاویر متنی به زبان فارسی

علیرضا نوحی[۱]، فرزین یغمایی[۲]
alireza.noohi@students.semnan.ac.ir[۱]
f-yaghmaee@semnan.ac.ir [۲]

چکیده

در سال های اخیر، تشخیص نوری متون فارسی و عربی به طور گسترده ای مورد توجه قرار گرفته است. در این مقاله سعی داریم روشی برای ساخت خودکار نمایه، از تصاویر متنی به زبان فارسی ارائه کنیم. از این رو، ابتدا به تفکیک لغات و خطوط به کمک هیستوگرام های عمودی و افقی پرداخته و برای بهبود کیفیت لغات جدا شده از عملگرهای مورفولوژی استفاده می کنیم، تا ریز فاصله های موجود بین کلمات را حذف کنیم. سپس هر کلمه بدست آمده را به عنوان تصویر نمونه در نظر می گیریم واز طریق محاسبه ضریب همبستگی آن را با سایر کلمات موجود در متن مقایسه می کنیم.

در صورت مشابه بودن کلمه ای از متن با تصویر نمونه ، آن را به همراه شماره صفحاتی که این کلمه در آنها آورده شده است به عنوان یک ورودی در جدول نمایه درج می کنیم. نتایج حاصل از الگوریتم ، نشان دهنده دقت حدود ۹۰ درصد الگوریتم در ایجاد نمایه بر روی متون فارسی است.

کلمات کلیدی

OCR ، تصاویر متنی فارسی، نمایه

۱- مقدمه

تشخیص نوری حروف (OCR) یکی از مهمترین کاربردهای تشخیص الگو است. در سال های اخیر تحقیقات زیادی در این زمینه صورت گرفته است. چندین الگوریتم برای تشخیص کاراکترهای چاپی و دست نویس پیشنهاد شده است. با این حال هنوز مسائل حل نشده ای در این زمینه وجود دارد. [۱] [۲] [۳]

با وجود کارهای فراوان و خوبی که برای تشخیص کاراکترهای زبان هایی مثل لاتین وچینی انجام شده، اما به علت مشکلاتی مثل پیوسته بودن حروف، کلمات و انحنای حروف در زبان فارسی و زبانهای مشابه هنوز زمینه تحقیقاتی برای محققان فراهم است و گروه های پژوهشی زیادی بر روی آنالیز متون فارسی/عربی تمرکز کرده اند. [۴] [۵] [۶]

با رشد و گسترش فعالیتهای علمی، جهت انجام هرگونه پژوهش علمی باید در وهله اول مطمئن شد که چنین فعالیتی در گذشته در زمینه مورد نظر انجام نگرفته و دروهله دوم از کلیه فعالیت هایی که می تواند در همین زمینه موثر باشد، استفاده گردد. بنابراین اطلاع از وجود مدارک علمی در زمینه های گوناگون وآگاهی از محتوای اطلاعاتی آن ها را می توان یکی از عوامل مهم پیشرفت های پژوهشی دانست.این فعالیت تا حدود زیادی در قلمرو نمایه سازی قرار می گیرد، به طوری که تعریف نمایه سازی نیز در تایید همین مطلب می باشد.

به عبارت دیگر نمایه، فهرستی از مطالب یا موضوعات و مفاهیم مهم است که معمولاً در آخرین صفحات کتاب آمده و به شماره صفحه ارجاع می دهد وچون بطور الفبایی مرتب می شود به سرعت می توان صفحه حاوی مطلب را درآن پیدا کرد. در واقع نمایه عامل ارتباطی بین منابع اطلاعاتی واستفاده کننده یا کاربر است.

هدف ما در این مقاله ارائه روشی برای تشخیص لغات فارسی وساخت نمایه از روی تصاویر متنی و یا کتاب های اسکن شده و یا کتاب های الکترونیکی است. با توجه به بررسی های صورت گرفته (توسط مولفان)، تاکنون کار مشابهی در زبان فارسی و حتی در زبان انگلیسی انجام نشده است.

ادامه مقاله به شرح زیر است :در بخش دوم، شرح الگوریتم ایجاد نمایه آورده خواهد شد، در بخش سوم نتایج الگوریتم ودر بخش چهارم نتیجه گیری وکارهای آینده بیان خواهد شد.

۲- شرح الگوریتم ایجاد نمایه

تشخیص و استخراج لغات در متن چاپی فارسی در ۵ گام نشان داده شده در شکل ۱ انجام می شود.