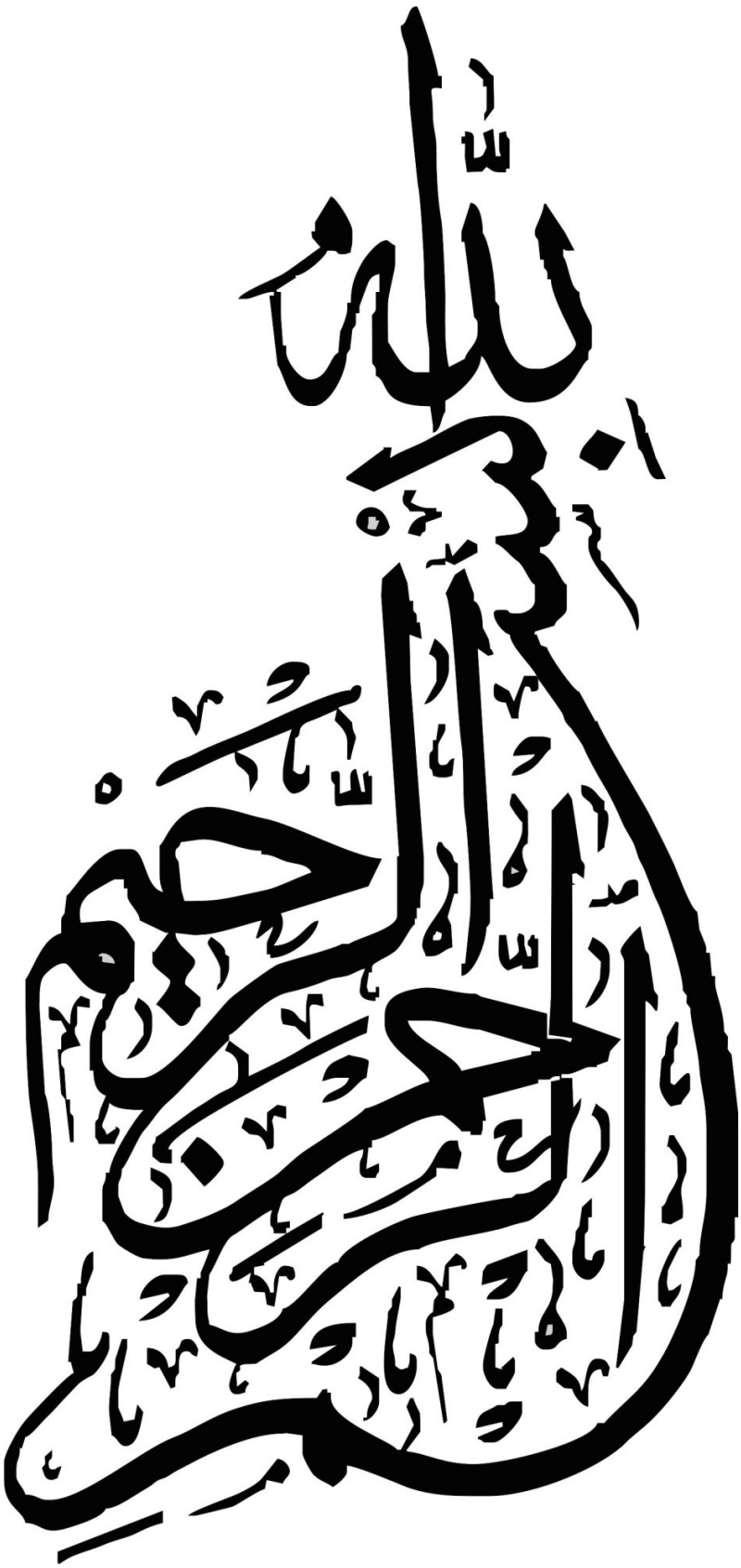

دانشگاه صنعتی خواجه نصیرالدین طوسی
تاسیس ۱۳۰۷

رساله دکترای الکترونیک

عنوان :

بازیابی متون چاپی فارسی بر اساس پرس و جوی کلمات

استاد راهنما : دکتر هوشنگ حسیبی

نگارنده : یعقوب پوراسد

زمستان ۱۳۹۰

چکیده

در این پایان نامه روشی جدید برای بازیابی و جستجوی کلمات فارسی از میان متون تصویری فارسی ارائه شده است. همچنین روشی مبتنی بر اندازه و شکل نقاط موجود در مستند پس از آستانه گذاری، جهت تشخیص قلم و اندازه قلم مستندات تصویری فارسی ارائه شده است. روش ارائه شده برای تشخیص قلم از اولین روش هایی می باشد که می تواند با دقت بسیار بالایی اندازه قلم مستند فارسی را هم تشخیص دهد. برای ارزیابی روشهای ارائه شده برای تشخیص قلم و نیز سیستم بازیابی کلمات، چندین پایگاه تصویری با استفاده از کامپیوتر ایجاد شدند. پایگاه تصویری اصلی بکار رفته برای ارزیابی، شامل ۴۴۸ تصویر تمیز و بدون نویز بود که سیستم ارائه شده با دقت بیش از ٪۹۸ قلم و اندازه قلم آنها را تشخیص داد. همچنین سیستم کلی بازیابی کلمات، با دقت ٪۸۶ در نرخ بازیابی ٪۸۲ قادر به بازیابی کلمات از مستندات تصویری بود. این نرخ دقت و بازیابی با ارزیابی سیستم بر روی ۲۰۰ کلمه فارسی بدست آمده است. همچنین یک پایگاه مستندات تصویری کوچک از مستندات تصویری اسکن شده (شامل ۱۳ مستند تصویری نوشته شده در نرم افزار Ms Word که به صورت کاملا تمیز و بدون کجی اسکن شده اند) هم برای بررسی امکان پیاده سازی سیستم بر روی تصاویر اسکن شده واقعی ایجاد شد که مشاهده شد که هر دو روش ارائه شده قابل پیاده سازی بر روی تصاویر اسکن شده واقعی هم هستند. علاوه بر پایگاه تصاویر گفته شده، عملکرد سیستم بر روی یک پایگاه تصویری شامل تعدادی تصویر اسکن شده در شرایط غیر ایده آل دارای نویز و کجی هم مورد آزمایش قرار گرفت که نتایج حاصل نشان دهنده ضعف سیستم تشخیص قلم در تشخیص قلم مستندات نویزی و کج می باشد.

کلمات کلیدی: بازیابی مستندات تصویری فارسی ، جستجوی کلمات کلیدی، تشخیص قلم، تشخیص اندازه قلم، آستانه گذاری، هیستوگرام

بسمه تعالی

تاسیس ۱۳۰۷ دانشگاه صنعتی خواجه نصیرالدین طوسی	**اظهارنامه دانشجو**	**شماره:** **تاریخ:**

اینجانب **یعقوب پوراسد** دانشجوی دکترای رشته برق گرایش الکترونیک دانشکده مهندسی برق و کامپیوتر دانشگاه صنعتی خواجه نصیرالدین طوسی گواهی می نمایم که تحقیقات انجام شده در پایان نامه با عنوان

بازیابی متون چاپی فارسی بر اساس پرس و جوی کلمات

با راهنمایی استاد محترم **جناب آقای دکتر هوشنگ حسیبی** ، توسط شخص اینجانب انجام شده و صحت و اصالت مطالب نگارش شده در این پایان نامه مورد تأیید می باشد، و در مورد استفاده از کار دیگر محققان به مرجع مورد استفاده اشاره شده است. بعلاوه گواهی می نمایم که مطالب مندرج در پایان نامه تاکنون برای دریافت هیچ نوع مدرک یا امتیازی توسط اینجانب یا فرد دیگری در هیچ جا ارائه نشده است و در تدوین پایان نامه چارچوب (فرمت) مصوب دانشگاه را به طور کامل رعایت کرده ام.

امضا دانشجو :

تاریخ :

بسمه تعالی

تاسیس ۱۳۰۷ دانشگاه صنعتی خواجه نصیرالدین طوسی	**حق طبع و نشر و مالکیت نتایج**	شماره: تاریخ:

۱- حق چاپ و تکثیر این پایان‌نامه متعلق به نویسنده آن می‌باشد. هرگونه کپی برداری بصورت کل پایان‌نامه یا بخشی از آن تنها با موافقت نویسنده یا کتابخانه دانشکده مهندسی برق و کامپیوتر دانشگاه صنعتی خواجه نصیرالدین طوسی مجاز می‌باشد.

ضمناً متن این صفحه نیز باید در نسخه تکثیر شده وجود داشته باشد.

۲- کلیه حقوق معنوی این اثر متعلق به دانشگاه صنعتی خواجه نصیرالدین طوسی می‌باشد و بدون اجازه کتبی دانشگاه به شخص ثالث قابل واگذاری نیست.

همچنین استفاده از اطلاعات و نتایج موجود در پایان نامه بدون ذکر مراجع مجاز نمی‌باشد.

بسمه تعالی

تاسیس ۱۳۰۷ دانشگاه صنعتی خواجه نصیرالدین طوسی	**تأییدیه هیأت داوران**	**شماره:** **تاریخ:**

هیأت داوران پس از مطالعه پایان نامه و شرکت در جلسه دفاع از پایان نامه تهیه شده تحت عنوان :

بازیابی متون چاپی فارسی بر اساس پرس و جوی کلمات

توسط آقای **یعقوب پوراسد** ، صحت و کفایت تحقیق انجام شده را برای اخذ درجه دکترای رشته برق گرایش الکترونیک در تاریخ مورد تأیید قرار می دهند.

۱- **استاد راهنما**	جناب آقای دکتر هوشنگ حسیبی	امضا
۲- **ممتحن داخلی**	جناب آقای دکتر محمد تشنه لب	امضا
۳- ممتحن داخلی	جناب آقای حمید ابریشمی مقدم	امضا
۴- **ممتحن خارجی**	جناب آقای دکتر احسان اله کبیر	امضا
۵- **ممتحن خارجی**	جناب آقای دکتر افشین ابراهیمی	امضا
۶- **نماینده تحصیلات تکمیلی**	جناب آقای دکتر مسعود علی اکبر گلکار	امضا

فهرست جداول

فهرست اشکال

فهرست علائم و اختصارات

OCR: Optical Character Recognition

CBD: City Block Distance

CORR: Correlation (Used to represent Correlation Similarity Measures)

DTW: Dynamic Time Warping

EDM: Euclidean Distance Measure

K-NN: K-Nearest Neighbor

NSHP-HMM: Non-Symmetric Half Plane Hidden Markov Model

P2D-HMM: Pseudo 2 Dimensional Hidden Markov Model

PHMM: Planar Hidden Markov Model (Another name of P2D-HMM)

SLH: Scot and Longuet Higgins algorithm

SSD: Sum of Squared Distances

SC: Shape Context algorithm.

SRF : Sobel and Robert Filters

WED : Weighted Euclidean Distance

SVM : Support Vector Machine

GSC : Gradient Structural Concavity Features

DP : Dynamic Programming

CSC : Character Shape Coding

WSC : Word Shape Coding

BCT : B Classification Tree

فصل اول

بازیابی مستندات تصویری

۱-۱ مقدمه

از زمان پیدایش نوشتار برای برقراری ارتباط میان افراد کاغذ به عنوان یک ابزار مهم برای درج اطلاعات مورد استفاده قرار گرفته است. اما امروزه محیط‌های الکترونیکی به دلیل نگهداری آسان و دسترسی سریع رواج فراوانی یافته و به عنوان جایگزینی برای کاغذ مطرح شده اند . از سوی دیگر سهولت استفاده، فراگیر بودن و وجود حجم عظیمی از اطلاعات کنونی به روی کاغذ، محققان را بر آن داشته است تا به دنبال روش‌هایی برای خواندن اتوماتیک این اطلاعات از روی کاغذ و تبدیل آن به شکل الکترونیکی باشند. امروزه کاربردهای فراوانی برای شناسایی متون وجود دارد از آن میان می‌توان به پردازش چک‌های بانکی شناسایی آدرس و کدهای پستی، کمک به نابینایان برای خواندن و... اشاره کرد. پردازش نوشتار یکی از شاخه‌های مربوط به شناسایی الگو است که هدف نهایی آن تقلید از انسان برای خواندن و درک متون است. اما با وجود پیشرفت‌های به دست آمده، هنوز قدرت این سیستم‌های پردازش‌گر در خواندن حروف با توانایی بشر فاصله دارد. کاربردهای عملی فراوان و طبیعت جذاب این رشته توجه پژوهشگران زیادی را به خود جلب کرده و پیشرفت‌های محسوسی در این زمینه به دست آمده است. با پیدایش کامپیوتر تلاش برای حل مسئله شناسایی نوشتار از دهه ۱۹۴۰ میلادی با تحقیقات در مورد شناسایی حروف و ارقام مجزای چاپی لاتین شروع شد، اولین ماشین‌های پردازش‌گر حروف که سیستم شناسایی کننده آن‌ها به طور مکانیکی کار می‌کرد، در اوایل دهه ۱۹۵۰ میلادی به بازار عرضه شدند. این ماشین‌ها فقط قادر به شناسایی اعداد لاتین تایپ شده با یک قلم خاص بودند. در دهه ۱۹۷۰ میلادی، سیستم‌های شناسایی کننده نرم‌افزاری با توانایی تشخیص متون تایپی تک قلمی پدید آمدند. در دهه ۱۹۸۰ میلادی پیشرفت زیادی در زمینه شناسایی متون لاتین صورت گرفت به طوری که سیستم‌های

موجود می‌توانستند متون تایپ شده با چندین قلم را نیز با سرعت بالا تشخیص دهند. امروزه سیستم‌های تجاری بازشناسی متون لاتین علاوه بر تشخیص متون تایپی قادر به درک متون دستنویس نیز هستند. در چند دهه اخیر کارهای زیادی برای شناسایی حروف زبان‌های لاتین، چینی و ژاپنی انجام شده است. اما برای شناسایی حروف عربی و فارسی، با اینکه تعداد زیادی از مردم جهان برای نوشتن از این حروف استفاده می کنند، کارهای انجام شده نسبتاً کم و پراکنده بوده است و نتایج به دست آمده چندان رضایت-بخش نیستند. شاید بتوان ویژگی‌های نوشتاری منحصر به فرد این زبان‌ها را به عنوان یکی از موانع موجود در این راه نام برد. در ادامه تعدادی از ویژگی‌های نوشتاری زبان فارسی به طور خلاصه ذکر می-گردد.

۲-۱ ویژگی‌های نوشتاری زبان فارسی

۱-۲-۱ پیوستگی حروف

بر خلاف نوشتار لاتین، متون فارسی هم در حالت تایپی و هم در حالت دستنویس به صورت پیوسته از راست به چپ نوشته می‌شوند. اگر چه به صورت عادی در هر یک کلمه، هر حرف به حرف قبلی و بعدی متصل است ، اما تعدادی ازحروف هستند(حروف ا. د . ذ.ر. ز. ژ. و) که فقط می‌توانند از سمت راست به حروف ماقبل خود وصل شوند. وجود یک یا چند حرف از این مجموعه باعث می‌شود یک کلمه به تعدادی بخش پیوسته به نام شبه کلمه تقسیم شود. به عنوان مثال کلمه(کتاب) از دو زیر کلمه(کتا) و (ب) تشکیل شده است که این بخشها با فضای خالی از یکدیگر جدا می‌شوند.

۲-۲- ۱ اشکال متفاوت یک حرف

شکل حروف در زبان فارسی تابعی از محل قرار گرفتن آن در کلمه است و هر حرف بر حسب موقعیت های مختلف در کلمه(اول. وسط. آخر و جدا) می‌تواند تا چهار شکل مختلف داشته باشد به عنوان مثال شکل حرف(ه) در حالت جدا باید به شکل (هـ)در ابتدای کلمه، شکل(ﻫ) در وسط کلمه و یا شکل(ه) در آخر کلمه تغییر پیدا کند.

۳-۲- ۱ همپوشانی بین حروف

حروف در یک کلمه فارسی می‌توانند حتی بدون اتصال با هم، همپوشانی عمودی داشته باشند. معمولاً حروف(ج، چ، ر، ژ، و) با حروف بعدی خود دارای همپوشانی هستند. علاوه بر این حروف قسمت‌های مکمل حروف (آ، ک) نیز می‌توانند بخشی از حروف مجاور خود را بپوشانند. این ویژگی در نوشتار فارسی، عمل تقطیع کلمات به حروف تشکیل دهنده آن‌ها را بسیار مشکل کرده است.

۴-۲- ۱ نقطه‌دار بودن حروف

بیش از نیمی از حروف فارسی نقطه‌دار هستند. به عبارتی دقیق‌تر ۱۰ حرف دارای یک نقطه۳ حرف دارای دو نقطه و ۵ حرف دارای سه نقطه هستند. در بعضی از موارد، وجود و یا عدم وجود ، تعداد و محل قرار گرفتن نقطه‌ها تنها عامل متمایز کننده بین حروف متشابه(ب، پ، ت، ث، نـ ، یـ) (جـ ،چـ ،حـ ،خـ)،(د،ذ)، (ر،ز،ژ)،(صـ ، ضـ) (س و ش)، (عـ ، غـ)، (فـ قـ) است. جدول ۱-۱مجموعه حروف فارسی و اشکال مختلف هر حرف در موقعیت‌های مختلف را نشان می‌دهد.

۵-۲- ۱ وجود اعراب

در نوشتار فارسی، هنگامی‌که احتمال تلفظ اشتباه یک کلمه وجود داشته باشد صدا به صورت اعراب به بالا یا پائین برخی از حروف اضافه می‌شود. علاوه بر اعراب در نوشتار فارسی، علائمی از قبیل تشدید، تنوین، همزه و مد نیز وجود دارند. با توجه به شباهت ظاهری اعراب و این علائم مکمل با نقاط، جدا کردن آن‌ها به خصوص در کلمات دستنوشته کار مشکلی است.

۶-۲-۱ اندازه متفاوت حروف

اندازه و عرض تمام حروف فارسی یکسان نیستند. مثلاً حروف (ب)، (س) در حالت جدا اندازه بزرگ‌تری نسبت به حروف(د) و (ه) دارند. این تنوع در اندازه حروف کار قطعه‌بندی حروف را مشکل می‌کند.

۷-۲-۱ ادغام حروف مجاور

در برخی از شیوه‌های نوشتاری زبان فارسی، دو یا چند حرف کنار هم می‌توانند به گونه‌ای با هم ترکیب شوند که شکل حاصل شباهتی به حروف تشکیل دهنده آن نداشته باشد. چنین مواردی در نوشتار دستنویس، بلکه در متون تایپی نیز وجود دارد. متداول‌ترین ترکیب در متون تایپی، ادغام دو حرف(ل) و (ا) به صورت (لا) است. در نوشته‌های دستنویس فارسی نیز بخاطر کمک به زیبایی نوشتار و همچنین سلیقه نویسنده، شکل بعضی از حروف کنار هم به کلی تغییر می‌کند.

۸-۲-۱ تنوع فراوان در شیوه‌های نگارش

در مقایسه با نوشتار لاتین، قلم‌های فارسی در حالت چاپی دارای واحدهای نوشتاری(شامل اعـداد، اشکال مختلف حروف و علائم خاص) بسیار زیاد و در حالت دسـتنویس دارای سـبک‌هـای متعـدد مـی باشند. جدول ۱-۱ مجموعه حروف فارسی را در موقعیت‌های مختلف نشان می‌دهد.

جدول ۱-۱ مجموعه حروف فارسی و شکل آن‌ها در موقعیت‌های مختلف

مجزا	ابتدا	وسط	انتها	مجزا	ابتدا	وسط	انتها
ض	ضـ	ضـ	ض	ا	ـ	ـ	ا
ط	طـ	طـ	ط	ب	بـ	بـ	ب
ظ	ظـ	ظـ	ظ	ت	تـ	تـ	ت
ع	عـ	عـ	ع	ث	ثـ	ثـ	ث
غ	غـ	غـ	غ	ج	جـ	جـ	ج
ف	فـ	فـ	ف	ح	حـ	حـ	ح
ق	قـ	قـ	ق	خ	خـ	خـ	خ
ك	كـ	كـ	ك	د	ـ	ـ	د
ل	لـ	لـ	ل	ذ	ـ	ـ	ذ
م	مـ	مـ	م	ر	ـ	ـ	ر
ن	نـ	نـ	ن	ز	ـ	ـ	ز
ه	هـ	هـ	ه	س	سـ	سـ	س
و	ـ	ـ	و	ش	شـ	شـ	ش
ي	يـ	يـ	ي	ص	صـ	صـ	ص

۳-۱ بازیابی اطلاعات

امروزه با ازدیاد کتابخانه های دیجیتال و نیز در راستای ایجاد ادارات بدون کاغذ، تعداد فزاینده ای از مستندات تایپی و دست نویس شامل کتاب ها، روزنامه ها، مجلات، مقالات و . . . با کیفیت های