ساخت يك نمايه ساز خودكار براي متون فارسي

مسعود تشكري
Tashakori@noavar.com

محمدرضا ميبدي
Meybodi@ce.aku.ac.ir

آزمايشگاه سيستمهاي نرم‌افزاري
دانشكده مهندسي كامپيوتر
دانشگاه صنعتي اميركبير

چكيده:

افزايش روزافزون حجم اطلاعات متني، توسعه تكنيكها و ابزارهاي كارآمدي را مي‌طلبد كه قادر به دستيابي به اطلاعات مورد نياز در مدت زمان مناسب باشند. در اين مقاله يك نمايه ساز خودكار براي متون فارسي ساخته شده است. براي آزمايش اين سيستم، ٤٥٠چكيده و ٣٢ پرس‌وجوي فارسي، كه در زمينه‌هاي تخصصي كامپيوتر هستند، جمع آوري شده است. بر روي متون جمع آوري شده، ابتدا شيوه توزيع واژگان مطالعه مي‌شود. سپس فهرستي از واژگاني كه در زبان فارسي، عمومي به شمار مي‌روند تهيه مي‌شود.

در بازيابي اطلاعات متني، استخراج خودكار ريشه واژگان از اهميت زيادي برخوردار است. از اين رو در اين مقاله سعي شده است با استفاده از قوانيني كه براي صرف واژگان در دستور زبان فارسي وجود دارد، يك ريشه‌ياب خودكار براي واژگان فارسي طراحي گردد. همچنين استثنائات موجود در قوانين صرف مورد بررسي قرار گرفته، و پيشنهادهايي نيز براي حل مشكل مرز بين واژه‌ها ارائه گرديده است. با استفاده از اين ريشه ياب خودكار، تغييرات كيفي سيستم در هنگام استفاده از ريشه يابي واژگان نيز مورد بررسي قرار مي‌گيرد. در نهايت سيستم نمايه ساز خودكار ارائه شده در اين مقاله، با استفاده از دو پارامتر "بازخواني" و "دقت" مورد ارزيابي قرار مي‌گيرد.

كليد واژه: نمايه سازي خودكار، بازيابي متون، ريشه يابي واژگان، پردازش خودكار متن، متون فارسي

١- مقدمه

امروزه حجم زياد و تقريباً نامحدود اطلاعات موجود باعث مي‌گردد تا استفاده از اطلاعات و مديريت آن با دشواريهايي روبرو باشد. با توجه به رشد و گسترش حجم اطلاعات و به موازات آن ضرورت به كارگيري و استفاده مؤثر از منابع اطلاعاتي، يكي از مهمترين و اساسي‌ترين نيازهاي موجود قابليت دستيابي به اطلاعات مورد نياز در مدت زمان مناسب است. در واقع جستجوي سريع و يافتن اطلاعات مورد نظر جستجوگر از اهميت فوق العاده‌اي برخوردار است .

با وجود آنكه اطلاعاتي كه امروزه عرضه مي‌شوند صورتهاي مختلفي از قبيل تصوير، صوت، انيميشن، و . . . به خود گرفته‌اند، هنوز هم پراستفاده‌ترين و حجيم‌ترين اطلاعات موجود، متون غير ساخت يافته هستند[١]. به طور مثال در بايگاني سازمانها اطلاعات متني زيادي درقالب نامه‌ها، جزوات، آئين نامه‌ها و دستورالعملها و قوانين(حقوقي، جزايي، مالياتي)، و ... وجود دارد؛ و يا در روزنامه‌ها و مجلات، كتابخانه‌ها، و محيطهاي وب[١] و پست الكترونيك[٢]، كه روزبروز نيز گسترش مي‌يابند، غالب اطلاعات به صورت متن مي‌باشد.

از زمان به كارگيري كامپيوتر با نگهداري متون به صورت الكترونيكي ، رشد قابل توجهي درجمع آوري اطلاعات و سازماندهي آنها به منظور بهره برداري هر چه بيشتر مشاهده مي‌گردد. اين نوع نگهداري باعث كاهش فضاي ذخيره سازي و هزينه نگهداري و نيز انعطاف بيشتر در كاربرد متنها شده است. از اين رو نگهداري الكترونيكي متون همواره مورد توجه مي‌باشد و تلاشهاي زيادي نيز براي بهبود سرعت و كيفيت روشهاي بازيابي چنين اطلاعاتي صورت مي‌پذيرد. منظور از بازيابي اطلاعات متني، فرايند جستجو و يافتن اين نوع اطلاعات در مجموعه بزرگي از متون است. سرعت و سهولت در جستجو و بازيابي اطلاعات نكته‌اي است كه

[١] WWW

[٢] E-Mail

همواره بايد مد نظر باشد تا از هزينه‌هاي غير ضروري و اتلاف وقت اجتناب شود. تكنيكهاي بازيابي اطلاعات به دليل ناتواني پايگاههاي داده‌اي سنتي در مديريت متون غير ساخت يافته، به شدت مورد توجه قرار گرفته‌اند[١].

يكي از مشكل‌ترين و مهمترين وظايف سيستم‌هاي بازيابي متن[١]، امكان جستجوي كارا روي اطلاعات متني مي‌باشد. در جستجو و بازيابي اطلاعات، يك نمايه[٢] مناسب در حقيقت نقش كليدي دارد. ساخت نمايه براي متنها، بازيابي آنها را براي محققان و خوانندگان آسان مي‌سازد. عمل ساخت نمايه در حال حاضر به دست نيروي انساني متخصص و ماهر انجام مي‌شود، كه اين روش دشوار، پرهزينه، و زمان‌بر است. اما در سيستم‌هاي بازيابي متن، چنين نمايه‌اي مي‌تواند كاملاً به صورت خودكار توليد شود. به عبارت ديگر يك سيستم بازيابي متن، مي‌تواند مجموعه اقداماتي را در جهت ساخت يك نمايه مناسب، مجهز، و كارآمد روي واژه‌هاي متن انجام دهد. پس از ساخت نمايه، سيستم مي‌تواند با استفاده از آن، در جواب به پرس‌وجوي[٣] كاربر، متوني را كه مربوط به واژه‌هاي مورد درخواست وي هستند را يافته و ارائه كند.

علاوه بر دشواري،كندي، و پرهزينه بودن نمايه سازي در شيوه دستي، اين شيوه مشكلات ديگري نيز دربردارد. در شيوه دستي هر فرد نمايه ساز، كه به خوبي هم آموزش ديده است، نسبت به فرد ديگر نمايه‌هاي متفاوتي، به يك متن مي‌دهد. حتي ديده شده است كه يك نمايه ساز نيز در زمانهاي مختلف نمايه‌هاي متفاوتي به يك متن مي‌دهد[٢]. اضافه بر اين هرچند افراد نمايه ساز درك بهتري از متن دارند، ولي ممكن است هنگامي كه با حجم زيادي از متون و واژه‌هاي نمايه سر و كار دارند، اشتباهات زيادتري نسبت به شيوه‌هاي خودكار نمايه سازي داشته باشند[٣]. با اين وجود روند نمايه سازي خودكار بهتر است تنها روي مجموعه‌هاي متني بزرگ به كار رود. در مورد مجموعه‌هاي كوچك، اين امكان وجود دارد كه روشهاي دستي سريعتر بوده و نتيجه كاراتري ارائه كنند[٤].

اصولاً هر متني داراي دو بخش است: بخش اول شامل اطلاعاتي است كه نسبت به محتواي متن خارجي محسوب مي‌شوند، مثل نام نويسنده، تاريخ و محل انتشار، و نام منتشر كننده؛ بخش دوم شامل محتويات اصلي متن است. در كتابخانه‌ها بخش اول با نـام دسته بندي توصيفي[٤] و بخش دوم با نام دسته بندي موضوعي[٥] شناخته مي‌شود. نمايه سازي به عمليات شـناخت موضـوع و محتـواي متن گفته مي‌شود؛ و هنگامي كه اين عمليات به كمك وسايل مدرن كامپيوتري صورت پذيرد، نمايه سازي خودكار ناميده مي‌شود[٤].

حساس‌ترين و مشكل‌ترين مرحله‌اي كه در روند نمايه سازي خودكار بايد طي شود، انتخاب واژه‌هايي است كه براي ساخت نمايه به كار مي‌روند. در عمل، نمايه سازي روي تمام واژه‌هاي متن داراي سربار بسيار زياد مي‌باشد، ضمن اينكه نمايه سازي روي تمام واژه‌ها يك كار غيرضروري است. كافي است تنها واژه‌هايي در نمايه به كار روند كه نشان‌دهنده محتواي متن مربوطه هستند، در واقع واژه‌ها يا مفاهيمي كه مورد علاقه كاربر بوده، توسط وي جستجو مي‌شوند. اولين و واضح ترين مكاني كه شناساننده‌هاي خوب محتوا ممكن است يافت شود، خود متن مستندات است[٤]. بنابراين به استخراج واژه‌هاي محتوايي از مستندات خواهيم پرداخت.

تحقيقات درباره ايجاد و يا بهبود روشهاي توليد خودكار نمايه و نيز جستجوي اطلاعات در متون براي زبانهاي مختلف همواره در جريان بوده است. ضرورت انجام تحقيق در مورد نمايه سازي خودكار متون فارسي، تفاوت نحوه ساخت و نيز به كارگيري واژگان در متون اين زبان است. به طور مثال در زبان فارسي بر خلاف زبان انگليسي افعال داراي شخص و زمان هستند؛ همچنين ضماير به واژه‌ها مي چسبند.

ادامه مقاله به صورت زير سازماندهي شده است: در بخش٢ در مورد معيارهاي ارزيابي كيفيت در سيتمهاي بازيابي متن بحث مي‌شود. در بخش٣ شيوه گرد آوري متون مورد نياز بانك متني اين مقاله بررسي مي‌شود. شيوه توزيع واژگان در متون فارسي و نيز تهيه واژه‌هاي عمومي زبان فارسي در بخش٤ توضيح داده مي‌شوند. سپس در بخش٥ در مورد شيوه پرس‌وجو و نيز نحوه گرد آوري پرس‌وجو براي ارزيابي بحث مي‌شود. در بخش٦ برخي شرايط و مفروضات نمايه سازي خودكار بررسي خواهدشد. در بخش٧ در مورد طراحي يك ريشه ياب خودكار براي واژگان فارسي مطالبي بيان مي‌شود. نتايج پياده سازي نمايه ساز خودكار ارائه شده در اين

[١] text retrieval

[٢] index

[٣] query

[٤] descriptive cataloging

[٥] subject cataloging

مقاله در بخش۸ مشاهده خواهد شد. در انتها نيز به نتيجه گيري پرداخته مي‌شود. در ضميمه نيز پرس‌وجوهاي جمع آوري شده براي ارزيابي سيستم آورده شده است.

۲- معيارهاي ارزيابي كيفيت در سيستمهاي بازيابي متن

براي ارزيابي كيفيت سيستمهاي بازيابي متن، تلاشهاي زيادي صورت مي‌پذيرد. دليل اول براي اين كار، مقايسه سيستمهاي موجود با يكديگر و نيز سيستمهاي خودكار با سيستمهاي دستي است. دليل دوم براي ارزيابي كيفيت اين گونه سيستمها اين است كه گاهي تغييرات خاصي براي بهبود سيستم، به اجزاء آن مي‌توان اعمال نمود؛ سنجش تأثير اين تغييرات بر روي سيستم، نياز به ارزيابي كيفيت سيستم دارد. در اين بخش روشهاي ارزيابي سيستمهاي بازيابي متن بررسي مي‌شود.

۲-۱- نوع واژه‌هاي به كار رفته در نمايه

براي تشخيص نوع واژه‌هاي به كار رفته در نمايه، دو پارامتر قابل تعريف است: پارامتر دربرگيري[۱]، بيانگر ميزاني است كه همه واژه‌هاي متن در نمايه ظاهر شده‌اند. در واقع هر چه واژه‌هاي بيشتري از متن در ساخت نمايه به كار روند، ميزان در برگيري نمايه و نيز نسبت موارد مربوطي كه مي‌تواند بازيابي شود بيشتر خواهد بود چرا كه جنبه‌هاي مختلف موضوعات مورد بحث تشخيص داده شده‌اند. پارامتر ديگر ميزان تعيين كنندگي[۲] نمايه است كه بيانگر سطح تعيين كنندگي واژه‌هاي به كار رفته در ساخت نمايه است. يعني در واقع هر واژه در نمايه تا چه حد دقيق، متنهاي مربوطه را مشخص مي‌كند. نمايه‌اي كه داراي سطح بالايي از تعيين كنندگي است، به واژه‌هاي به كار رفته در آن، موارد نامربوط را نگاشت نمي‌كند، و در چنين نمايه‌اي واژه‌ها دقيق و تعيين كننده‌اند[٤و٦].

معيارهاي ديگري نيز مانند زمان پاسخگويي به پرس‌و جو، شكل خروجي و سهولت استفاده از سيستم گاهي به معيارهاي ارزيابي اضافه مي‌شوند. اما كارايي سيستمهاي بازيابي عمدتاً با دو پارامتري كه در ذيل معرفي خواهند شد مقايسه مي‌گردد و در بسياري از موارد ارزيابي، تنها اين دو پارامتر اهميت دارند[٤و٦]. در اين مقاله نيز ارزيابي نتايج تنها با اين دو پارامتر انجام مي‌گيرد.

۲-۲- پارامترهاي اصلي ارزيابي

همانطور كه گفته شد براي تعيين كيفيت بازيابي در سيستمهاي بازيابي متن، پارامترهاي زيادي تعريف شده‌اند. مهمترين اين پارامترها به اين شرح مي‌باشند[٤]:

"بازخواني" [3]: اين پارامتر به بيان ساده، بيانگر قابليت سيستم براي ارائه موارد مربوط به درخواست كاربر مي‌باشد. در واقع اين پارامتر را مي‌توان به صورت زير تعريف كرد:

$$\text{"بازخواني"} = \frac{\text{تعداد موارد بازيابي شده و مربوط}}{\text{تعداد كل موارد مربوط درمتن}}$$

مقدار اين پارامتر همواره بين صفر و يك است و نسبت مستقيمي با ميزان دربرگيري واژه‌ها در نمايه دارد. يعني هر چه نمايه واژه‌هاي بيشتري را در خود داشته باشد، تعداد موارد بازيابي شده مربوط("بازخواني") بيشتر مي‌شود. از طرف ديگر ميزان اين پارامتر به مجموعه متون، مجموعه كاربران، و نحوه فرموله كردن پرس و جو توسط كاربران بستگي دارد[٤و٦].

"دقت" [٤]: اين پارامتر به بيان ساده بيانگر قابليت سيستم براي ارائه فقط موارد مربوطه به درخواست كاربر (يا به عبارت ديگر رد موارد نامربوط) مي‌باشد. در واقع اين پارامتر را نيز مي‌توان به صورت زير تعريف كرد:

$$\text{"دقت"} = \frac{\text{تعداد موارد بازيابي شده و مربوط}}{\text{تعداد كل موارد بازيابي شده}}$$

مقدار اين پارامتر نيز بين صفر و يك مي‌باشد و نسبت مستقيمي با ميزان تعيين كنندگي واژه‌ها در نمايه دارد. يعني هر چقدر واژه‌هاي به كار رفته در نمايه دقيقتر و تعيين‌كننده‌تر باشند، تعداد موارد نامربوط كمتري بازيابي مي‌شوند. از طرف ديگر ميزان اين پارامتر نيز به مجموعه متون، مجموعه كاربران، و نحوه فرموله كردن پرس وجو توسط كاربر هم بستگي دارد [٤و٦]. در حالتي كه

---

[۱] exhaustivity

[۲] specifity

[۳] recall

[٤] precision

تعداد كل موارد بازيابي شده صفر است، نمي‌توان مقدار "دقت" را با استفاده از فرمول بالا محاسبه نمود. منطقاً در اين حالت مقدار پارامتر "دقت" يك در نظر گرفته مي‌شود.

مقدار دو پارامتر "بازخواني" و "دقت" غالباً نسبت معكوس با هم دارند و بهبود يكي موجب افت ديگري مي‌شود. اما مـي‌تـوان در هر كاربردي مقدار بهينه‌اي براي هر دو پيدا نمود. اگر مقدار هر دو پارامتر در پرس و جوي اول بيشتر از پـرس و جـوي دوم باشـد؛ نتايج پرس و جوي اول بهتر خواهد بود. اما اگر در يك پرس و جو يكي از اين دو پارامتر كمتر و ديگري بيشتر باشد؛ قضاوت در مـورد اينكه كدام پرس و جو نتيجه بهتري داشته است، به نياز كاربر وابسته خواهد بود. بـا توجـه بـه اينكـه عمليـات ارزيـابي بـا اسـتفاده از مجموعه‌اي از پرس‌وجوها انجام مي‌شود، روشي به عنوان روش بهتر در نظر گرفته خواهد شد كه براي مجموعه پـرس‌وجوهـا ميـانگين بهتري داشته باشد؛ حتي اگر روش ديگر كه ميانگين بدتري دارد، براي بعضي از پرس‌وجوهاي خاص نتيجه بهتري داشته باشد.

۳ - گرد آوري متن

به طور كلي وجود يك واژه خاص در هر زمينه‌اي معناي خاص خود را دربردارد. اگر متون ورودي يك سيستم نمايه سازي به زمينه و موضوع خاصي مثل پزشكي، حقوق، كامپيوتر، ... محدود شود، واژه نامه سيستم تخصصي‌تر خواهد بود، كه در بهبود نتايج نقش بسزايي مي‌تواند داشته باشد. از اين رو براي متون فارسي ورودي در محيط آزمايش اين مقاله، شرط در حوزه كامپيوتر بودن قرار داده شد.

اولين نياز براي ارزيابي سيستم نمايه سازي خودكار ارائه شده در اين مقاله، تشكيل يك بانك متني مناسب متشكل از متون متنوع فارسي در زمينه تخصصي كامپيوتر است. طبيعتاً محتويات اين بانك بايد به شكلي در كامپيوتر ذخيره شود كه بتوان واژگان متن را به طور خودكار از فايل استخراج نمود. بنابراين به خاطر در دسترس نبودن فايلهايي كه از تصوير متن روي كاغذ[۱] ساخته شده باشند نمي‌توانند وارد بانك متني شوند. گردآوري اين بانك متني با چنين خصوصياتي كاري بس دشوار به نظر مي‌رسيد. چرا كه در اكثر كنفرانسها، همايشها، و كنگره‌هاي داخلي اگر هم متن مقالات به صورت كامپيوتري منتشر شده باشد، غالباً از روي تصوير متن چاپ شده بوده است. همچنين در دانشگاهها و مراكز تحقيقاتي نيز، تا قبل از كار بر روي اين مقاله، فايل متن پايان نامه‌ها، سمينارها، و مقالات تحقيقاتي جمع آوري نشده بود. اين مقوله مشكل بزرگي بر سر راه تحقيقات مختلف در زمينه آناليز متون فارسي به نظر مي‌رسد.

باتوجه به هزينه ذخيره سازي و نيز كمبود متون تايپ شده، استفاده از متن كامل در سيستمهاي بازيابي متن، دشوار به نظر مي‌رسد. در بسياري ازكاربردهاي عملي، استفاده از گلچين متون (نظير چكيده‌ها) كافي به نظر مي‌آيد. نتايج بررسي‌هاي موجود بيانگر اين است كه استفاده از چكيده‌ها در كارآمد بودن عمليات بازيابي نقش بسيار مؤثري دارد، هر چند استفاده از متن كامل مستندات، بهبود بسيار كمي را نسبت به چكيده‌ها در پي خواهد داشت[٤]. علاوه بر اين با توجه به تعداد متنها در سيستمهاي بازيابي متن، ارزيابي پرس‌وجوها بر روي متون غير كوتاه، كاري بسيار مشكل است. به همين علت در اين مقاله، نمايه سازي خودكار فقط بر روي متن چكيده‌ها بررسي مي‌گردد.

بانك متني جمع آوري شده جهت اين مقاله، كه اولين بانك در نوع خود به شمار مي‌رود، شامل ٤٥٠متن چكيده مقالات، سمينارها و نيز پايان نامه‌هاي دانشجويي است، كه همگي در زمينه كامپيوتر هستند و تا حد امكان با تلاش نگارندگان جمع آوري شده است. در بين اين ٤٥٠چكيده، ۲۲۰ چكيده به وسيله يك پايان نامه كارشناسي[٥] از مقالات كنفرانسهاي سالانه انجمن كامپيوتر ايران جمع آوري گرديده است. اين بانك متني از اين پس "مچك" (مـتون چـكيده كـامپيوتري) ناميده مي‌شود.

استفاده از نرم افزارهاي گوناگون در تايپ متون فارسي يكي از مسايل جدي بر سر راه استخراج خودكار واژگان متون فارسي به شمار مي‌رود. سيستم نمايه سازي خودكار اين مقاله، كه از اين پس از آن با عنوان "نمايه ساز" ذكر خواهد شد، در ابتدا فرمتهاي "زرنگار"، "PE2فارسي"، و "MS WORD تحت ويندوزهاي مختلف" را به يك فرمت داخلي تبديل[۲] مي‌نمايد. لازم به ياد آوري است كه در استخراج واژه‌هاي متن فرض مي‌شود كه واژه‌ها درست تايپ شده‌اند؛ و بنابراين خطاهاي تايپي احتمالي ناديده گرفته مي‌شود. براي اطمينان از صحت واژه‌ها، مي‌توان به طور جداگانه از يك بررسي كننده صحت واژه‌ها قبل از ورود متن به "نمايه ساز" استفاده نمود.

---

[۱] hard copy

[۲] convert

٤ - واژه‌هاي عمومي در زبان فارسي

با توجه به شيوه استفاده از واژگان زبان، بعضي واژه‌ها در همه متنها زياد تكرار مي‌شوند. اين گونه واژه‌ها، واژه‌هـاي عمـومي زبان[1] ناميده مي‌شوند. هنگامي كه واژه‌ها به ترتيب نزولي تعداد تكرار (واژه‌هاي با تكرار بالاتر در ابتدا) مرتب شـوند طبـق قـانونzipf رابطه زير برقرار است[٤و٦]:

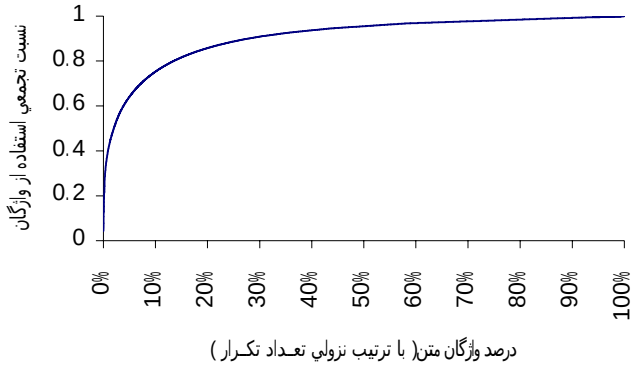
$$Frequency.rank \approx Constant$$

بر طبق نظريه لان، واژه‌هايي كه بيشترين تكرار را دارند (و كمترين رتبه) واژه‌هاي عمومي هستند كه ارزش استفاده از آنهـا كم است[٧] (در زبان فارسي واژه‌هايي مثل "از" ، "به" ، "و" ، ...). براي اولين بار قانونzipf بـر روي متـون مختلـف فارسـي در زمينـه كامپيوتر در[٩] بررسي شده است. در جدول١ تعداد تكرار و نيز ثابت zipf در "مجك" براي ٤٠واژه كه بالاترين تعداد رخداد را دارند نشان داده شده است. شكل١ نيز بيانگر توزيع تجمعي كاربرد واژه‌ها در متون مذكور مي‌باشد.

جدول١- تعداد تكرار و ثابتZipf براي ٤٠ واژه با بالاترين تعداد رخداد در "مجك"

(تعداد كل واژگان = ٧٠٤٢٧ ، تعداد متمايز واژگان = ٥٤٥٦)

رتبه(R)	واژه	تعداد رخداد(F)	R × F/N
1	و	3007	0.042697
2	در	2475	0.070286
3	از	2030	0.086473
4	به	1640	0.093146
5	اين	1619	0.114942
6	است	1338	0.11399
7	كه	1077	0.107047
8	شده	1053	0.119613
9	با	1020	0.130348
10	براي	960	0.136311
11	يك	846	0.132137
12	را	792	0.134948
13	استفاده	660	0.121828
14	سيستم	528	0.10496
15	هاي	525	0.111818
16	روش	505	0.114729
17	بر	448	0.10814
18	ميشود	420	0.107345
19	شبكه	394	0.106294
20	آن	391	0.111037
21	مورد	337	0.100487
22	ميباشد	331	0.103398
23	ارائه	311	0.101566
24	طراحي	309	0.105301
25	الگوريتم	295	0.104718
26	بررسي	288	0.106323
27	مقاله	286	0.109645
28	سازي	278	0.110526
29	انجام	272	0.112002
30	داده	227	0.096696
31	روي	218	0.095958
32	مدل	218	0.099053
33	قرار	218	0.102148
34	هر	208	0.100416
35	دو	206	0.102376
36	روشهاي	204	0.104278
37	آنها	201	0.105599
38	مختلف	195	0.105215
39	پردازش	180	0.099678
40	جهت	177	0.10053



شكل١- نسبت استفاده از واژگان در متون چكيده

[1] stop list, negative dictionary

همانطور كه در شكل۱ مشخص است، در "مچك"۲۰% از واژگان متمايز بيش از ۸۵% از كل واژگان را تشكيل داده‌اند. اشاره شد كه واژه‌هاي عمومي زبان در نمايه سازي ارزش كمي دارند. اين واژه‌ها ۳۰ تا۵۰ درصد متون را تشكيل مي‌دهند و داراي خاصيت تعيين كنندگي در مورد متنهايي كه در آنها ظاهر شده‌اند نيستند. در واقع اين واژه‌ها، واژه‌هايي مثل ضماير، قيود، حروف اضافه و ربط هستند كه كاربر مايل به جستجوي آنها نيست. بنابراين براي ساخت نمايه ابتدا بايد اين واژه‌ها از بين واژه‌هاي استخراج شده از متون حذف شوند. در هر زبان، يك مجموعه از اين واژه‌ها قابل شناسايي است. در زبان انگليسي در حدود ۲۵۰ واژه از اين نوع مي‌توان يافت، واژه‌هايي مانند 'a' ، 'the' ، و 'is'. به مجموعه اين واژه‌ها، معمولاً واژه‌هاي عمومي زبان(stop list يا negative dictionary) گفته مي‌شود[٤و٦]. در اين مقاله براي واژه‌هاي عمومي زبان فارسي، از يك فهرست ۱۵۰واژه‌اي تهيه شده است كه در جدول۲ نشان داده شده‌اند. بعد از حذف واژه‌هاي عمومي زبان، نوبت به نمايه سازي واژه‌هاي باقيمانده مي‌رسد.

جدول۲- فهرست واژه‌هاي عمومي در "نمايه ساز"

الان	اگرچه	اگر	است	از	آنها	آنگاه	آنكه	آنچه	آنجا	آنان	آن
باشد	با	اينها	اينگونه	اينست	اينجا	اين	ايشان	اي	او	انجام	اما
تايي	تا	پس	بين	بيشتر	به	بنابراين	بنابر	براي	بر	بدون	بايد
چنين	چندين	چندگانه	چندان	چند	چنانچه	چنان	چگونه	چگونگي	چرا	توسط	تو
خودش	خودتان	خودت	خود	خواهيم	خواهد	حتي	چيست	چيزي	چيز	چون	چه
ديگر	دو	درباره	در	داشته	دارد	داراي	داده	خويش	خودمان	خودم	خودشان
شود	شما	شده	شد	شايد	شامل	سپس	زيرا	روي	را	ديگري	ديگران
مابين	ما	ليكن	لكن	گرفته	گرچه	كه	كردن	كرد	كدام	فقط	صورت
ميكنند	ميكند	ميشوند	ميشود	ميدهد	ميتواند	ميتوان	ميباشد	مورد	من	مختلف	مانند
هرچه	هر	هايي	هاي	ها	نيستند	نيست	نيز	نياز	نوع	نميتوان	نظر
همديگر	همچون	همچنين	همانند	همانطور	همان	هم	هستيم	هستند	هست	هريك	هر
وگرنه	وجود	و	هيچگونه	هيچكدام	هيچ	هنوز	همين	هميشه	هميشگي	همواره	همه
						يكي	يكديگر	يك	يا	وي	ولي

۵ - گرد آوري پرس‌وجو

در بازيابي متن هدف زير مجموعه‌اي از متون است كه نسبت به پرس‌وجويي خاص مربوط به نظر مي‌رسند. به عبارت ديگر عمليات جستجو مقايسه‌اي بين پرس‌وجو و متون ذخيره شده انجام مي‌دهد[٦]. مدلهاي رياضي زيادي جهت عمليات در سيستمهاي بازيابي اطلاعات پيشنهاد شده است. رايجترين آنها مدل دو ارزشي[1] است كه جملات پرس‌وجوي دو ارزشي (متشكل از واژه‌هاي كليدي وارد شده توسط جستجوگر) را با مجموعه واژه‌هاي شناساننده محتواي متن مقايسه مي‌كند و متني بازيابي مي‌شود كه عبارت پرس‌وجو براي آن صحيح باشد.

عبارات پرس‌وجوي دو ارزشي شامل مجموعه‌اي از واژه‌ها است كه با عملگرهاي دو ارزشي عطفي ("و")، فصلي ("يا")، و نقض ("غيراز") به هم مرتبط شده‌اند. در واقع از "يا" جهت بيان ترادف، از "و" جهت ساخت عبارت، و از "غيراز" جهت محدود ساختن جستجو استفاده مي‌شود[۸]. مثلاً در پرس‌وجوي "موسيقي و كامپيوتر" ، متوني مورد نظر است كه راجع به هر دو واژه "موسيقي" و "كامپيوتر" مطلب دارند. اگر فرض كنيم كه **A** و **B** نشان دهنده مجموعه متنهايي باشند كه به ترتيب واژه‌هايA و B نمايه آنها هستند؛ بازيابي متن در مدل دو ارزشي به شرح زير خواهد بود[٤]:

- پرس‌وجوي"A يا B" مجموعه متنهاي (**A∪B**) را بازيابي خواهد نمود. به عبارت ديگر متنهايي كه واژهA يا واژهB و يا هر دو نمايه آنها باشد بازيابي مي‌شود.
- پرس‌وجوي"A و B" مجموعه متنهاي (**A∩B**) را بازيابي خواهد نمود؛ يعني متنهايي كه هم واژهA و هم واژهB نمايه آنهاست بازيابي مي‌گردد.
- پرس‌وجوي"A غيراز B" مجموعه متنهاي (**A - B**) ، يعني متنهايي كهA نمايه آنهاست ولي واژهB نمايه آنها نيست، را بازيابي خواهد كرد.

[1] Boolean

پردازش در پرس‌وجوهايي كه بيش از يك عملگر داشته باشند نيز به صورت بازگشتي انجام مي‌شود، كه با شروع از داخلي‌ترين جمله(پرانتز) هر بار يك جمله پردازش مي‌شود.

براي ارزيابي "نمايه ساز" نياز داريم تا افراد متخصص در علم كامپيوتر، خود اقدام به طرح پرس‌وجوهاي مختلف كنند. سپس همان افراد بايد تعيين نمايند كه كدام متنها نسبت به پرس‌وجوي آنها مربوط است. براي زبانهاي گوناگون (مانند انگليسي، عربي، آلماني، و ...) بانكهاي مختلفي با موضوعات متنوع ايجاد گرديده‌اند. اين بانكها همانند استانداردهاي مختلفي هستند، كه مقايسه نتايج الگوريتمهاي نمايه سازي مي‌تواند با استفاده از اين بانكها انجام شود. لكن براي زبان فارسي چنين مجموعه‌اي تهيه نشده بود.

بنابراين از ۳۲ كاربر (كه دانشجوي كامپيوتر بودند) خواسته شد تا با تخصص خود به طرح پرس‌وجو و ارزيابي متنها بپردازند. پرس‌وجوهاي جمع آوري شده و شماره متنهاي مربوط در ضميمه مشاهده مي‌شوند. اين پرس‌وجوها مي‌توانند تك يا چند واژه‌اي باشند. همچنين مي‌توان در صورت لزوم، از پرانتز و عملگرهاي منطقي در پرس‌وجوها استفاده كرد.

٦ - شرايط نمايه سازي خودكار

به طور كلي قبل از شروع ساخت نمايه، بايد به بعضي موارد مهم توجه داشت. به مهمترين اين موارد در ذيل اشاره مي‌شود.

در نمايه سازي دستي وسايل مختلفي وجود دارد تا به نمايه ساز در كنترل عمليات نمايه سازي كمك نمايد. واضح است كه اين گونه وسايل كه به شكل متون زبان طبيعي هستند به سادگي نمي‌توانند در يك سيستم نمايه سازي خودكار مورد استفاده قرار گيرند[٤].

نمايه سازي مي‌تواند به دو روش كنترل شده و كنترل نشده انجام شود. در روش نمايه سازي كنترل نشده، همه واژه‌هاي متن در ساخت نمايه به كار مي‌روند. اين كار معمولاً علاوه بر ايجاد سربار زياد، مي‌تواند باعث ايجاد كج فهمي و خطا در روند جستجوي كاربر شود. بنابراين روش نمايه سازي كنترل شده كه در آن واژه‌هاي نمايه كاملاً محدود هستند، غالباً مورد حمايت قرار مي‌گيرد. در روش كنترل شده صحت املاء واژه‌ها بررسي مي‌شود و نيز براي واژه‌هاي مترادف يك واژه واحد انتخاب مي‌شود. مثلاً واژه "نوسان" نماينده واژه‌هاي "ارتعاش"، "تموج"، "نوسان"، "تپش"، "چرخش"، و ... مي‌گردد. لكن استفاده از اين نوع نمايه باعث مي‌گردد تا اشخاص آموزش ديده‌اي براي فرموله كردن جملات پرس و جو مورد نياز باشد. از اين روش در ساخت نمايه به صورت دستي استفاده مي‌شود. شواهد نشان مي‌دهد كه در صورت استفاده از روش نمايه سازي كنترل نشده، نتايج نزديك به روش كنترل شده خواهد بود[٤و٦].

در نمايه سازي مي‌توان از واژه‌هاي ساده و يا مركب استفاده كرد. در روشpostcoordination از واژه‌هاي ساده براي ساخت نمايه استفاده مي‌شود، و در موقع جستجو كاربر مي‌تواند با تركيب واژه‌هاي ساده و بر اساس نمايه، موارد متني مورد نظرش را بيابد. اين روش بيشتر در سيستمهايي كه ساخت نمايه به صورت خودكار انجام مي‌شود، به كار مي‌رود. در روشprecoordination واژه‌هاي تركيبي هم در ساخت نمايه به كار مي‌روند. از اين روش بيشتر در سيستمهايي كه نمايه به صورت دستي ساخته مي‌شود، استفاده مي‌شود. مثلاً نمايه سازي متن حاضر تحت عنوان "نمايه سازي خودكار" استفاده از روش precoordination محسوب مي‌شود؛ در حالي كه نمايه سازي واژه‌ها به صورت مجزا در روشpostcoordination صورت مي‌پذيرد. در سيستمهاي نمايه سازي خودكار غالباً از واژه‌هاي مجزا استفاده مي‌شود؛ زيرا انتساب خودكار اين گونه از واژه‌ها به متون بهتر انجام مي‌شود[٤و٦].

٧ - طراحي يك ريشه‌ياب خودكار براي واژگان فارسي

در زبانهاي طبيعي از بين كل واژگان مي‌توان تعداد معدودي واژه يافت كه واژه‌هاي ريشه در نظر گرفته مي‌شوند، و بقيه واژگان از اين ريشه‌ها اشتقاق مي‌شوند. به عبارت ديگر هر واژه داراي ريشه‌اي است كه ايده و معنا و مفهوم اصلي آن واژه را در بر مي‌گيرد. اشتقاق يك واژه از ريشه اصلي سبب مي‌شود تا مفهوم اصلي واژه شكل بهتري به خود بگيرد، و يا واژه براي بروز نقش نحوي خود در جمله آماده شود. در زبان فارسي (و خانواده زبانهاي هند و اروپايي) عمليات اشتقاق و ساخت واژگان با تركيب واژه‌هاي ريشه و الحاق پسوند و پيشوندهاي مختلف صورت مي‌گيرد. هدف از ريشه‌يابي، زدودن الحاقات و يافتن جوهره اصلي واژه است. هر چند در واقعيت گاهي الحاقات واژه معناي آن را چنان تغيير مي‌دهند كه حذف آنها موجب از بين رفتن معناي اصلي مي‌شود[١٠].

در پردازش زبانهاي طبيعي و سيستمهاي بازيابي اطلاعات متني، به مواردي برمي‌خوريم كه بايد ريشه واژگان به صورت خودكار استخراج گردد. به همين منظور ريشه‌يابهاي خودكار متنوعي براي زبانهاي مختلف ساخته شده است؛ مثلاً براي زبان انگليسي چندين ريشه‌ياب خودكار ساخته شده است[١١ و١٢ و١٣ و١٤]. لكن تا جايي كه نگارندگان اين مقاله اطلاع دارند، يك ريشه‌ياب

كامل براي زبان فارسي تاكنون طراحي نشده است. ريشه‌ياب پيشنهادي، " بن" ناميده مي‌شود.

در عمليات ريشه‌يابي، معمولاً واژگان منفرد (و مستقل از متن) مد نظر قرار گرفته مي‌شود؛ هرچند در برخي از ريشه‌يابها، از نقش (نحو) واژگان در جمله نيز كمك گرفته مي‌شود. ريشه‌يابهايي كه از نقش واژگان نيز استفاده مي‌كنند، اگرچه از نتايج بهتري برخوردار هستند؛ ليكن به دليل بار سنگين محاسباتي با اقبال كمتري مواجه شده‌اند.

اولين قدم براي ريشه‌يابي واژگان يك متن، جدا كردن واژه‌ها از يكديگر است. به عبارت بهتر بايد مرز واژه‌ها در متن دقيقاً معين باشد؛ تا بتوانيم ريشه هر واژه را استخراج نمائيم. در اكثر زبانهاي طبيعي علامتهايي مثل فاصله خالي، نقطه، كاما، علامت پرسش، علامت تعجب، سمي كولن، و ... بيانگر مرز بين واژه‌هاست. اما هنگام حروفچيني متون فارسي گاهي بين واژكهاي يك واژه نيز فاصله خالي گذاشته مي‌شود. در اين مقاله پيشنهادهايي براي حل اين مشكل ارائه شده است.

اصولاً ريشه‌يابهاي خودكار را مي‌توان در دو گروه "جدولي"[1] و "حذف پيشوند و پسوند"[2] جاي داد. به دليل مزايائي كه در بخش بعدي گفته خواهد شد، روش "حذف پيشوند و پسوند" را براي "بن" برگزيديم. همچنين به جهت كاهش زمان ريشه‌يابي از نقش (نحو) واژگان در جمله صرف نظر نموده‌ايم. به عبارت ديگر در اين مقاله با بررسي شيوه صرف واژگان در زبان فارسي، قوانين ريشه‌يابي خودكار را استخراج نموده‌ايم. به اين ترتيب در "بن"، ضمن پايين بودن ميزان خطا، زمان ريشه‌يابي نيز بسيار اندك خواهد بود. علاوه بر اين با كنار گذاشتن روش جدولي، نياز به نيروي انساني متخصص به حداقل ممكن خواهد رسيد.

در اين بخش پس از بررسي انواع روشهاي ريشه‌يابي خودكار، به شناسايي روشهاي ريشه‌يابي از واژگان زبان فارسي مي‌پردازيم و سپس الگوريتمي براي ريشه‌يابي ارائه مي‌دهيم. شيوه‌هاي ارزيابي يك ريشه‌ياب و نتايج به دست آمده از "بن"، نيز در اين بخش بررسي خواهد شد.

۷-۱- انواع ريشه‌يابهاي خودكار

همانطور كه ذكر شد، هر واژه داراي ريشه است. ريشه‌يابي به عملياتي اطلاق مي‌شود كه نهايتاً ريشه يك واژه را مشخص مي‌نمايد. چنانچه اين عمليات به كمك كامپيوتر انجام شود به آن ريشه‌يابي خودكار گفته مي‌شود. ريشه‌يابهاي خودكار در يكي از دو گروه ريشه‌يابهاي "جدولي" و ريشه‌يابهاي "حذف پيشوند و پسوند" جاي مي‌گيرند[۱۵].

ريشه‌ياب جدولي: ساده‌ترين روشي كه براي ريشه‌يابي خودكار به نظر مي‌رسد، نگهداري ريشه هر واژه در يك جدول است. در اين روش با جستجوي واژه در اين جدول، ريشه واژه مشخص مي‌گردد. هر چند از اين روش مي‌توان نتايج خوبي گرفت؛ اما استفاده از اين روش مشكلاتي به همراه دارد. اولاً براي واژگان فارسي اين اطلاعات فعلاً وجود ندارد. ثانياً نگهداري اين جدول سربار زيادي براي سيستم خواهد داشت. اصولاً ذخيره سازي نتايج پردازش هنگامي مفيد است كه پردازشها تكراري و يا پر هزينه باشند[۱۵]. بنابراين بايد به دنبال روشهاي ديگري براي ريشه‌يابي بود.

گاهي از دو روش گوناگوني پسين[3] و n-gram با عنوان ريشه‌ياب ذكر مي‌شود؛ اما از آنجا كه در اين روشها ريشه حقيقي توليد نمي‌شود، عنوان ريشه‌ياب براي آنها كمي‌گمراه كننده است. بنابراين در مواردي كه يافتن ريشه حقيقي واژه لازم است؛ روشهاي گوناگوني پسين و n-gram جوابگو نيستند، و بايد به دنبال روشهايي مثل "حذف پيشوند و پسوند" بود.

ريشه‌ياب حذف پيشوند و پسوند: اين نوع ريشه‌يابها با حذف پيشوند و يا پسوند به ريشه واژه مي‌رسند. در اكثر موارد ريشه به دست آمده ريشه حقيقي است؛ ولي گاهي هم اين ريشه‌يابها مجبور به تغيير واژه به دست آمده مي‌شوند، تا ريشه حقيقي حاصل شود. الگوريتم اين ريشه‌يابها از تعدادي قوانين تشكيل مي‌شود كه با يافتن اولين قانون مناسب و ممكن براي حذف پيشوند و يا پسوند، آن قانون مورد استفاده قرار مي‌گيرد. اكثر ريشه‌يابهاي موجود از اين نوع، اقدام به زدودن بزرگترين دنباله ممكن از حروف واژه بر طبق قوانين مي‌نمايند. اين فرايند آنقدر ادامه مي‌يابد تا هيچ حرف ديگري نتواند زدوده شود. همانطور كه گفتيم، پس از حذف پيشوند و يا پسوند يك واژه ممكن است ريشه اصلي دقيقاً به دست نيايد. براي جبران اين امر از روش تغيير دوباره[4] استفاده مي‌شود[۱۵].

تغيير دوباره يك تغيير وابسته به زمينه به شمار مي‌رود. در اين تغيير با توجه به دنباله حروفي كه در واژه حاصله وجود دارد، طبق قواعدي بعضي از حروف واژه تبديل به حروف ديگري مي‌شوند. به طور نمونه در قاعده AxC→AyC كه A و C زمينه را

---

[1] table lookup
[2] affix remove
[3] successor variety
[4] recoding

نشان مي‌دهند، دنباله حروفx به دنباله حروفy تبديل مي‌شود. مثلاً اگر واژه‌اي به "گ" و پس از آن به "ان" ختم شود، پـس از حـذف "ان" به عنوان نشانه جمع، "گ" تبديل به "ه" مي‌شود.

همانطور كه ملاحظه شد بهترين نوع ريشه‌يابي كه مي‌توان پيشنهاد نمود، ريشه‌ياب "حذف پيشوند و پسوند" خواهد بود. بدين منظور به بررسي قواعد صرف در زبان فارسي و انواع پيشوندها و پسوندها در اين زبان خواهيم پرداخت.

۷-۲- واژگان فارسي

پرفسور حسابي در مقدمه[١٦] يك مقايسه بسيار جالب بين زبانهاي سامي (با واژگان قانونمند مانند زبان عربي) و زبانهاي هند و اروپايي (مانند فارسي) انجام داده است. آنچه مهم است تأثير پيشوند و پسوند در ساخت واژه در زبان فارسي مي‌باشد. وي بيان مي‌دارد كه زبانهاي هند و اروپايي داراي شمار كمي ريشه، در حدود ١٥٠٠ ريشه[با توجه به ورود اصطلاحات از زبانهاي ديگر، فهرست "بن" در حدود ٧٠٠٠ واژه دارد]، و داراي تقريباً ٢٥٠ پيشوند و ٦٠٠ پسوند هستند كه با اضافه نمودن آنها به اصل ريشه مي‌توان واژه‌هاي ديگري ساخت. جمع واژه‌هايي كه فقط از تركيب واژه‌ها با پيشوندها و پسوندها توليد مي‌شود (بدون در نظر گرفتن تركيب خود واژه‌ها و روش‌هاي ديگر توليد واژه) حدوداً ٢٢٦ ميليون واژه مي‌شود.

استادان دستور زبان فارسي واژگان اين زبان را به هفت گروه مختلف تجزيه نموده‌اند كه عبارتند از : اسم، صفت، ضمير، قيد، حرف، صوت، و فعل[١٧ و ١٨ و ١٩]. پس از بررسي متوجه خواهيم شد كه غير از حرف، هريك از واژگان اين گروهها قبل از آنكه وارد جمله شوند و مورد استفاده قرار گيرند، پيشوندها، و يا پسوندهايي به ريشه اصلي آنها مي‌تواند اضافه شود تا شكل مناسب خود را پيدا كنند. به طور مثال براي جمع بستن اسامي پسوندهايي به اسم اضافه مي‌شود، و يا جهت توليد فعل براي زمان و شخص موردنظر، تغييراتي در مصدر فعل صورت مي‌پذيرد. نگارندگان مقاله به اين موضوع پي بردند كه واژگاني كه صفت، قيد، و يا صوت هستند، با صرف نظر كردن از گروهي كه در آن قرار دارند، از لحاظ ريشه يابي كاملاً در گروه اسم يا گروه فعل مي‌توانند قرار گيرند. بنابراين كافي است تا قوانين ريشه‌يابي تنها براي سه گروه اسم، فعل، و ضمير بررسي شود.

براي استخراج قوانين ريشه‌يابي، بايد به بررسي دستور زبان فارسي و شيوه ساخت واژگان در اين زبان بپردازيم. اما متأسفانه يك قانون كلي براي ساخت واژگان در زبان فارسي وجود ندارد و براي هر مورد استثناهاي زيادي يافت مي‌شود كه كار را مشكل مي‌كند. بدين ترتيب هر قانوني كه بخواهيم براي ريشه‌يابي به كار ببريم، به راحتي بي‌قاعدگيها و استثنائات فراواني براي آن قانون پيدا مي‌كنيم. شيوه ساخت واژگان در سه گروه اسم، فعل، و ضمير كه براي استخراج قوانين ريشه‌يابي بايد مورد توجه قرار گيرد، در ذيل بررسي مي‌شود.

۷-۲-۱- فعل

مراد از مشتقات فعل كليه ساختهايي است كه از بن ماضي و بن مضارع ساخته مي‌شوند، اعم از فعل، اسم، و صفت. شيوه ساخت اسامي و صفاتي كه از بن ماضي و بن مضارع ساخته مي‌شوند را مي‌توان در [١٧ و ١٨ و ١٩] مشاهده نمود.

ساخت افعال فارسي: در فعلهاي فارسي پس از حذف "ن" از انتهاي مصدر، بن ماضي باقي مي‌ماند. از طرف ديگر نيازي نيست كه براي هر مصدر بن مضارع آن نيز نگهداري شود؛ چون مي‌توان در اكثر افعال (كه جزء استثنائات نيستند) از بن ماضي به بن مضارع و يا بالعكس رسيد كه اين كار حجم فضاي مورد نياز براي نگهداري فهرست مصادر را كاهش خواهد داد. وابسته به نوع تغييري كه براي ساخت بن مضارع روي بن ماضي انجام مي‌شود، افعال را مي‌توان در هشت گروه جاي داد[١٧ و ١٨]. جدول١ تغييرات فعلها را در گروههاي هشت‌گانه نشان مي‌دهد. در هر يك از گروهها استثناهايي وجود دارد.

جدول ١- تغييرات افعال در گروههاي هشت گانه

شماره گروه	حروف پاياني	بن مضارع
١	يدن	حذف "يدن" از انتهاي مصدر
٢	دن	حذف "دن" از انتهاي مصدر
٣	ودن	حذف "دن" از انتهاي مصدر و تبديل "و" پاياني به "ا"
٤	ادن	حذف "ادن" از انتهاي مصدر
٥	ختن	حذف "تن" از انتهاي مصدر و تبديل "خ" پاياني به "ز"

٦	ستن	حذف "ستن" از انتهاي مصدر
٧	شتن	حذف "تن" از انتهاي مصدر و تبديل "ش" پاياني به "ر"
٨	فتن	حذف "تن" از انتهاي مصدر و تبديل "ف" پاياني به "ب"

شيوه ساخت زمان فعل و شيوه ساخت شخص(شناسه) در افعال فارسي در [١٧ و ١٨ و ١٩] ذكر شده است.

ساختمان فعل: افعال را در زبان فارسي از جهت ساختمان به چهار گروه مي‌توان تقسيم كرد: فعلهاي ساده، فعلهاي پيشوندي، فعلهاي مركب، و عبارتهاي فعلي[١٧ و ١٨]. حال به اختصار هر يك از اين گروهها را بررسي مي‌نماييم.

فعلهاي ساده: فعلهاي ساده افعالي هستند كه مصدر آنها بيش از يك واژه نيست. به عبارت ديگر اين افعال داراي اجزايي نيستند كه بتوان آنها را جدا كرد و در تركيب با واژه‌اي ديگر به‌كار برد. جهت ريشه‌يابي خودكار افعال ساده، تنها فهرست مصادر آنها را نگهداري مي‌نمائيم. در ساخت فعل، گاهي از افعال معين استفاده مي‌شود. با توجه به اينكه معمولاً قبل و بعد از افعال معين، يك فاصله خالي گذاشته مي‌شود؛ بايد اين مسئله را در طراحي الگوريتم ريشه‌يابي خودكار در نظر داشت.

فعلهاي پيشوندي: افعال پيشوندي از يك پيشوند و يك فعل ساده ساخته مي‌شوند. مهمترين پيشوندهايي كه فعل و يا مصدر پيشوندي مي‌سازند عبارتند از: "باز"، "بر"، "در"، "فرا"، "فرو"، "وا"، و "ور"[١٧ و ١٨]. نگارندگان مقاله اعتقاد دارند كه علي‌الظاهر جزء پيشوندي در اينگونه افعال بايد به صورت چسبيده و متصل به فعل ساده نوشته شود. همچنين هرچند گاهي فعل پيشوندي با فعل ساده سازنده آن تشابه معنايي دارد؛ لكن اين امر كليت ندارد (به طور مثال در مورد مصدر "برخوردن") و بهتر است كه براي ريشه‌يابي خودكار، مصادر افعال پيشوندي نيز اضافه بر مصادر افعال ساده وارد فهرست مصادر گردد. در مواردي كه بخواهيم ريشه فعل پيشوندي، همان مصدر فعل ساده سازنده آن باشد، مي‌توان از رابطه ترادف بين مصادر استفاده نمود.

شيوه ساخت فعل از مصادر پيشوندي نيز تفاوتهايي دارد. به طور كلي حروفي كه معمولاً به ابتداي ساخت فعل ساده افزوده مي‌شوند، در اين گونه افعال، بعد از پيشوند و در ابتداي فعل ساده ظاهر مي‌گردند. "بن" با اين فرض كه بين اجزاء اين گونه افعال هيچ فاصله‌اي وجود ندارد، عمليات ريشه‌گيري را انجام مي‌دهد.

فعلهاي مركب: افعال مركب از يك واژه (جزء غير صرفي) به همراه يك فعل ساده (جزء صرفي) ساخته مي‌شوند و مجموعاً معني واحدي را مي‌رسانند. در فعلهاي مركب جزء غير صرفي معمولاً اسم يا صفت يا بن فعل است. مصدرهاي ساده مانند: "كردن"، "آوردن"، "ساختن"، "خوردن"، و ... را كه فعل مركب مي‌سازند، "همكرد" مي‌گويند. مهمترين همكردهاي فارسي عبارتند از[١٧]:

آمدن، آوردن، بردن، بستن، پيوستن، خواستن، خوردن، دادن، داشتن، ديدن، رفتن، زدن، ساختن، شدن، فرمودن، كردن، كشيدن، گرديدن، گرفتن، گشتن، نمودن، نهادن، يافتن.

از آنجايي كه در افعال مركب معمولاً بين جزء صرفي و جزء غير صرفي يك فاصله خالي وجود دارد، از اين رو بايد در "بن"، مصدر كليه افعال مركب نيز در فهرست مصادر نگهداري شود. با توجه به اينكه افعال مركب بسيار گسترده و فراوان هستند، تهيه فهرست كامل آنها كمي مشكل است؛ همچنين بر سر مركب بودن و مركب نبودن بسياري از واژگان هنوز بين زبانشناسان اختلاف رأي وجود دارد [٢٠].

عبارتهاي فعلي: عبارت فعلي به دسته‌اي از واژگان اطلاق مي‌شود كه از مجموع آنها معني فعل واحدي حاصل مي‌شود[١٧]. عبارتهاي فعلي سه واژه دارند كه معمولاً يكي از واژگان حرف اضافه است و مجموع عبارت نيز معني مجازي دارد؛ مانند: "از چشم افتادن". بنابراين بايد براي ريشه‌يابي خودكار، مصدر تمامي عبارتهاي فعلي در فهرست مصادر وارد شود.

٧-٢-٢- اسم

به اسمي كه از تركيب بن فعل و يك يا چند حرف ديگر ساخته شود اسم مشتق مي‌گويند. نحوه ساخت اسامي مشتق در بخش فعل بررسي شد.

در زبان فارسي قوانين مختلفي براي جمع بستن اسم وجود دارد. در فارسي براي جمع بستن اسم در فارسي از دو علامت "ان" و "ها" استفاده مي‌شود. اگر اسمي به حرف "ه" ختم شود هنگام جمع بستن با "ان" حرف "ه" تبديل به "گ" مي‌شود. گاهي "ان" نشانه جمع نيست و جزيي از خود واژه است. علاوه بر اين بعضي اسامي كه از زبان عربي وارد زبان فارسي شده‌اند به صورت ديگري جمع بسته مي‌شوند كه به آن جمع مكسر مي‌گويند و داراي قانوني در زبان فارسي نمي‌باشد.

همچنين واژگان عربي و نيز بعضي واژگان فارسي با نشانه‌هاي جمع عربي يعني "ون"، "ين"، "ات" جمع بسته مي‌شوند. در اين مورد اسامي كه مختوم به "ا"، "ه"، "ي"، "و" باشند به جاي علامت "ات" با "جات" جمع بسته مي‌شوند[۱۷]. به آخر بعضي اسامي نيز براي رساندن مفهوم خردي و كوچكي "چه" و يا "ك" اضافه مي‌شود تا اسم مصغر ساخته شود.

وندها: مراد از وندها كليه پيشوندها و پسوندها مي‌باشد، كه به اسم يا صفت يا ... افزوده مي‌شوند. كاربرد پسوندها و پيشوندهاي معروف در [۱۷ و ۱۸ و ۱۹] ذكر شده است.

۷-۲-۳- ضمير

ضمايرر بر دو نوع منفصل و متصل مي‌باشند. جهت پيدا كردن ريشه واژه لازم است تا ضمايرر متصل در صورت وجود شناسايي شوند و از اصل واژه جدا گردند. ضمايرر متصل كه به انتهاي واژه مي‌چسبند در [۱۷ و ۱۸ و ۱۹] مشخص شده است.

اگر ضمايرر متصل، به آخر واژگان مختوم به حرف "ا" يا حرف "و" بچسبند، بين آنها و واژه حرف "ي" اضافه مي‌گردد. همچنين اگر ضمايرر مفرد به واژه مختوم به "ه" اضافه شوند، ميان آنها و واژه حرف "ا" اضافه مي‌شود[۱۷ و ۱۸].

۷-۳- الگوريتم ريشه يابي

قبل از آنكه به الگوريتم "بن" بپردازيم، بايد مشكل مرز بين واژگان را حل كنيم. امروزه تمام سيستمهاي كامپيوتري ذخيره و بازيابي اطلاعات، با زبان فارسي از نظر فاصله ميان واژه‌ها و عبارتها مشكل دارند. در صورتي كه خط فارسي صدها سال است كه توانسته ارتباط ميان افراد را برقرار كند و هيچ مشكل غير قابل حلي هم نداشته باشد. مشكل از جايي شروع مي‌شود كه پاي فناوري نوين به ميان مي‌آيد. بنابراين مي‌توان نتيجه گرفت كه راه حل اين مشكل را يا بايد در رسم‌الخط فارسي و يا در فناوري نوين يافت.

يكي از منطقي‌ترين راه حلها براي رفع اين مشكل آن است كه منتظر باشيم فرهنگستان زبان و ادب فارسي رسم‌الخط دقيق، منضبط، و مدون فارسي را ارائه كند. اگر چنين رسم‌الخطي تدوين گردد؛ حتماً مشكل فاصله ميان واژه‌ها و عبارتها نيز حل خواهد شد. در آن زمان احتمالاً همه يكسان عمل خواهند كرد و مشكلات رسم‌الخط فارسي كاهش خواهد يافت. ويژگي عمده اين راه حل آن است كه هم سيستمها و هم كاربران از يك رسم‌الخط تبعيت مي‌كنند و ميزان ناهماهنگي در پايين‌ترين حد خواهد بود. اين راه حل يك مشكل اساسي دارد؛ يعني معلوم نيست چه زماني رسم‌الخط يكسان شده نهايي تدوين خواهد شد. در مورد پيوسته نويسي يا جدا نويسي دهها سال است كه بحث و گفتگو وجود دارد و هنوز به نتيجه قطعي نرسيده است. حتي اگر در زماني خاص نتيجه نهايي به دست آيد و رسم‌الخط دقيق فارسي مشخص شود؛ براي اينكه همه آن را اجرا كنند باز هم به ساليان بسياري وقت نياز است[۲۰].

راه حل ديگري نيز براي حل مشكل رسم‌الخط فارسي پيشنهاد شده است. در اين راه حل نرم‌افزارها به گونه‌اي طراحي مي‌شوند كه هرگاه وارد كننده اطلاعات به بود يا نبود فضاي خالي ميان واژه‌ها شك داشته باشد با زدن علامتي ميان آن واژه‌ها، ورود اطلاعات را ادامه دهد[۲۱]. اما به نظر نمي رسد چنين پيشنهادي بتواند مشكل مرز بين واژه‌ها را حل نمايد. با تمام اين احوال به اعتقاد نگارندگان مقاله، به‌كارگيري پيشنهادهاي زير معقول به نظر مي‌رسد:

۱- قبل و بعد از هر واژه فاصله خالي درج گردد.

۲- بين حروف اصطلاحات و واژه‌هاي لاتين كه دقيقاً منعكس كننده لفظ خارجي است فاصله خالي درج نشود.

۳- انواع پيشوندها و پسوندها پيوسته يا بدون فاصله خالي نوشته شوند. اگر نياز باشد تا پسوند يا پيشوند به واژه اصلي كاملاً متصل نشود (مثلاً پسوند "اند" هنگام اتصال به واژه "شده"، به جاي "شده‌اند" به صورت "شده‌اند" درآيد.)؛ بايد از كاراكترهاي كنترلي موجود در نرم‌افزارهاي حروفچيني استفاده نمود و از درج فاصله خالي اجتناب كرد.

همانطور كه در بخش۷-۳ اشاره شد، هر قانوني كه براي ريشه‌يابي به كار رود، به راحتي بي‌قاعدگيها و استثنائات فراواني براي آن قانون مي‌توان يافت. براي حل چنين مشكلي دو راه حل به نظر مي‌رسد. يكي اينكه فهرستي از استثنائات نگهداري شود، ديگر اينكه از يك فهرست براي تشخيص واژه‌هاي درست استفاده شود. با توجه به فراواني استثنائات در قوانين ريشه‌يابي، راه حل دوم در عمل مناسبتر به نظر مي‌رسد.

اكنون به الگوريتم "بن" مي‌پردازيم. در اين الگوريتم ابتدا واژه در فهرست واژگان كه فقط اسامي جامد را در بر مي‌گيرد و از قبل آماده شده است جستجو مي‌شود. در صورت وجود واژه در فهرست واژگان، عمليات ريشه‌يابي انجام نمي‌شود و خود واژه به عنوان ريشه اصلي در نظر گرفته مي‌شود. اما در صورتي كه واژه در فهرست واژگان موجود نباشد؛ تا حد ممكن الگوريتم ريشه‌يابي روي آن

واژه اجرا مي‌گردد. چنانچه واژه حاصل شده در هر مرحله از الگوريتم، در فهرست واژگان موجود بود، عمليات مورد قبول واقع مي‌گردد و كار به پايان مي‌رسد.

در طراحي ريشه‌ياب سعي بر آن بوده است كه ضمن حفظ انسجام، اجزاء مستقل از يكديگر باشند؛ بنابراين سيستم از چندين بخش تشكيل شده است. اين بخشها عبارتند از:

الف) مجموعه قوانين ريشه‌يابي (براي زدودن پسوند و پيشوند) كه با استفاده از قوانين ساخت واژه در زبان فارسي استخراج شده‌اند.

ب) فهرست مصادر فارسي (اعم از افعال ساده، پيشوندي، و مركب و نيز عبارتهاي فعلي)

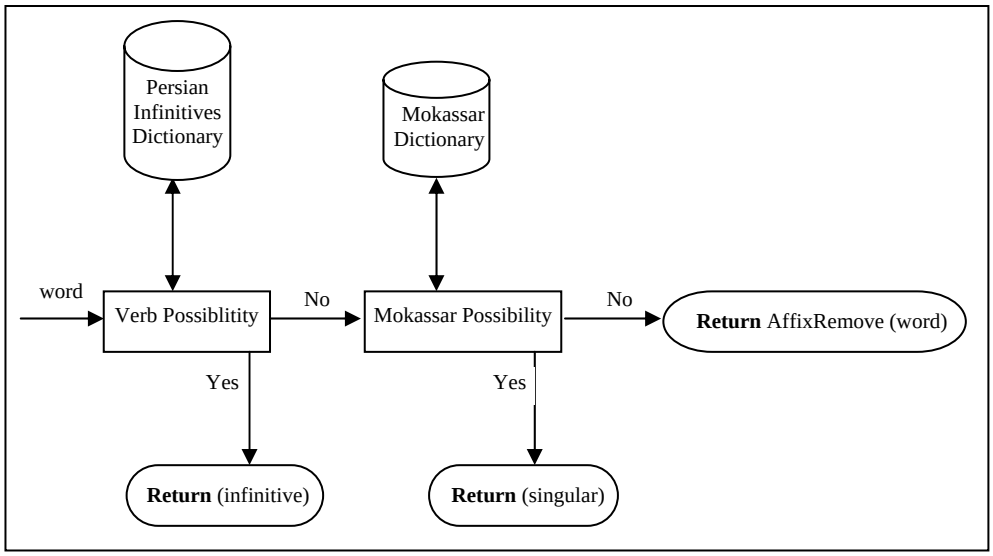
ج) فهرست واژگان جمع مكسر كه براي هر واژه جمع مكسر، مفرد آن را مشخص مي‌كند.

د) فهرست واژگان جامد كه كليه واژگان فارسي به غير از "افعال و مشتقات آن، جمعهاي مكسر و نيز مابقي واژگان مشتق" را نگهداري مي‌نمايد. به دليل اينكه چنين فهرستي براي زبان فارسي موجود نبوده است؛ به طور موقت يك فهرست آزمايشي ساخته شد. اين فهرست آزمايشي را، كه اكنون حدود ۹۰۰۰ واژه را دربردارد، مي‌توان به مرور تكميل‌تر نمود. براي تهيه فهرست اوليه، ابتدا واژه‌هاي موجود در ٤٥٠ چكيده از حوزه كامپيوتر استخراج گرديد. سپس با به كارگيري قوانين (خودكار) ريشه‌يابي، واژه‌هايي كه مشتق واژه‌هاي ديگر بودند از فهرست واژگان حذف شدند. پس از آن هنگام آزمايش ريشه‌ياب به تدريج واژه‌هاي جامدي كه در فهرست موجود نبود را به آن اضافه كرديم، و اگر واژه مشتقي نيز در فهرست باقيمانده بود، آن را حذف نموديم.

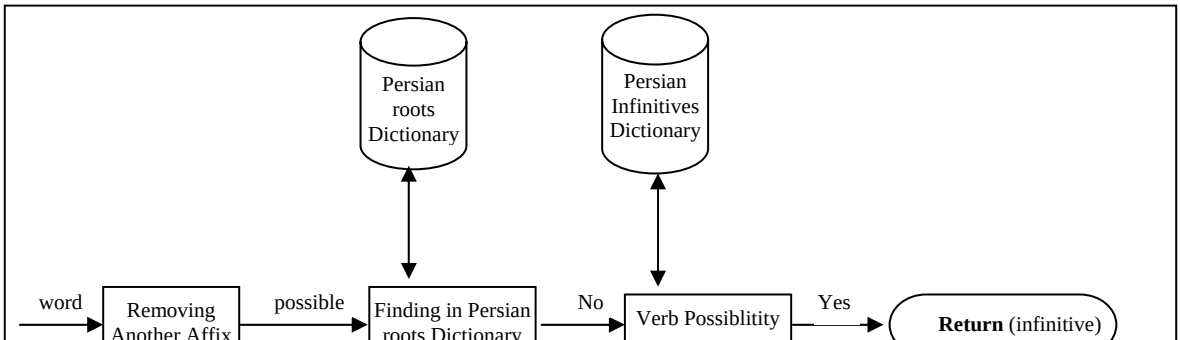
با مشخص شدن بخشهاي تشكيل دهنده ريشه‌ياب، نگارندگان مقاله از الگوريتمي كه در شكل۲ مشخص شده است استفاده كرده‌اند. در اين الگوريتم براي يافتن ريشه كلمه *word* ، *Stem(word)* فراخواني مي‌شود.

براي آنكه بتوان ريشه واژگاني كه از زبان عربي وارد فارسي شده‌اند را استخراج نماييم، دو روش به نظر مي‌رسد. يكي اينكه قواعد و ريشه‌هاي واژگان زبان عربي را نيز، هرچند محدود، در سيستم وارد كنيم؛ و يا اينكه با استفاده از يك فرهنگ ترادف تمامي واژگان هم خانواده را فهرست نماييم. در "بن"، راه حل اول به كار گرفته شده است.

Stem(word):



AffixRemove(word):



شكل۲- الگوريتم بن

پس از ساخت يك ريشه‌ياب خودكار، آگاهي از ميزان صحت نتايج آن، نخستين مطلبي است كه به ذهـن خطـور مـي‌كنـد. بنابراين به ارزيابي "بن" خواهيم پرداخت.

۷ - ٤ - ارزيابي ريشه‌ياب

به منظور آگاهي از قدرت يك ريشه‌ياب، بايد نتايج آن مورد ارزيابي قرار گيرد. هنگام آزمايش هر ريشه‌ياب روي نمونه‌اي از واژه‌ها كه از متون طبيعي گرفته شده‌اند، مي‌توان ارزيابي را بر پايه تعداد كل واژه‌ها و يا بر پايه تعداد واژه‌هاي متمايز قرار داد. در ايـن مقاله نتايج هر دو شيوه بررسي خواهد شد. ارزيابي نتايج ريشه‌يابها به كمك يكي از سه روش زير انجام مي‌گيرد:

۱- در روش اول، خطاهاي ريشه‌يابي به دو نوع خطاهاي تفريطي[1] و خطاهاي افراطي[2] تقسيم مي‌شوند. واژه‌هـاي هـم‌ريـشه كه در حقيقت ريشه و مفهوم يكساني دارند، خطاي تفريطي به شمار مي‌روند؛ اگر در عمل داراي يك ريشه واحد نباشند. واژه‌هايي نيز خطاي افراطي به حساب مي‌آيند؛ كه در حقيقت ريشه و مفهوم متفاوتي دارند، لكـن در عمـل داراي يـك ريـشه واحـد مـي‌شـوند. در طراحي الگوريتمهاي ريشه‌يابي، بايد تعادلي بين اين دو خطا برقرار شود. يـك ريـشه‌يـاب سـبك[3] از بـروز خطاهـاي افراطـي اجتنـاب مي‌كند؛ لكن توجهي به خطاهاي تفريطي نخواهد داشت. برعكس يك ريشه‌ياب سنگين[4] از بروز خطاهاي تفريطي جلوگيري مـي‌كنـد؛ هرچند موجب بروز خطاهاي افراطي مي‌گردد[۱۵].

مهمترين مشكل اين روش دسته بندي واژه‌هاي هم‌ريشه حقيقي و سپس قضاوت در مورد واژه‌هاي به دست آمده از ريشه‌ياب است. بنابراين بهتر است روشهاي ديگر ارزيابي نيز بررسي گردد.

۲- در روش دوم، از ميزان فشردگي فهرست ريشه‌هاي به دست آمده نسبت به فهرست واژه‌هاي اوليـه، بـه عنـوان واحـدي براي اندازه‌گيري قدرت بين دو ريشه‌ياب استفاده مي‌شود[۲۲]. به دليل اينكه ريشه‌ياب ديگري بـراي واژگـان زبـان فارسـي پيـشنهاد نشده است، اين روش ارزيابي را نيز كنار مي‌گذاريم.

۳- در روش سوم، ميزان صحت[5] ريشه‌هاي حاصل از ريشه‌ياب، جهت ارزيابي قدرت آن مورد بررسي قرار مي‌گيرد[۱۵]. بـا مناسب به نظر رسيدن اين روش، به كمك ده چكيده مقاله، كه از ده مقالـه اول مجموعـه مقـالات سـومين كنفـرانس سـالانه انجمـن كامپيوتر ايران(۱۳۷٦) انتخاب شده‌اند، ميزان صحت ريشه‌هاي به دست آمده از "بن" بررسي گرديد. نتايج در جدول۲ آمده است:

[1] understemming
[2] overstemming
[3] light stemmer
[4] heavy stemmer
[5] correctness

جدول ۲- نتايج حاصل از "بن"

	كل واژه‌ها	واژه‌هاي متمايز
تعداد واژه‌ها	۱۱٤۱	٤۱۳
تعداد ريشه‌هاي صحيح	۱۱۳۸	٤۱۰
درصد ريشه‌هاي صحيح	%۹۹/۷٤	%۹۹/۲۷

همانطور كه مشاهده مي‌شود "بن"، نتيجه بسيار خوبي داشته است. اجراي اين ريشه‌ياب، بر روي مجموعه فوق‌الذكر با يك كامپيوترPIII 500 Mhz با 64Mbحافظه، فقط ۱۲ ثانيه به طول انجاميد كه از لحاظ زماني نيز جالب توجه است. به عبارت ديگر زمان ريشه‌يابي هر واژه بر روي چنين كامپيوتري به طور متوسط تنها در حدود ۰/۰۳ ثانيه بود.

۸- نتايج

"نمايه ساز" پس از استخراج واژه‌هاي متن، واژه‌هاي عمومي را از بين آنها حذف مي‌كند. سپس روند نمايه سازي بر روي مجموعه‌اي از ٤۵۰چكيده در زمينه كامپيوتر با ۳۲پرس‌وجوي مختلف آزمايش گرديد. نتايج بازيابي متن در سيستم "نمايه ساز"، با استفاده از پرس‌وجوهاي ضميمه و بدون استفاده از ريشه يابي واژگان، در جدول۳ مشخص شده است.

در بازيابي متن استفاده از ريشه يابي واژگان، ميزان پارامتر "بازخواني" را افزايش مي‌دهد، هرچند پارامتر "دقت" اندكي كاهش مي‌يابد. ارزيابي "نمايه ساز" بر روي پرس‌وجوهاي ضميمه و با استفاده از ريشه يابي واژگان، در جدول۳ آمده است.

جدول۳- نتايج بازيابي متن در "نمايه ساز"

	"بازخواني"	"دقت"
بدون ريشه‌يابي واژگان	0.3595258	0.8974702
با ريشه‌يابي واژگان	0.5421372	0.8397220

در هنگام عدم استفاده از ريشه يابي واژگان، پس از حذف واژگان عمومي، حداكثر تعداد رخداد براي يـك واژه در "مچـك"، ٦٦۰بار مي‌باشد. پس از بررسي مشخص گرديد كه اگر واژه‌هايي كه تعداد رخداد آنها در مجموعه فوق الذكر بيش از ۲۰بار است حذف شوند، هيچ تغييري در نتايج به وجود نمي‌آيد؛ و فقط حجم نمايه ساخته شده بر روي مجموعه، كاهش مي‌يابد. اما اگـر واژه‌هـايي كـه تعداد رخداد آنها ۲۰ و يا كمتر از ۲۰بار است حذف شوند، به ازاي كاهش "بازخواني"، "دقت" در سيـستم افـزايش مـي‌يابـد. در هنگـام استفاده از ريشه يابي واژگان نيز، پس از حذف واژگان عمومي، حداكثر تعداد رخداد براي يك واژه، ۱۰۷۲بار مي‌باشد. اين بار نيز پـس از بررسي مشخص گرديد كه اگر واژه‌هايي كه تعداد رخداد آنها در "مچك" بيش از ۲۵بار است حذف شوند، هيچ تغييري در نتـايج بـه وجود نمي‌آيد. اما به طور كلي انتخاب مقدار تعداد رخداد براي حذف واژه‌هاي با تعداد رخدا بالا براي هر مجموعه‌اي متفاوت است.

۹- نتيجه گيري

در اين مقاله براي اولين بار يك محيط آزمايشي مناسب براي ارزيابي سيستمهاي نمايه سازي خودكار فارسي گردآوري شد. اين محيط آزمايشي، كه "مچك" ناميده مي‌شود، متشكل از ٤۵۰چكيده و ۳۲پرس‌وجوي مختلف فارسي در زمينه كامپيوتر است. به همراه پرس‌وجوها متون مربوط به آنها نيز مشخص شده است. همچنين با بررسي تكرار واژه‌ها در "مچك" نشان داده شد كه ۲۰% از واژگان متمايز استفاده شده، بيش از ۸۵% مجموع واژگان نامتمايز متنها را تشكيل مي‌دهند. بنابراين مي‌توان گفت كه بخش زيادي از واژگان متون را واژه‌هاي عمومي تشكيل مي‌دهند. از آنجا كه روند نمايه سازي خودكار، پس از شناسايي تك‌تك واژگان متن با حذف واژگان عمومي زبان از فهرست واژگان متن آغاز مي‌شود. در اين مقاله براي زبان فارسي يك فهرست ۱۵۰واژه‌اي از واژگان عمومي اين زبان تهيه گرديد. نتايج به دست آمده از "نمايه ساز" بر روي "مچك" بسيار مطلوب به نظر مي‌رسد.

در بازيابي متن استفاده از ريشه يابي واژگان، با افزايش "بازخواني"، "دقت" را اندكي كاهش مي‌دهد. به طـور كلـي مـي‌تـوان گفت كه در مواردي كه بالا بودن مقدار "دقت" اهميت بيشتري نسبت به مقدار "بازخواني" دارد، استفاده از ريشه يـابي واژگـان توصـيه نمي‌شود. اما اگر بالا بودن مقدار "بازخواني" از اهميت بيشتري برخوردار باشد، ريشه يابي واژگان كمك شاياني به ايـن امـر مـي‌نمايـد. براي افزايش بيشتر مقدار "بازخواني" پس از ريشه يابي واژگان، گروه بندي واژه‌هـا راه حـل مناسـبي بـه شـمار مـي‌رود كـه مـي‌توانـد موضوعي براي تحقيقات آينده مي‌باشد . علاوه بر اين در تحقيقات بعدي مي‌توان به بررسي روشهاي خودكار وزن دادن به واژگان مـتن و تأثير آن بر روي نتايج بازيابي نيز پرداخت. همچنين مدلهاي بازيابي متن بر روي متون فارسي مي‌توانـد مـورد بررسـي بيـشتر قـرار گيرد.

با بررسي روشهاي ريشه‌يابي واژگان، و مزايا و معايب هر يك مي‌توان گفت براي يافتن ريشه حقيقي، روش "حذف پيشوند و پسوند" بهترين روش به شمار مي‌رود. براي استفاده از اين روش در زبان فارسي، بايد با بررسي شيوه ساخت واژگان، قوانين ريشه‌يابي استخراج گردد. در اين قوانين به نقش(نحو) واژه در جمله توجهي نمي‌شود؛ زيرا تجزيه نحوي جمله منجر به افزايش زمان ريشه‌يابي مي‌شود، ضمن آنكه در ميزان صحت نتايج تأثير بسيار اندكي خواهد داشت. با توجه به اينكه در زبان فارسي واژه‌هاي مركبي وجود دارد كه مركب بودن آنها با صرف واژگان مشخص نمي‌شود، و نيز انواع و اقسام استثنائاتي كه در قوانين ساخت واژه و در نتيجه قوانين ريشه‌يابي به چشم مي‌خورد، طراحي يك الگوريتم مناسب، مشكل به نظر مي‌رسد. در اين مقاله با بررسي چنين مشكلاتي، الگوريتم ريشه‌يابي واژگان زبان فارسي پيشنهاد گرديد. با توجه به اهميت ارزيابي نتايج ريشه‌ياب، با انتخاب بهترين روش ارزيابي، نتايج "بن" نيز مورد بررسي قرار گرفت.

براي ارزيابي دقت كلي[1] ريشه‌يابها مي‌توان نتايج آنها را با روش قطع طولي[2] واژه‌ها مقايسه نمود. در روشهاي قطع طولي، qحرف ابتدايي و يا انتهايي هر واژه به عنوان ريشه آن در نظر گرفته مي‌شود[۱۵]. واضح است كه نتايج به دست آمده از "بن" با نتايج روشهاي قطع طولي قابل مقايسه نيست. بنابراين بحث در مورد يك ريشه‌ياب قطع طولي براي زبان فارسي ضروري به نظر نمي‌رسد. ذكر اين نكته لازم به نظر مي‌رسد كه حتي با تجزيه نحوي جمله نيز نتيجه ۱۰۰% قابل حصول نخواهد بود؛ چرا كه مواردي مي‌توان يافت كه نقش واژگان نيز كمكي به ريشه‌يابي نمي‌كند و بايد به دنبال روشهاي ديگري بود.

تشكر و قدرداني

دوستان زيادي با در اختيار گذاشتن متون تايپ شده خود و نيز طرح پرس‌وجو در انجام اين مقاله نقش داشته‌اند. با توجه به كثرت اين دوستان، بدون ذكر نام تك‌تك آنان، در اينجا از همگي اين عزيزان سپاسگزاري مي‌شود.

مراجع

[۱] عليشاهي، افرا، و لوكس، كارو. "يك الگوريتم خوشه بندي فازي براي طبقه بندي خودكار مفاهيم استخراج شده از متون غير ساختيافته". در مجموعه مقالات چهارمين كنفرانس سالانه انجمن كامپيوتر ايران، ص ۱۱۰-۱۱۷، بهمن۱۳۷۷.

[2] Chen H., Schatz B., Ng T., Martinez J., Kirchhoff A., and Lin C., “A Parallel Computing Approach to Creating Engineering Concept Spaces for Semantic Retrieval: The Illinois Digital Library Initiative Project”, *IEEE Trans. on Pattern Analysis and Mach. Intell.*, Vol. 18, No. 8, pp. 771-782, 1996.

[3] Hmeidi I., Kanaan G., and Evens M., “Design and Implementation of Automatic Indexing for Information Retrieval with Arabic Documents”, *Journal of the American Society for Information Science*, Vol. 48, No. 10, pp. 867-881, 1997.

[4] Salton G. and Mc Gill M. J., *Introduction to Modern Information Retrieval*, Mc Graw Hill, New York, 1983.

[۵] يوسفي گرجي، نبي. تحليل آماري دقت نتايج يك سيستم بازيابي متون فارسي، پايان نامه كارشناسي نرم افزار، دانشكده مهندسي كامپيوتر، دانشگاه صنعتي امير كبير، ۱۳۸۰.

[6] Rijsbergen C.J. Van, *Information Retrieval*, http://www.dcs.gla.ac.uk/Keith

[7] Luhn H. P., “A Statistical Approach to Mechanized Encoding and Searching of Literary Information”, *IBM Journal of Research and Development*, Vol. 1, No. 4, pp. 309-317, 1957.

[8] Salton G., *Automatic Text Processing*. Reading, Mass.: Addison-Wesley, 1989.

[۹] تشكربي، مسعود. بررسي و ارزيابي روشهاي شاخص گذاري خودكار براي متون فارسي در قلمرو كامپيوتر، پايان نامه كارشناسي ارشد نرم افزار، دانشكده مهندسي كامپيوتر، دانشگاه صنعتي اميركبير، ۱۳۸۰.

[1] general accuracy
[2] length truncation

[10] Paice C. D., “An Evaluation Method for Stemming Algorithms”, *Proc. of the seventeenth annual international ACM-SIGIR conference on Research and development in information retrieval*, pp. 42-50, 1994.
[11] PORTER M. F., "An Algorithm for Suffix Stripping", *Program*, vol. 14, no. 3, pp. 130-137, 1980.
[12] PAICE, C. D., "Another Stemmer", *ACM-SIGIR Forum*, vol. 24, no. 3, pp. 56-61, 1990.
[13] DAWSON J., "Suffix Removal and Word Conflation", *ALLC Bulletin*, Michelmas, pp. 33-46, 1974.
[14] LOVINS J. B., "Development of a Stemming Algorithm", *Mechanical Translation and Computational Linguistics*, vol. 11, no. 1-2, pp. 22-31, 1968.
[15] Frakes W. B., *Stemming Algorithms*, http://matrix.nbu.bg/books/books/book5/chap08.htm
[١٦] حسابي، محمود. فرهنگ واژگان انگليسي به فارسي. تهران: چاپ اول، ١٣٧٢.
[١٧] انوري، حسن، و احمدي گيوي، حسن. دستور زبان فارسي١. تهران: فاطمي، چاپ دوازدهم، پاييز١٣٧٤.
[١٨] انوري، حسن، و احمدي گيوي، حسن. دستور زبان فارسي٢. تهران: فاطمي، چاپ دوازدهم، پاييز١٣٧٤.
[١٩] سجادي، علي محمد، حاكمي والا، اسماعيل، استخري، منصور، و سلطاني گردفرامرزي، علي. نگارش و دستور زبان فارسي(٥). تهران: چاپ و نشر ايران، ١٣٧٦.
[٢٠] حري، عباس. " كامپيوتر و رسم‌الخط فارسي". فصلنامه پيام كتابخانه، سال سوم، شماره اول، ص ٦-١١، بهار١٣٧٢
[٢١] اكبري نژاد، سعيد. "فاصله خالي ميان واژه‌ها در ذخيره و بازيابي اطلاعات". فصلنامه كتاب، دوره هشتم، شماره ١و٢، ص ٤٩-٥٦، بهار و تابستان١٣٧٦.
[22] Lennon M., Pierce D. S., Tarry B. D., and Willett P., “an Evaluation of Some Conflation Algorithms for Information Retrieval”, *Journal of Information Science,* no. 3, pp. 177-183, 1981.

ضميمه

پرس‌وجوهاي جمع آوري شده از افراد به همراه تعداد و شماره متنهاي مربوط هر يك

ردیف	پرس‌وجو	تعداد متنهاي مربوط	شماره متنهاي مربوط
١	آموزش و كامپيوتر	٢	١١٠ ١٣٢
٢	اتوماتاي يادگير و پردازش تصاوير	٢	١٨٨ ١٩٢
٣	(اتوماتاي يادگير و كنترل درخواست ورود) يا ربات	١٩	١٦ ١٨ ٢٤ ٣٣ ٣٤ ٦٧ ١٤٦ ١٧١ ١٧٩ ٢٢٣ ٢٢٤ ٢٢٨ ٣٤٩ ٣٦٧ ٣٧٠ ٣٨٩ ٣٩٠ ٣٩١ ٣٩٧
٤	اثبات خودكار قضايا	٣	٢١٤ ٢٦١ ٢٧٥
٥	استخراج ويژگي تصاوير و پزشكي	٧	٩ ٥٤ ٥٧ ٢٤٧ ٢٥٢ ٣١٠ ٣١٧
٦	اسكلت موازي	٢	٢٦٤ ٢٧٨
٧	بازشناسي گفتار و مدل مخفي ماركف يا شبكه عصبي	٨	٨٢ ١٨٥ ٢٠٤ ٢٠٥ ٣٢٤ ٣٢٥ ٣٧١ ٣٧٤
٨	پايگاه داده توزيع شده يا پايگاه داده فازي	٥	١٠٧ ١٢٩ ١٣٩ ١٩٤ ٢٣٢
٩	پردازش تصاوير و تبديل هاف	٢	٩ ٢٣٦
١٠	پردازش زبان و نظريه گشتاوري	٢	٣٢٣ ٤٤٩
١١	پردازش گفتار و شناسائي گفتار يا شناسائي هويت گوينده	١٥	٦٤ ٨٢ ١٧٠ ١٨٥ ٢٠٣ ٢٠٥ ٢٨٧ ٢٨٨ ٣٢٤ ٣٢٥ ٣٦٢ ٣٧١ ٣٧٤ ٣٧٥ ٤١٧
١٢	پردازش موازي و برنامه نويسي موازي	١٢	٣ ٣٠ ٥٩ ٩٣ ٩٤ ١١٢ ١٢٩ ٢٣٧ ٣٣٨ ٣٤٠ ٤٢٥ ٤٢٨
١٣	پژواك و تلفن	١	٣٠٦
١٤	ترافيك و پردازش تصوير	٧	٤٢ ٤٨ ٦٣ ٧٣ ٣٧٩ ٤٢٧ ٤٤٧
١٥	تصديق هويت و تلفن	١	١٧٠
١٦	جاوا و پردازش توزيعي	٤	٥٩ ٧٧ ٢١٢ ٣٣١
١٧	ربات و بينايي ماشين	٧	١٦ ٢٤ ٣٣ ١٠١ ١٤٦ ٢٢٤ ٢٢٨
١٨	رمز نگاري	٤	٤ ١٦٨ ٣١٩ ٣٢٠
١٩	رمز نگاري و امنيت و شبكه هاي كامپيوتري	٤	٤ ٣٧ ٨١ ٨٦
٢٠	زلزله نگاري و شبكه هاي عصبي	١	٨
٢١	سيستم عامل توزيع شده غيراز سيستم عامل شريف	٣	١٠ ٥٣ ٣٩٨
٢٢	شبكه هاي عصبي و اثر انگشت	١	١٢٥
٢٣	شبكه هاي عصبي و تشخيص حروف	٩	٣٢ ٦٠ ٦٢ ٧٩ ٨٩ ٣٣٤ ٣٥١ ٣٥٢ ٤٢١
٢٤	شبكه هاي عصبي و فشرده سازي تصوير	٢	٦٦ ٧٦
٢٥	فشرده سازي متن يا فشرده سازي تصوير	٧	٢٣٩ ٢٦٦ ٣٢١ ٣٥٣ ٣٥٨ ٣٨٨ ٤٢٢
٢٦	(شبكه يا اينترنت) و امنيت	٦	٣٧ ٨١ ٨٦ ١٢٠ ٢٢٠ ٢٩٨
٢٧	فشرده سازي و فركتال	١	١١
٢٨	(كنترل پروژه يا مديريت پروژه) و سيستمهاي خبره	١	١

٣٦٥ ٣٦٩	٢	كنترل ترافيك تلفني	٢٩
١٨٥ ٢٠٤ ٢٤٠ ٢٨٨ ٣٢٤ ٣٧١	٦	مدل مخفي ماركف و (پردازش گفتار يا شناسايي حروف)	٣٠
٦ ٢٧ ٩٠ ١١٣ ٣١٤	٥	يادگيري و پايگاه فوائين	٣١
٢١ ٢٢ ١٧٩ ١٨٠ ١٨١ ١٨٢ ١٨٣ ١٨٨ ١٩٢ ٢٤١ ٢٥٦ ٢٥٧ ٢٦٠ ٢٧١ ٢٧٤ ٣١١ ٣٨٩ ٣٩٠ ٣٩١ ٣٩٤ ٣٩٧	٢١	(يادگيري يا آموزش) و (اتوماتون يا اتوماتا يا اتوماتون)	٣٢