به نام خدا

یک سیستم بازیابی اطلاعات متنی برای زبان فارسی
بر پایه نمایه گذاری معانی پنهان

به وسیله‌ی :

احمد یوسفان نجف‌آبادی

پایان نامه

ارائه شده به معاونت تحصیلات تکمیلی به عنوان بخشی

از فعالیت‌های تحصیلی لازم برای اخذ درجه‌ی کارشناسی ارشد

در رشته‌ی

کامپیوتر

از دانشگاه شیراز

شیراز ، ایران

ارزیابی شده توسط کمیته پایان نامه با درجه: عالی

دکتر منصور ذوالقدری جهرمی ، استادیار بخش کامپیوتر (رئیس کمیته) .....

دکتر حسن اقبالی جهرمی ، دانشیار بخش کامپیوتر (رئیس کمیته) .....

دکتر جلال رحیمیان ، دانشیار بخش زبانهای خارجی و زبان‌شناسی .....

شهریور ماه 1382

چکیده

یك سیستم بازیابي اطلاعات متني براي زبان فارسي
بر پایه نمایه گذاري معاني پنهان

به وسیله‌ی :

احمد یوسفان نجف‌آبادی

با افزایشِ روزافزونِ حجمِ اطلاعاتِ متنی به صورتِ دیجیتال ، همزمان نیاز به سیستمهای دقیق و کارآ برای بازیابی اطلاعات افزایش می‌یابد.

در این پایان نامه یک سیستم بازیابی اطلاعات متنی برای زبان فارسی بر پایه‌ی نمایه گذاری معانی پنهان طراحی و پیاده سازی شده است. برای کمک به یافتن فهرست واژه‌های سراسری و ریشه‌یابی ، یک زبان برنامه‌نویسیِ ساده طراحی و بر پایه‌ی قاعده‌های زبان فارسی روشی نوین برای شناسایی خودکار فعلهای فارسی پیشنهاد شده است.

کارآیی سیستم با ریشه‌یابی و بدون ریشه‌یابی با استفاده از یک مجموعه اسناد گردآوری شده به این منظور و به کمک معیارهای دقت و یادآوری مورد ارزیابی قرارگرفته است.

سپاسگزاري

بر خود می‌دانم از همه‌ی بزرگوارانی که مـرا در انجـام ایـن پایـان نـامه راهنمـایی فرموده‌انـد ، سپاسگزاری کنم. بویژه از دکتر منصور ذوالقدری ، دکتر حسن اقبالی ، مهندس هـومن تحیـری (استادان بخش کامپیوتر دانشگاه شیراز) و دکتر منصور رستگار فسایی ، دکتر کاووس حسن لی ، دکتر اکبر صیادکوه (استادان بخش ادبیات فارسی دانشگاه شـیراز) و دکـتر جلال رحیمیـان ، دکتر لطف الله یارمحمدی (استادان بخش زبان و زبان‌شناسی دانشـگاه شـیراز) و دکـتر فرهـاد اورومچیان (استاد بخش کامپیوتر دانشگاه تهران) و مهندس احمد خاتون آبادی ، مهندس احمد محمدی ، مهندس مازیار صالحی (دانش آموختگان بخش کامپیوتر دانشگاه شـیراز) و مهنـدس مهران ترحمی ، مهنـدس مرتضـی تـاجبخش ، مهنـدس مهـدی احمـدی (دانشـجویان بخـش کامپیوتر دانشگاه شیراز) ، سپاسگزاری می‌کنم.

همچنین جـا دارد از کارمنـدان بخـش کـامپیوتر بـویژه خـانم سـیم‌و زر ، کارمنـدان کتابخانه‌ی دانشکده‌ی مهندسی بویژه آقای دشتی و از آقای قنبرپور کارمند تحصیلات تکمیلـی دانشکده‌ی مهندسی که بیش از انجام وظیفه با من همکاری شایسته‌ای داشته‌اند ، سپاسـگزاری کنم.

همچنین بر خـود می‌دانـم از همه‌ی دوسـتان دوران دانشـجویی‌ام در دانشـگاه شـیراز سپاسگزاری کنم ، که همواره مرا یاری رسانده‌اند و چیزهای بسیاری از آنها آمـوخته‌ام. بـویژه از مهدی صالحپور ، کوروش پارسایی ، علی یزدان‌پنـاه ، حسـن چیـذری ، غلام یعقـوبی ، مهـدی افتخاری ، رضا حقیقی ، علی جعفری و کامران خداپرستی سپاسگزارم.

همچنین از خانواده‌ام که همواره پشتیبانم در زندگی بوده‌اند ، بسیار سپاسگزارم.

فهرست مطالب

عنوان صفحه

فهرست شکلها

عنوان صفحه

فصل اول

مقدمه

1- مقدمه

انسانها نزدیک به چهار هزار سال است که اطلاعات را برای به کارگیری در آینده ، سازماندهی و دسته بندی می‌کنند تا بتوانند در صورت نیاز در آینده آنها را به سادگی بیابند یا به عبارت دیگر بازیابی نمایند [baeza-yates,1999]. با افزایش دانش و رشد روز افزون اندازه‌ی اطلاعات در گردش و بایگانی شده ، دردسر بزرگی فراروی انسان امروزی به وجود آمده است [Kowalski,1998]. امروزه بیش از ۹۳ درصد از اطلاعات در جهان به صورت رایانه‌ای است و اندازه‌ی اطلاعات رایانه‌ای هر سال دو برابر می‌شود [Tang,2003]. نیاز به یافتن نوشته‌هایی در یک زمینه‌ی ویژه ، یکی از نیازهای روزمره است و هر روز بیش از پیش این روند افزایش می یابد. به همان اندازه نیز یافتن نوشته‌هایی که نیاز است ، سخت‌تر می‌شوند. روشهای جستجو باید گسترش یابند تا این دشواریها کاهش یابند یا دستِ کم بر تعداد این دشواریها افزوده نشود. به روند جستجو در گردایه‌ای از نوشته‌ها و یافتن نوشته‌هایی در پاسخ به پرس و جوی کاربر[1] ، بازیابی اطلاعاتِ نوشتاری[2] گفته می‌شود. بازیابی اطلاعات روشهای گوناگونی برای سازمان دهی و مدل کردن اطلاعات و نیازهای کاربر پیشنهاد می‌دهد.

1.1- بازیابي اطلاعات چیست ؟

کوشش برای خودکار کردن همه‌ی کارهای جستجو ، دانش ویژه‌ای را به نام بازیابی اطلاعات به وجود آورده است. به کمک آمار و تئوری احتمالات و روشهای یادگیری ماشین ، مدلهای گوناگونی از دانش برای بازیابی اطلاعات ایجاد شده است [salton,1983]. نوشته‌های گوناگون با قالبهای گوناگون (ساخت یافته ، نیمه ساخت یافته و ساخت نیافته) ، زبانهای گوناگون و

[1] 1_ query

[2] 2_text information retrieval

کاربران بسیاری که رده‌های گوناگونی از دانش را دارند؛ پیچیـدگی‌های بازیـابی اطلاعـات را دو چندان کرده‌اند [salehi,2000].

1.2- نقش زبان در بازیابي اطلاعات

همه‌ی روشهای بازیابی اطلاعاتِ نوشتاری کم وبیش به زبانی که نوشته‌ها به آن زبـان (یکـی از زبانهای زنده‌ی دنیا) نگاشته شده‌اند ، بستگی دارند ؛ بنابراین ساختار و دیگر ویژگیهای آن زبان باید بخوبی بررسی گردند. گرچه روشهایی نیز وجود دارند که به زبان نوشته‌ها بستگی ندارنـد ؛ ولی این روشها هنوز در دستِ آزمایش هستند و پاسخ شایسته‌ای از آنها گرفته نشـده اسـت. در این پایان نامه بازیابی اطلاعات ، بر روی زبان فارسی به کار گرفته شده است.

شکی نیست که سیستمهای اطلاعاتی در دنیای امروز که بدرستی عصر اطلاعات نامیده شده، نقشی حیاتی در کارهای روزمره‌ی اداره‌ها، محیطهای اداری و صنعتی دارند. بـومی سازی این سیستمها برای پوشش دادن به اطلاعات فارسی و گنجاندن مـوارد زبـان شـناختی و فرهنگی بویژه پس از آمدن رسانه‌ای توانمند و فراگیری همچون شبکه‌ای جهانی[1] ، امری مهـم و غیر قابل چشم پوشی است. متأسفانه نبودِ استانداردهای ملی در زمینه فارسـی چـه از کُـدهای حروف الفبا ، چه قالبِ سـندها[2] و چـه ذخیـره و بازیـابی دردسـرهای بزرگـی را بـرای کارهـای رایانه‌ای فارسی آفریده است. این دشواریها ناهماهنگی‌های زیادی میان سـازندگان نـرم افـزار و سرویسهای خدماتی شبکه‌ی جهانی به وجود آورده‌اند. دربـاره‌ی بسـیاری از پروژه‌هـای فارسـی هنوز حتی نیازمندیهای واقعی مشخص نشده است تا بتوان کارهای انجام شـده‌ی کنـونی را بـا هم مقایسه کرد [زارع ۱۳۸۰].

1.3- ساختارپایان نامه

در نگارش پایان نامه کوشش شده است که ساختاری پیوسته برای کارهـای انجـام شـده نشـان داده شود ، درحالی که انجام این پایان نامه به هیچ روی ساختار پیوسته‌ای نداشته اسـت. بـرای جلوگیری از سردرگمی خواننده‌ی این نوشتار ، ساختار نوشته شده بر پـایه‌ی آخریـن ویرایشـها آماده شده است. بنابراین بسیاری از پیچیدگیها و درهم تنیدگیهایی که این پایان نـامه بـا آنهـا

[1] 1_ internet

[2] 2_ documents format

روبرو بوده است ، از دیده خوانندهی این نوشتار پوشیده شده است. افزون بر این ، اندازهی یک پایان نامه گنجایش نگارش همهی روشهای آزموده شده را ندارد.

برای گذر از سردرگمیهایِ بسیارِ پایان نامه بیش از یک سال و اندی از زمان کار بر روی آن به سر زدن به جاهای گوناگون برای یافتن پاسخِ پرسشهای بیشمار ، که پایان نامه با آنها روبرو بود ، گذشت. همچنین در آن زمان به چگونگی انجام کار نیز بسیار اندیشیده میشد و بسیاری از روشهای اندیشیده شده نیز آزموده میشد و اغلب پاسخی از آنها گرفته نمیشد. پس از آن نخستین گام ، که پایهی ساختار کنونی است ، گردآوری بیش از هفتاد مگابایت صفحهی فارسی از شبکهی جهانی بود. واژههای فارسی به دست آمده از درون این صفحهها نخستین دادههای در دسترس برای پایان نامه بودند. کوششهای پس از این گام هدفمندتر از پیش انجام شد و نیز برنامه ریزی امکان پذیر گشت. پس از آن ، دادههای فراوان دیگری نیز گردآوری شد. گرچه بهتر دیده شد که در نگارش ، این بخش پس از چند بخش دیگر گذاشته شود.

فصل دوم با گذری بر بازیابی اطلاعات آغاز شده است. سپس به روشهای ارزیابی سیستم بازیابی اطلاعات پرداخته شده است و فشردهای نیز از روش نمایه گذاری معانی پنهان آورده شده است. به دلیل کاستن از اندازهی پایان نامه ، این فصل بسیار فشرده شده نوشته شده است. گرچه کوشش شده است تا خواننده بتواند با خواندن این فصل ، دیدِ شایستهای از چهارچوپ نظری این پروژه به دست آورد. منابع این فصل اطلاعات بیشتری در بارهی این فصل در اختیار خواننده خواهند گذاشت.

در فصل سوم برخی از دشواریهایی که پایان نامه با آنها روبرو بوده است ، آورده شده است. همچنین برخی از نکتهها و روشهای به کار گرفته شده برای برطرف کردن این دشواریها نیز نوشته شده است. برای نمونه در بند استاندارد خط فشردهای از چگونگی گسترش و پیاده سازی خطِ فارسی در رایانه آورده شده است.

در فصل چهارم به پیاده سازی یک سیستم بازیابی اطلاعاتِ فارسی پرداخته شده است. میشد این فصل را به چندین فصل جداگانه بخش کرد ولی بهتر دیده شد که بیشترِ کارِ پیاده سازی در این فصل یک جا آورده شود. همانگونه که پیش از این گفته شد تنها نسخهی پایانی کارهای انجام شده ، در این فصل آورده شده است. چون که ساختار فصلها روشن است ، افزودن توضیحهایِ بیشتر در اینجا به نظر زیادی است.

فصل دوم

چهارچوب نظري

2- چهارچوب نظري

2.1- بازيابي اطلاعات

بازیابی اطلاعات با نمـایش ، ذخیـره سـازی ، سـازماندهی و چگـونگی دسـتیابی بـه داده‌هـای اطلاعاتی سر و کار دارد. نمایش و سازماندهی شایسته‌ی داده‌ها به کاربر کمک می‌کند تا بتواند به آسانی داده‌هایی را که به آنها نیاز دارد ، بیابد [Baeza-Yates,1999].

بازیابی اطلاعات دربردانده‌ی بازیابی تصویر ، صدا ، نوشته و ... اسـت کـه در هـر کـدام روشهای ویژه‌ای برای بازیابی به کار گرفته می‌شود. در اینجا بازیابی اطلاعات بـر روی نوشـته‌ها انجام می‌شود و بنابراین به آن سیستم بازیابی اطلاعات نوشتاری گفته می‌شود.

امروزه کاربران به هزاران منبع اطلاعاتی و میلیونها سند دسترسی دارند. همچنیـن بـه گسترش کتابخانه‌های الکترونیکی به عنوان وسیله‌ای آسان برای دسترسـی بـه اطلاعـات بهـای زیادی داده می‌شود. پرسشی که پیش می‌آید این است که چگونه امکاناتی برای دسترسی آسان به این اطلاعات فراهم شود. اگر بنا نباشد که کـاربران در دریـایی از سـندها رهـا شـوند ؛ بایـد امکاناتی برای دسترسی و جستجو بر روی سندها به آنها داده شود.

کاربران سیستم بازیابی اطلاعات نیازهای اطلاعاتی گوناگونی دارند. دانشمندانی که بـه دنبال مقاله‌هایی در یک زمینه‌ی ویژه‌ی آزمایشگاهی هستند ، مهندسانی که می‌کوشند روشـن کنند که آیا ایده‌ای که در نظر دارند یک ایده‌ی نوین است یا پیش از این دیگـران ایـن ایـده را پیشـنهاد داده‌انـد و حـق مـؤلف[1] آن را بـه دسـت آورده‌انـد ، خریـداران وسـیله‌های گونـاگون می‌کوشند تا اطلاعاتی درباره‌ی وسیله‌های جدید و کارآیی و بهای آنها به دست آورند. به عبارت دیگر کاربران سیستم بازیابی اطلاعات ، پیش زمینه‌های گوناگونی دارند و به دلیلهای گوناگونی نیز از امکانات سیتم بازیابی اطلاعات بهره می‌برند [salton,1983].

[1] 1_ copy right

برای نمونه فهرست آغازین یک کتاب برای نشان دادن سازماندهی نوشته‌های آن کتاب به کار گرفته می‌شود. چون کـه انـدازه‌ی اطلاعـات همـواره در حـال افزایـش اسـت ، بنـابراین ساختمان داده‌های بهتری برای دسترسی سریعتر به داده‌های ذخیره شده نیاز اسـت. «نمـایه»[1] یک ساختمان داده‌ی کُهن و سراسری برای افزایش سرعت در بازیابی اطلاعات از یک گـردایه‌ی اطلاعات یا سندها است. نمایه‌ها هسته‌ی اصلی هر سیستم بازیابی اطلاعات نوین هستند. بـرای قرنها نمایه‌ها به صورت دستی ساخته می‌شدند و سندها (برای نمونه کتابهای یک کتابخانه) در ساختاری دسته بندی شده و سلسله مراتبی نگهـداری می‌شـدند. امـروزه نمایه‌هـای بـزرگ بـه کمک رایانه‌های پیشرفته ساخته می‌شوند. کتابخانه‌ها از نخسـتین جاهـایی بودنـد کـه بازیـابی اطلاعاتِ رایانه‌ای در آنها به کار گرفته شد. در نخستین نسل از برنامه‌های رایانه‌ای به کار گرفته شده در کتابخانه‌ها تنها می‌شد بر پایه‌ی نـام کتـاب و نـام نویسـنده جسـتجو را انجـام داد. در دومین نسل از این برنامه‌ها ، کارهایی همچون جستجو بـر پـایه‌ی واژه‌هـای کلیـدی و موضـوع کتاب (زمینه‌های کتاب) به این برنامه‌ها افزوده شد. همچنین برخی از امکاناتِ پردازشِ پـرس و جوی کاربر نیز فراهم گردید. در اینجا کاربر کسی است که برای یـافتن یـک یـا چنـد کتـاب ، سیستم کتابخانه را به کار می‌گیرد و برای یافتن کتاب یا کتابهایی که نیاز دارد ؛ پرس و جـویی را به سیستم بازیابی اطلاعات (در اینجا سیستم کتابخانه) می‌دهد تا سیسـتم بتوانـد کتـاب یـا کتابهایی را که مورد نظر کاربر است ، به کمک پرس و جوی کـاربر بیابـد. هـم اکنـون سـومین نسل از برنامه‌های رایانه‌ای کتابخانه‌ها در حال گسترش است. در ایـن نسـل بیشـتر بـر سـاخت رابطهای گرافیکی شایسته ، صـفحه‌های الکـترونیکی ، ویژگیهـای اَبَـر متن[2] و سیسـتمهای بـاز تأکید می‌شود [Baeza-Yates,1999].

2.1.1- طرح کلي يک سيستم بازيابي اطلاعات

سیستم بازیابی اطلاعات دربردارنده‌ی انبوهی از سندها یا چکیده‌ی آنها یا نمایه‌ی سندها است. این سندها یکی از ورودیهای سیستم بازیابی هستند. این سندها بیشتر از روی شبکه‌ی جهـانی و بصورت خودکار گردآوری می‌شوند. سپس بـه شـکل ویـژه‌ای تبـدیل می‌شـوند و در سیسـتم نگهداری می‌شوند. کاربر ، سیستم بازیابی اطلاعات را به کار می‌گیرد تا سندهایی را که به آنهـا علاقه‌مند است ، بیابد. کاربر با دادن پرس و جو یا پرس و جوهایی پیاپی به سیسـتم ، رسـته‌ی سندهایی را که نیاز دارد نشان می‌دهد ، سیستم می‌کوشد تا این سندها را یافته و اشاره‌گرهایی به این سندها را در دسترس کاربر بگذارد. به کمک ایـن پاسـخهای(اشـاره‌گرهای) سیسـتم بـه

[1] 1_ index

[2] 2_ hyper text

احتمالِ زیاد ، کاربر می‌تواند سندهایی را که نیاز دارد ، بیابد. پس پرس و جوی کـــاربر ، ورودی سیستم بازیابی اطلاعات است و خروجی سیستم یک دسته از اشاره‌گرها به سندهایی است کـــه به احتمال زیاد کاربر به آنها علاقه‌مند است.

بازیابی اطلاعات نوشتاری را می‌توان بـه شـکلهای گونـاگونی بخـش بنـدی کـرد. در [baeza-yates,1999] یک سیستم بازیابی اطلاعات به شکل زیر بخش بندی شده است.

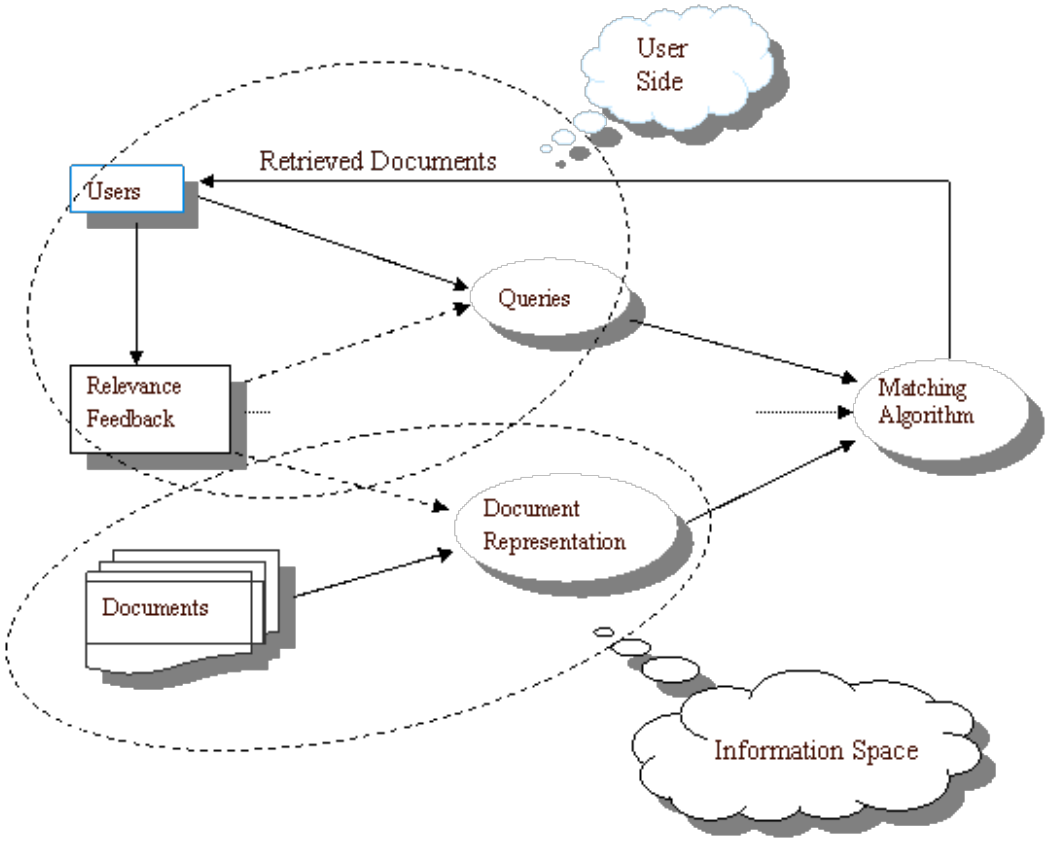
۱- گردایه‌ای از سندها

۲- گردایه‌ای از پرس و جوها

۳- یک قالب برای چگونگی نمایش درونی سندها و پرس و جوها (نمایه گذاری سندها و پـــرس و جوها)

۴- نگاشتی که ما را از پرس و جوی کاربر بـه زیـر گـردایه‌ای از سـندها می‌بـرد کـه بیشـترین نزدیکی را با پرس و جوی کاربر دارند. به این نگاشت ، نگاشت شباهت گفته می‌شود.

همچنین می‌توان سیستم بازیابی اطلاعات را به دو بخش کاربر و اطلاعات افـراز نمـود. در شکل صفحه‌ای بعد طرح کلی یک سیستم بازیابی اطلاعات نشان داده شده اسـت. بـازخورد ارتباطی برای بهتر نمودن پاسخ سیستم به کار گرفته می‌شـود [Gordon,1988]. بـرای آشـنا شدن با بازخورد ارتباطی می‌توانید [دستغیب،۱۳۸۰] را ببینید.



شکل 2-1- طرح کلی یک سیستم بازیابی اطلاعات

2.1.2- نمایه گذاري

نخستین کار در نمایش و ذخیره‌ی سندها در درون سیستم بازیابی اطلاعات ، نمایه گذاری است. اگر بنا بود سندها به همان شکل آغازین خود در سیستم بازیابی اطلاعات نگهداری شوند ، دستِ کم دشواریهای زیر برای سیستم به وجود می‌آمد:

۱- جای بسیار بزرگی برای نگهداری سندها نیاز می‌بود. با افزایش روزافزون اطلاعات ، فن آوری رایانه‌ها پاسخگوی ذخیره سازی همه‌ی سندها نمی‌بود.

۲- جستجو بسیار کُند انجام می‌شد. زیرا داده‌ها (سندها) در ساختاری شایسته برای جستجو ذخیره نشده بودند و می‌بایست در هنگام جستجو کارهایِ اضافی فراوانی انجام می‌شد.

۳- سیستم پاسخ شایسته‌ای نمی‌داد. زیرا برخی واژه‌ها معنای نوشته را در برندارند و بسیاری از واژه‌ها ریشه‌ی یکسانی دارند که یک معنا را می‌رسانند. بنابراین سندهایی که باید در پاسخ

سیستم می‌آمدند ، از دید سیستم پوشیده می‌شدند. همچنین ارزش گذاریِ سندها بـر پـایه‌ی نزدیکی یا دوری از پرس و جوی کاربر چندان ساده نمی‌بود.

نمایه گذاری سندها ، یک نمایش جدید از سندها را می‌سـازد. نمـایه گـذاری بـر روی پرس و جوی کاربر نیز انجام می‌شود تا پرس و جوی کاربر بـه صـورت نمـایش درونـی سـندها درآید . با نمایه گذاری پرس و جوی کاربر ، سنجش دوری یا نزدیکـی سـند بـه پـرس و جـوی امکان پذیر می‌شود.

نمایه گذاری سندها بخشهای زیر را دارد:

۱- تبدیل قالب پرونده[1] : در نخستین گام باید قالبهای گوناگون پرونده‌ها به یک قالب تبدیل شوند.

۲- بخش بندی نوشته[2] : با به کار بردن برخی از ویژگیهای نوشته می‌توان کار بخش بندی را انجام داد. برای نمونه واژه‌هایی که با قلم ویژه یا با اندازه‌ی بزرگتر نوشته شده‌اند یا زیـر نویسـها در بخش بندی نوشته به کار گرفته می‌شوند. همچنین برخی از واژه‌های کلیدی یا قراردادهـای قالب پرونده می‌توانند به بخش بندی نوشته کمک کننـد. در [خـاتون آبـادی،۱۳۸۰] برخـی از روشهای خودکار بخش بندی نوشتار به کمک ویژگیهای پرونده بخوبی بررسی شده است.

۳- جدا سازی واژه‌ها[3] : در اینجا واژه به یک رشته نویسه[4] گفته می‌شود که با فاصـله[5] از هـم جدا می‌شوند [salton,1983]. واژه‌ها پایه‌های اصلی سیستم بازیابی اطلاعات هستند. به تعداد واژه‌های جداگانه در یک گردایه از سندها بُعدهای آن گردایه گفته می‌شود. جدا نمـودن واژه‌هـا در زبان انگلیسی بسیار ساده است ، زیرا وجود هر یک از نویسه‌های فاصله در دو طرف یک واژه مرزهای آن واژه را نشان می‌دهد. در حالی که این کار در فارسی با دشواری فراوانی روبرو است. که در فصل بعد به این دشواری پرداخته می‌شود.

۴- برداشتن واژه‌های سراسری[6] : برخی از واژه‌ها در جمله و در زبـان تنهـا نقـش دسـتوری دارند و دربردارنده‌ی اطلاعات بنیادی نوشتار (معنای نوشتار) نیستند ؛ به این واژه‌ها ، واژه‌هـای سراسری گفته می‌شود [salton,1983]. بـرای نمـونه فعلهـای کمکـی ، حرفهـای اضـافه و ... واژه‌های سراسری هستند. واژه‌های سراسـری بایـد از سـندها برداشـته شـوند تـا هـم انـدازه‌ی

---

[1] 1_file

[2] 2_ text segmentation

[3] 3_ term extraction

[4] 4_ character

[5] 5_ tab , space , newline

[6] 6_ stop word elimination

اطلاعات کاهش یابد و هم سیستم پاسخ بهتری بدهد. یکی از کارهای انجام شده در این پایان نامه شناسایی واژه‌های سراسری فارسی است.

۵- ریشه‌یابی[1] : جایگزینی ریشه‌ی واژه‌ها به جای خود واژه‌ها اندازه‌ی اطلاعات را کاهش می‌دهد. همچنین با جایگزین کردن واژه‌هایی که دارای یک ریشه هستند با ریشه‌ی آنها ، نوشته‌هایی که یک معنا را می‌رسانند ؛ بیشتر به هم نزدیک می‌شوند. در کارِ ریشه‌یابی بسیاری از پیشوندها و پسوندهای واژه برداشته می‌شوند. برای نمونه اگر کاربر واژه‌ی «گویند» را در پرس و جوی خود به کار برد ؛ به احتمال زیاد او می‌خواهد که سیستم بازیابی نوشته‌هایی را هم که در آنها «گفتیم»،«گفتند»،«می‌گویند»،«گفتن» یا .. نوشته شده است را نیز در پاسخ نشان دهد. یکی دیگر از کارهای انجام شده در این پایان نامه ریشه یابی واژه‌ها در زبان فارسی است.

۶- وزن گذاری واژه[2] : برای نشان دادن وزن یا درجه‌ی ارزش یک واژه در یک سند ، عددی به آن واژه در آن سند نسبت داده می‌شود؛ به این روند وزن گذاری واژه گفته می‌شود. وزن گذاری واژه اثر بسزایی در بهبود کارآیی سیستم بازیابی اطلاعات دارد. فراوانی واژه‌ها در هر سند و رابطه‌ی میان واژه‌ها در هر سند به یافتن درجه‌ی ارزش یک واژه کمک می‌کند.

۷- ساخت عبارت[3] : پس از انجام تبدیلهایی که پیش از این گفته شد ، عبارتهایی ساخته می‌شوند که جایگزین سندها می‌شوند. این عبارتها نمایانگر شکل نوین سندها هستند. همواره در کنار عبارتی که تبدیل شده‌ی یک سند است ، اشاره‌گری به اصل سند گذاشته می‌شود.

۸- ذخیره‌ی نمایه[4] : سر آخر نمایه‌های ساخته شده در پایگاه داده‌ی سیستمِ بازیابیِ اطلاعات ذخیره می‌شوند. جستجوی پرس و جوی کاربر در این پایگاه انجام خواهد شد.

بر روی پرس و جوی کاربر نیز کارهای زیر انجام می‌شود تا نمایش پرس و جوی کاربر نیز همانند نمایش درونی سندها گردد.

۱- جداسازی واژه

۲- برداشتن واژه‌های سراسری

۳- ریشه‌یابی

۴- وزن گذاری واژه

---

[1] 1_ stemming

[2] 2_ term weighting

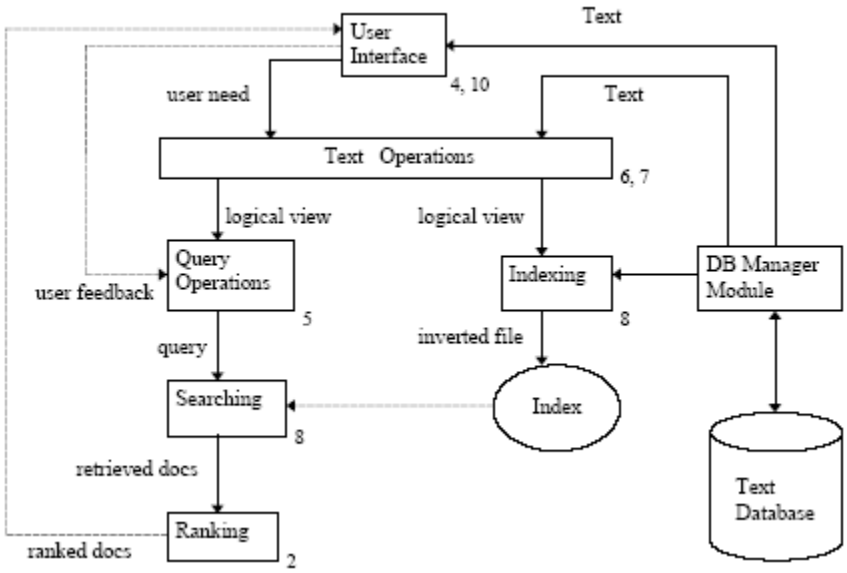
[3] 3_ phrase construction

[4] 4_ index storage

۵- ساخت عبارت

۶- فرستادن عبارت ساخته شده به تابع شباهت

بخشهای جداسازی واژه‌ها ، برداشتن واژه‌های سراسری و ریشه‌یابی از بخشهای پایه‌ای هر سیستم بازیابی اطلاعات نوشتاری هستند. این سه بخش اثر بسزایی بر روی دیگر بخشها و همچنین کارآیی سیستم بازیابی اطلاعات دارند. این بخشها بسیار به زبان نوشته‌ها بستگی دارند [Honglan,2002]. این سه بخش ، کارهای پیش پردازشی[1] بر روی سندها و پرس و جوها نامیده می‌شوند. این بخشها در بسیاری از کاربردهای دیگر نیز به کار گرفته می‌شوند و بخش پایه‌ی بسیاری از سیستمهای پردازش نوشتار هستند. برای نمونه در دستیابی به معنای نوشته‌ها ، خلاصه نمودن نوشته‌ها ، دسته بندی نوشته‌ها و ... این بخشها به کار گرفته می‌شوند. ارزش این بخشها آن چنان است که برخی از کتابهای در زمینه بازیابی اطلاعات نوشتاری بر جدا نمودن این بخشها از دیگر بخشهای بازیابی اطلاعات تأکید می‌کنند. برای نمونه در [Baeza-Yates,1999] طرح کلی یک سیستم بازیابی اطلاعات در شکل 2-2 نشان داده شده است. همان گونه که دیده می‌شود بخش پردازش نوشته در قلب این طرح از بازیابی اطلاعات گذاشته شده است.



شکل 2-2- طرح دیگری از یک سیستم بازیابی اطلاعات

[1] 1_ preprocessing

2.1.3- تابع شباهت

هدف اصلی از هر سیستم بازیابی اطلاعات یافتن سندهایی است که بیشترین نزدیکی را با پرس و جوی کاربر داشته باشند. پس از تبدیل شدن سندها به نمایش درونی و ذخیره سازی آنها سیستم آماده‌ی پاسخگویی به پرس و جوی کاربران است. زمانی که کاربری پرس و جویی به سیستم بازیابی اطلاعات می‌دهد ، نخست آن پرس و جو به نمایش درونی تبدیل می‌شود، سپس تابع شباهت به کار گرفته می‌شود. تابع شباهت دو بخش زیر را دارد :

۱- چگونگی یافتن بهترین سندها در پاسخ به پرس و جوی کاربر. سندهای یافته شده باید بیشترین نزدیکی را با پرس و جوی کاربر داشته باشند. به جای جستجوی پرس و جو بر روی گردایه‌ی سندها ، پرس و جوی تبدیل شده بر روی گردایه‌ی تبدیل شده‌ی سندها جستجو می‌شود. این کار دشواریهای فراوانی دارد. اغلب ، کاربر به روشنی نمی‌داند که برای پرس و جوی خود چه واژه‌های کلیدی را به کار برد که نشان دهنده‌ی منظور وی باشد [Blair,1985].

۲- فهرست سندهای یافته شده (اشاره‌گرهایی به آنها) باید بگونه‌ای آرایش یابند که نزدیکترین سند یافته شده به پرس و جوی کاربر ، در آغاز فهرست گذاشته شود. بنابراین الگوریتم باید همواره عددی به نام عدد شباهت به هر سند بازیابی شده نسبت دهد و اغلب آرایش سندهای بازیابی شده بر پایه‌ی همین عدد خواهد بود. چگونگی نمایش سندهای بازیابی شده اثر بسزایی در دیدگاه کاربر نسبت به سیستم بازیابی اطلاعات دارد.

تابع شباهت روشهایی را به کار می‌گیرد تا بتواند نزدیکترین سندها را به پرس و جوی کاربر بیابد و به هر کدام از این سندها عدد شباهت را نسبت دهد. در چند بند پس از این ، یکی از این روشها آورده شده است.

2.1.4- روش فضاي برداري

این روش را نخستین بار salton و همکاران در دهه‌ی شصت میلادی (۱۹۶۰) در دانشگاه cornel (امریکا) پیشنهاد دادند. سپس سیستمی به نام SMART را که یک سیستم بازیابی اطلاعات است ، بر پایه‌ی این روش پیاده سازی کردند [Salehi,2000]. سیستم SMART هنوز هم پایه‌ی بسیاری از کارهای پژوهشی در زمینه‌ی بازیابی اطلاعات به کمک روش فضای برداری است.

در آغاز همانند همه‌ی روشهای بازیابی اطلاعات کارهای پیش پردازشی روی سندها انجام می‌شوند. اگر تعداد همه‌ی واژه‌های باقی مانده از سندها برابر t باشد ، آنگاه می‌توان هر سند را به صورت یک آرایه به صورت زیر پنداشت.

$$D = \{f_1, f_2, ..., f_t\} \quad 2\text{-}1$$

در این آرایه f_i فراوانی (تعداد تکرار) واژه i اُم در این سند است. روشن است که فراوانی واژه‌هایی که در این سند نیستند برابر صفر است. اکنون اگر برای همه‌ی سندها چنین کاری انجام شود و همه‌ی آنها در یک ماتریس گذاشته شوند ، آنگاه در این ماتریس تعداد سطرها برابر تعداد سندها (N) خواهد بود و تعداد ستونها برابر تعداد واژه‌ها (t) خواهد بود. به جای به کارگیری فراوانی واژه ، وزنی به هر واژه در هر سند نسبت داده می‌شود که بهتر بتواند نشان دهنده‌ی ارزش آن واژه در آن سند باشد. با انجام این کارها ماتریسی به نام ماتریس سند-واژه ساخته می‌شود. این ماتریس در زیر نشان داده شده است.

	T_1	T_2	$\cdots$	T_t
D_1	W_{11}	W_{12}	$\cdots$	W_{1t}
D_2	W_{21}	W_{22}	$\cdots$	W_{2t}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
D_N	W_{N1}	W_{N2}	$\cdots$	W_{Nt}

2-2

در این ماتریس W_{ik} وزن واژه k اُم در سند i اُم است و N تعداد سندها می‌باشد و t تعداد واژه‌های جداگانه‌ای به کار رفته در همه‌ی سندها است.

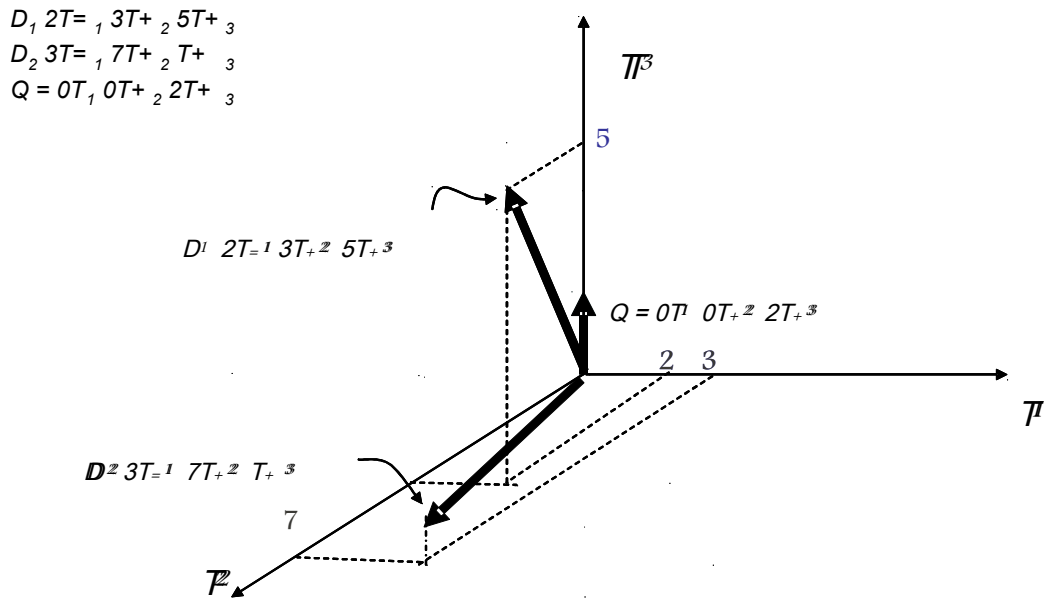
اگر واژه‌های جداگانه‌ی سندها ، یک فضای برداری t بُعدی را بسازند ، آنگاه این واژه‌ها ، بردارهای پایه‌ی (یکه‌ی) این فضای برداری هستند. همچنین سندها را می‌توان بردارهایی در این فضای برداری پنداشت و نقطه‌ی آغازِ بردارِ همه‌ی سندها مبدأ مختصاتِ این فضای برداری است. پیکان(پایان) بردار i اُم (سند i اُم) نقطه‌ای به مختصات زیر است.

$$D_i = [W_{i1} \quad W_{i2} \quad \cdots \quad W_{it}] \quad 2\text{-}3$$

بر روی پرس و جوی کاربر نیز کارهای پیش پردازشی انجام می‌شود. سپس پرس و جو نیز به صورت زیر نوشته می‌شود.

$$Q = [q_1 \quad q_2 \quad \cdots \quad q_t] \quad 2\text{-}4$$

در اینجا باز t تعداد واژه‌های جداگانه در گردایه‌ی سندها است. q_j وزن واژه j اُم در پرس و جو است. پرس و جو نیز همانند سندها به صورت یک بردار در فضای برداری پنداشته می‌شود. در شکل 2_3 یک فضای برداری با تعداد واژه‌های جداگانه t=۳ نشان داده شده است.



شکل 2-3- فضای برداری

برای سادگی کار تنها یک پرس و جو برای سیستم در نظر گرفته می‌شود. گرچه اغلب سیستم بازیابی اطلاعات در یک زمان با پرس و جوهای زیادی روبرو است. می‌توان ماتریسی را نیز برای پرس و جوها پنداشت. در زیر این ماتریس فرضی نشان داده شده است.

$$\begin{array}{c|cccc} & T_1 & T_2 & \cdots & T_t \\ \hline Q_1 & W_{11} & W_{12} & \cdots & W_{1t} \\ Q_2 & W_{21} & W_{22} & \cdots & W_{2t} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ Q_q & W_{q1} & W_{q2} & \ldots & W_{qt} \end{array} \qquad 2\text{-}5$$

به هر حال روند کار چه با یک پرس و جو و چه با دسته‌ای از پرس و جوها یکسان است زیرا هر پرس و جو جداگانه پردازش می‌شود.

2.1.4.1- وزن گذاري واژه

در این بخش به کمک رابطه‌ای زیر وزن واژه به دست می‌آید.

$$W_{ij} = L_{ij} * G_i * N_j \qquad 2\text{-}6$$

L_{ij} برابر وزن محلی واژه i اُم در سند j اُم است. به طور ساده در وزن دهی محلی باید به واژه‌هایی که در یک سند فراوانی بیشتری دارند ، بهای بیشتری داده شود. G_i وزن سراسری

واژه‌ی i اُم در همه‌ی سندها است. بسیاری از روشهای وزن دهی سراسری بـر ایـن بـاور اسـتوار هستند که واژه‌ی با فراوانی کمتر در همه‌ی سندها ارزش بیشتری در بازیـابی اطلاعـات داشـته باشد. N_j ضریب نرمال کننده برای سند j اُم است که بـا تـوجه بـه طـول سـند مقـدار وزن را بگونه‌ای تغییر می‌دهد که در هنگام جستجو به سندهای بزرگ بهای بیشتری داده نشود.

در پرس و جو اغلب وزن واژه با فراوانـی واژه در آن پـرس و جـو برابـر اسـت. در روش دیگری بسادگی وجود یک واژه در پرس و جو با یک نشان داده می‌شود و نبود آن واژه بـا صـفر نشان داده می‌شود. در این روش بردار پرس و جو تنها دربردارنده‌ی صفر و یک خواهد بود.

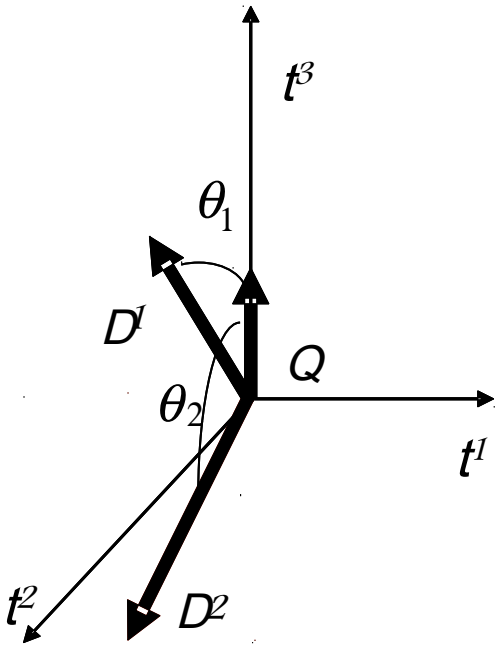
2.1.4.2- تابع شباهت

در روش فضای برداری تا اینجا پرس و جو و همه‌ی سندها به صورت آرایه و بردار تبدیل شدند. اکنون باید به دنبال تابعی بود که بتواند عددی را به هر سند نسبت دهـد کـه نشـان دهنـده‌ی اندازه‌ی نزدیکی پرس و جو به آن سند باشد.

معیارهای سنجش گوناگونی در ریاضی وجود دارند کـه از آن میـان می‌تـوان فاصـله‌ی اقلیدسی را نام برد. فاصله‌ی اقلیدسی معیار شایسته‌ای برای یافتن نزدیکترین سندها به پرس و جو در مدل فضای برداری نیست.

زاویه‌ی میان دو بردار معیار بهتری در فضای برداری است. در این روش دو بردار که بـا هم زاویه‌ی کوچکتری بسازند بیشتر به هم نزدیک انگاشـته می‌شـوند. بـه جـای بـه کـارگیری زاویه‌ی میان دو بردار ، کسـینوس زاویه‌ی میـان دو بـردار معیـار فاصـله گذاشـته شـده اسـت. کسینوس زاویه‌ی میان دو بردار یک تابع نزولی از زاویه در بازه‌ی ۰ تا ۱۸۰ درجه اسـت. بـرای زاویه‌ی صفر (بردار پرس و جو و بردار سند در یک جهت) این تابع مقدار یـک را برمی‌گردانـد و برای زاویه‌ی ۱۸۰ درجه (دو بردار در دو جهت روبروی هم و بیشترین فاصله) ایـن تـابع مقـدار ۱- را برمی‌گرداند. روش محاسبه‌ی این تابع بر پایه‌ی ضرب داخلی دو بردار در زیر نشـان داده شده است.

$$CosSim(d_j, q) = \frac{\vec{d_j} \bullet \vec{q}}{|d_j| \cdot |q|} = \frac{\sum_{i=1}^{t} w_{ij} \cdot w_{iq}}{\sqrt{\sum_{i=1}^{t} w_{ij}^2 \cdot \sum_{i=1}^{t} w_{iq}^2}} \qquad 2\text{-}7$$



شکل 2-4- زاویه‌ی میان سندها و پرس و جو

Q = 0T1 + 0T2 + 2T3

D1 = 2T1 + 3T2 + 5T3 $\quad CosSim(D_1, Q) = \frac{10}{\sqrt{(4+9+25)(0+0+4)}} = 0.81$

D2 = 3T1 + 7T2 + 1T3 $\quad CosSim(D_2, Q) = \frac{2}{\sqrt{(9+49+25)(0+0+4)}} = 0.13$

2.1.4.3- برگزیدن تعداد پاسخها

سیستم بازیابی باید تعدادی از اشاره‌گرها به سندها را در پاسخ به پرس و جوی کاربر بـه کـاربر نشان بدهد. تا اینجا به کمک تابع شباهت اندازه‌ی نزدیکی هر یک از سندها بـه پـرس و جـوی کاربر روشن شد. برگزیدن تعداد سندهایی که باید به کاربر نشان داده شوند ، اغلب بـه دو روش زیر انجام می‌شود.

الف- در این روش سندهایی به کاربر نشان داده می‌شوند که عدد به دست آمده از تابع شباهت برای آن سندها بیشتر از یک آستانه[1] باشند. در این حالت تعداد پاسخهای سیستم ثابت نیست و بسته به اندازه‌ی نزدیکی سندها به پرس و جوی کاربر تغییر می‌کند. بنابراین با این روش ، نمی‌توان مدلهای گوناگون بازیابی اطلاعات را ارزیابی نمود. برگزیدن اندازه‌ی آستانه یکی دیگر از دشواریهای این روش است.

ب- همواره تعداد ثابتی از سندها در پاسخ به هر پرس و جوی کاربر ، نشان داده می‌شوند. این عدد ثابت که نشان دهنده‌ی تعداد پاسخهای سیستم است ، DCVنامیده می‌شود. برای نمونه اگر DCV برابر ۱۵ باشد آنگاه سیستم بازیابی همواره ۱۵ سند با بیشترین نزدیکی به پرس و جو را به کاربر نشان می‌دهد.

این روش بر این باور استوار است که اغلب کاربر همان ۱۵ تا ۲۰ پاسخ نخست سیستم را نگاه می‌کند و دیگر پاسخهای سیستم را نخواهد دید. به همین دلیل سندهایی که در این بازه هستند (برای نمونه از ۱ تا ۱۵ برای DCV=۱۵) بیشتر ارزشمند هستند. ولی برای پرس و جویی که تعدادِ بسیار کمی از سندها (برای نمونه تنها ۵ تا از سندها) به آن مربوط باشند ، پاسخهای پایانی سیستم بازیابی (پاسخهای از ۶ تا ۱۵ در این نمونه) از پرس و جوی کاربر بسیار دور هستند. بزرگترین سودمندی این روش این است که به کمک این روش می‌توان مدلهای گوناگون بازیابی اطلاعات را نسبت به هم ارزیابی نمود [Salehi,2000].

2.2- ارزیابي سیستم بازیابي اطلاعات

مانند هر کار دیگر در سیستمهای هوش مصنوعی باید گردایه‌ی آزمایشی فراهم گردد تا پاسخ سیستم به کمک آن سنجیده شود یا به عبارت دیگر باید برای سیستم هوشمند یک یا چند آزمون آماده شود. برای ارزیابی یک سیستم بازیابی اطلاعات نیز باید یک گردایه‌ی آزمایشی از سندها و پرس و جوها آماده شود. همچنین در این گردایه باید نگاشت میان سندها و پرس و جوها روشن باشد یا به عبارت دیگر باید روشن باشد که چه سندهایی به چه پرس و جوهایی مربوط هستند. به کمک این گردایه ، سیستم بازیابی اطلاعات ارزیابی می‌شود. معیارهای گوناگونی برای ارزیابی یک سیستم بازیابی اطلاعات به کار می‌رود. در اینجا دو معیار که بیشتر در سیستمهای بازیابی اطلاعاتِ تحقیقاتی به کار می‌رود ، آورده شده است. این دو معیار در

[1] 1_ threshold

سیستمهای بازیابی اطلاعات به صورت استاندارد درآمده‌اند. سپس روشی آورده می‌شود که امروزه بیشترِ نمودارهای ارزیابی سیستمهای بازیابی اطلاعات به کمک آن نشان داده می‌شوند.

2.2.1- دقت[1]

دقت یک سیستم بازیابی اطلاعات نشان دهنده‌ای سودمندی فهرستی است که سیستم در پاسخ به پرس و جوی کاربر در خروجی می‌دهد. هر چه تعداد سندهایی که در خروجی سیستم آمده است، بیشتر به پاسخهای واقعی نزدیکتر باشند این معیار بهتر خواهد شد و اگر تعداد سندهایی که به پرس و جوی کاربر ربطی ندارند زیادتر باشد ، دقت سیستم کاهش می‌یابد. دقت برابر است با :

$$precision = \frac{\mathrm{Re}t\,\mathrm{Re}l}{no\,\mathrm{Re}t} \quad 2\text{-}8$$

در این رابطه RetRel برابر است با تعداد سندهای بازیابی شده و مربوط و noRet برابر است با تعداد همه‌ای سندهای بازیابی شده.

2.2.2- یادآوری[2]

یادآوری نشان دهنده‌ای کامل بودن فهرست خروجی سیستم در پاسخ به پرس و جوی کاربر است. هر چه تعداد سندهایی که در خروجی سیستم آمده است بیشتر باشد این معیار بهتر خواهد بود. این معیار اندازه‌ی پوشا بودن سیستم را نشان می‌دهد. یادآوری برابر است با :

$$\mathrm{Re}call = \frac{\mathrm{Re}t\,\mathrm{Re}l}{no\,\mathrm{Re}l} \quad 2\text{-}9$$

در این رابطه noRel برابر است با تعداد سندهای مربوط در گردایه.

یادآوری نشان می‌دهد که جویشگر[3] در پیدا کردن سندهای مربوط چقدر خوب کار می‌کند. هنگامی که همه‌ی سندهایی که مربوط هستند بازیابی شوند یادآوری صد در صد خواهد بود. در تئوری برگرداندن همه‌ی سندهایی که مربوط به پرس و جوی کاربر هستند (در

[1] 1_ precision

[2] 2_ Recall

[3] 1_ search engine

گردایه‌ای بازیابی اطلاعات) و به دست آوردن یادآوری شایسته ، ساده است. ولی در عمل برای پرس و جوهای گوناگون با معناهای چندگانه‌ی واژه‌ها در یک زبان و نیز معناهای گوناگون واژه‌ها در کنار یکدیگر و در ترکیبهای گوناگون و دیگر عاملهای اثر گذار ، به دست آوردن یادآوری شایسته بسیار سخت است [جباری‌فر،۱۳۸۰].

شاید سیستمی یادآوری بالا داشته باشد (تعداد سندهایی که به پرس و جوی کاربر ربط دارند وبازیابی شده‌اند ، زیاد باشد) ، ولی همزمان تعداد بسیار زیادی پاسخهای بی‌ربط نیز بازیابی شده باشند که دقت سیستم را پایین می‌آورد. این دو معیار (یادآوری و دقت) اغلب در برابر هم قرار گرفته‌اند و افزایش یکی ، دیگری را کاهش خواهد داد. بنابراین یادآوری به تنهایی سنجش خوبی برای یک جویشگر نیست.

2.2.3- نمودار یادآوري-دقت

یادآوری و دقت بگونه‌ای که تا کنون تعریف شد ، درباره‌ی جای پاسخها و مرتبه‌ی آنها اطلاعاتی نمی‌دهد. در اینجا دقت و یادآوری بگونه‌ای تعریف می‌شود که دربردارنده‌ی جای پاسخها نیز باشد [Salehi,2000].

به کمک فهرست مرتب شده‌ای از اشاره‌گرها به سندهایی که سیستم در پاسخ به پرس و جوی کاربر یافته است ، معیار دیگری تعریف می‌شود. سالانه کنفرانسی به نام TREC[1] در دنیا برگزار می‌شود که به ارزیابی و رتبه بندی جویشگرها می‌پردازد. این معیار یکی از معیارهای پایه‌ی تعریف شده در این کنفرانس برای ارزیابی جویشگرها است. به جای به دست آوردن یک یادآوری به کمک پاسخهای مرتب شده‌ی سیستم فهرستی از یادآوریها آماده می‌شود یا به عبارت دیگر برای هر یک از پاسخها یک یادآوری به دست می‌آید. یادآوری در سند i اُم (از سندهای بازیابی شده و دارای رتبه‌ی i اُم در این سندها بر پایه‌ای عدد شباهت) برابر است با :

$$r_i = \frac{\mathrm{Re}t\,\mathrm{Re}l_i}{no\,\mathrm{Re}l} \quad 2\text{-}10$$

در این رابطه $\mathrm{Re}t\,\mathrm{Re}l_i$ تعداد سندهای مربوط و بازیابی شده تا سند i اُم را نشان می‌دهد. $no\,\mathrm{Re}l$ تعداد همه‌ای سندهای مربوط و بازیابی شده تا سند i اُم را نشان می‌دهد. دقت در سند i اُم برابر است با :

[1] 2_ Text REtrieval Conference

$$p_i = \frac{\mathrm{Re}t\,\mathrm{Re}l_i}{i} \quad 2\text{-}11$$

همانگونه که دیده می‌شود هر یادآوری r_i متناظر یک دقت p_i است و دقتها با افزایش i (تعداد سندهای بازیابی شده تا این سند) کاهش می‌یابند. تعداد یادآوریها و دقتهای به دست آمده به کمک دو رابطه‌ای 2-10 و 2-11 ، به تعداد پاسخهای سیستم بستگی دارند. برای به دست آوردن تعداد ثابتی از این عددها ، اغلب یازده نقطه‌ای یادآوری به نام نقطه‌های استاندارد در نظر گرفته می‌شوند. میان صفر تا یک به ده بخش می‌شود و یازده عدد (0،0.1،...،0.9،1) ، نقطه‌های ثابت در نظر گرفته می‌شوند. از این پس این عددها را (نقطه‌های استاندارد یادآوری) با x_k (k=0,1,...,10) نشان می‌دهیم. اکنون باید برای هر کدام از این یازده نقطه‌ای یادآوری ، یک دقت به دست آورده می‌شود.

به کمک رابطه‌ای 2-12 دقت k اُمین نقطه‌ای استاندارد یادآوری به دست آورده می‌شود.

2-12

$$xp_k = \max\{p_j | x_{k-1} \le r_j \le x_k \,, 1 \le \mathrm{j} \le \mathrm{noRet}\},\ 0 \le \mathrm{k} \le 10 \,,\ \mathrm{x_k} \in \{0,0.1,...,0.9,1\}$$

در رابطه‌ای پیش ، $no\,\mathrm{Re}t$ تعداد سندهایی است که سیستم در پاسخ به پرس و جوی کاربر بازیابی کرده است. برای نمونه برای به دست آوردن دقت متناظر با نقطه‌ای استاندارد یادآوری 0.2 (x_2) کارهای زیر انجام می‌شود. از میان دقتهای به دست آمده به کمک فرمول 2-11 ، دقتهایی که یادآوری متناظر آنها از 0.2 کوچکتر و از 0.1 (x_1) بزرگتر هستند ، برگزیده می‌شوند. بزرگترین دقت از میان این دقتهای برگزیده شده ، دقت متناظر با یادآوری 0.2 است.

اغلب برای ارزیابی یک سیستم بازیابی اطلاعات بیش از یک پرس و جو با پاسخ از پیش آماده (سندهای مربوط به این پرس و جو در گردایه‌ای بازیابی) به سیستم داده می‌شود. به کمک رابطه‌های نوشته شده تا کنون برای هر پرس و جو یک دسته دقت (متناظر با نقطه‌های یادآوری) به دست آورده می‌شود. سپس در هر کدام از نقطه‌های استاندارد یادآوری از دقتهای متناظر آنها (که از هر پرس و جو به دست آمده‌اند) میانگین گرفته می‌شود. این دقت میانگین گیری شده ، «شبه دقت» نامیده می‌شود. رابطه‌ای به دست آوردن شبه دقت در رابطه‌ای 2-13 نوشته شده است.

$$\overline{p}(x_k) = \frac{\sum_{l=1}^{noQuery} xp_k(l)}{noQuery} \quad 2\text{-}13$$

در این رابطه $noQuery$ تعداد پرس و جوها است و $xp_k(l)$ دقت به دست آمده به کمک رابطه‌ای 2-12 برای l امین پرس و جو در نقطه‌ای استاندارد یادآوری k اُم است. به این

ترتیب برای هر نقطه‌ای استاندارد یادآوری x_k (صفر تا یک به ده بخش) یک شبه دقت به دست می‌آید. اکنون می‌توان سیستمهای گوناگون بازیابی اطلاعات را به هم مقایسه نمود.

2.3- نمایه گذاري معاني پنهان

تا به اینجا یکی از روشهایی که بسیار در بازیابی اطلاعات به کار گرفته می‌شود، بررسی گردید. داده‌ها و سندها به آرایه‌ها تبدیل شدند و ادامه‌ی کار نیز با روشهای ریاضی انجام گرفت. ولی بازیابی اطلاعات نوشتاری هنوز دشواریهای فراوانی دارد.

هنوز دو دشواری بزرگ وجود دارد :

الف ــ کاربران بیش از آن که واژه‌های درون یک نوشته را در نظر داشته باشند به معنای نوشته توجه دارند.

ب ــ واژه‌های یک نوشته به طور جداگانه دربردارنده معنای جمله نیستند.

اگر کاربر به دنبال نوشته‌ای در زمینه «آموزش نقاشی به کودکان» باشد ؛ او اغلب می‌خواهد که اگر در نوشته‌ای به جای کودکان ، بچه‌ها یا دبستانیها باشد سیستم بازیابی اطلاعات بتواند آن نوشته را نیز بیابد. همچنین اگر به جای آموزش ، یاد دادن یا تعلیم در نوشته‌ای باشد سیستم آن نوشته را هم در پاسخ داشته باشد. زیرا بیشتر به دنبال هر نوشته‌ای است که همان معنا را در برداشته باشد. به حالتی که چند واژه یک معنا را برساند ، synonymy گفته می‌شود. افزون بر این یک واژه در جاهای گوناگون معناهای گوناگونی دارد و بیشتر وقتها معنای یک واژه وابسته به واژه‌های پس و پیش آن است. همچنین معنای واژه به جمله‌ای که در آن قرار دارد ؛ نیز بستگی دارد. همچنین به جمله‌های دیگری که معنای این جمله به آنها وابسته است و موضوعی که نوشته در آن زمینه است؛ نیز بستگی بسیاری دارد. به این حالت که واژه‌ای چندین معنا را برساند polysemy گفته می‌شود. دو دشواری synonymy و polysemy از پایه‌ای‌ترین دشواریهایی هستند که هر طرح نمایه گذاری معنایی باید به گونه‌ای به آنها پاسخ دهد [salehi,2000]. بنابراین زمانی که به دنبال پاسخ کاربر هستیم ، جستجوی واژه به واژه نمی‌تواند پاسخ دقیقی را برای درخواست کاربر بیابد.

با داشتن لغت‌نامه‌ای از واژه‌های هم معنی می‌توان از دشواری نخست کاست. ولی ساختن چنین لغت‌نامه‌ای چندان هم ساده نیست. همچنین هر روز می‌بینیم که واژه‌های جدید به زبان افزوده می‌شوند یا معنای واژه‌ای سراسر دگرگون می‌شود و معنای واژه به واژه‌های دیگر

نزدیک می‌شود. بنابراین باید چاره‌ی دیگری اندیشید. برای برداشتن دشواری دوم ، نوشته باید تحلیل معنایی شود. که کار بسیار سنگینی است. تحلیل معنایی برای بیشتر زبانها هنوز در دست آزمایش است. و تا دستیابی به پاسخی شایسته راه درازی مانده است.

دقت فرمایید که در نمایه گذاری معانی پنهان ترانهاده‌ی ماتریس سند-واژه به کار گرفته می‌شود. به عبارت دیگر ماتریس واژه-سند به کار گرفته می‌شود.

2.3.1- پایه‌های ریاضی

روش نمایه گذاری معانی پنهان در بازیابی اطلاعات روشی در جبر خطی به نام تجزیه کننده‌ی مقدار ویژه[1] را به کار می‌گیرد. این روش یکی از روشهای فاکتورگیری است که بر پایه تئوری طیفی[2] می‌باشد. روش SVD چه در کاربرد و چه در تئوری به گذشته‌ای به نسبت دور برمی‌گردد. می‌توان جای پای این روش را در نوشته‌های بلترامی یافت [Nakos,1998]. در این روش هدف تجزیه‌ی ماتریس A به ابعاد m*n به شکل $U\sum V$ است. که U یک ماتریس m*m است و V ماتریسی n*n می‌باشد. ماتریسهای U و V هر دو متعامد هستند. ماتریس متعامد ماتریسی است که

۱ـ مربعی باشد.

۲ـ ستونهای ماتریس متعامد یکه[3] هستند.

تعدادی بردار (r_i) در صورتی متعامد یکه هستند که

$$\forall i, \|V_i\| = 1 \text{ and } \forall i, j \text{ where } i \neq j, V_i \bullet V_j = 0 \quad 2\text{-}14$$

ماتریس $\sum$ یک ماتریس m*n قطری است. این ماتریس چهار بخش دارد که سه بخش آن ماتریسهای صفر هستند و بخش سمت چپ بالایی آن یک ماتریس قطری است که عنصرهای روی قطر آن عددهای مثبتی هستند که ترتیب نزولی دارند.

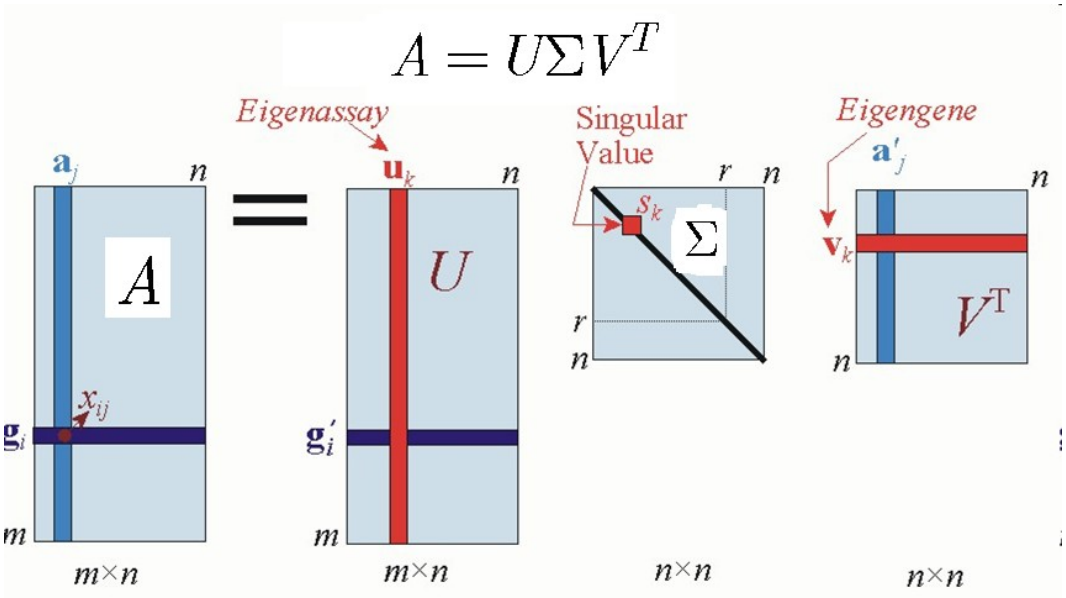
[1] 1_ Singular Value Decomposition

[2] 2_ spectral theorem

[3] 3_ orthonormal

$$A = U \Sigma V^T$$
$$\Sigma = \begin{bmatrix} D & 0 \\ 0 & 0 \end{bmatrix}$$
$$D = \begin{bmatrix} \sigma_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sigma_r \end{bmatrix} \qquad 2\text{-}15$$
$$\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r > 0 \ , \ r \leq m,n$$

اکنون با برداشتن صفرها از ماتریس $\sum$ (گذاشتن $\Sigma = D$) و همچنین با برداشتن تعدادی از σ_i_های پایانیِ کوچک از ماتریس $\sum$ ، وابستگی میان واژه‌ها بیشتر روشن می‌شود. در شکل زیر این روند نشان داده شده است.



شکل 2-5- روند برداشتن تعدادی از σ_i_های پایانی کوچک

U و V ماتریسهای دربردارنده‌ی واژه‌ها و سندها هستند و ماتریس $\sum$ مقدارهای ویژه ماتریس A را داراست. نقطه‌های سایه‌دار در ماتریسهای U و V و خط قطری در $\sum$ به کمک مقدار بهینه K تعیین می‌شود. U_k یک ماتریس با ابعاد t*k است که ستونهایش k ستون نخست U هستند. V_k یک ماتریس d*k است که سطرهایش k سطر نخست ماتریس V است و $\sum_k$ یک ماتریس k*k است که عناصر قطرش ، k مقدار بزرگ نخست از مقدارهای ویژه

ماتریس A می‌باشند. مرتبه‌ای[1] ماتریس r_A برای ماتریس A ، تعداد مقدارهای ویژه غیر صفر آن ماتریس می‌باشد با انجام دادن کارهای بالا ماتریس زیر به دست می‌آید.

$$A_k = U_k \sum_k V_k \quad 2\text{-}16$$

باید مقدار k شایسته‌ای برای محاسبه A_k برگزیده شود. همانطور که کاهش k می‌تواند نوفه[1] را بیشتر برطرف کند، اگر تعداد بسیار کمی از مقدارهای ویژه نگه داشته شوند ؛ اطلاعات باارزشی را از دست می‌دهیم [Dumais,1991]. بهترین گزینه برای k زمانی است که اگر مقدار k را افزایش دهیم تغییرِ کارآیی سیستم ناچیز باشد. در این حالت عنصرهای اصلی پایگاه داده کاهش نیافته و یا برداشته نمی‌شوند.

2.3.2- چگونگي عملكرد

نمایه گذاری معانی پنهان با یافتن وابستگی معنا میان واژه‌ها تا اندازه‌ای دشواری مقایسه را کم می‌کند [Buckley,1996]. تجزیه کننده‌ای مقدار ویژه تابعی است که برای تخمین معنای واژه‌ها در سند به کار گرفته می‌شود. در این روش ، بازیابی به کمک ماتریس SVD انجام می‌شود [Glub,1996]. یکی از عیبهای این روش ، نیاز آن به محاسبه‌های بسیار زیاد است که هزینه‌ای زمانی زیادی را (و همچنین هزینه‌ای فراهم نمودن رایانه‌های ویژه با توانایی‌های بالا برای گردایه‌های بزرگ) برای به دست آوردن ماتریس SVD می‌خواهد [Berry,1996]. روش LSI در مدل فضای برداری کار می‌کند و ماتریس دیگری ایجاد می‌کند که دربردارنده اطلاعات ماتریس واژه-سند است [Berry,1995]. تجزیه کننده‌ای مقدار ویژه برای حل مسأله مربعهای خطی کوچک و تخمین درجه‌ای ماتریس و بررسی همبستگی ذاتی به کار گرفته می‌شود [Berry,1992]. این روش ریاضی در سال ۱۹۸۸در یافتن معانی پنهان به کار گرفته شد.

هدف از به کارگیری روش نمایه گذاری معانی پنهان جایگزین کردن نمایه‌های بر پایه‌ی واژه با نمایه‌های بر پایه‌ی معنا است. در این رهیافت نگاشتی از ماتریس واژه-سند در فضای برداری به فضایی دیگر انجام می‌شود و این کار به کمک تابع تجزیه کننده‌ی مقدار ویژه (SVD) انجام می‌شود [Arms,2002].

---

[1] 1_ rank

[1] 2_ noise

روش نمایه گذاری معانی پنهان بر پایه‌ی این باور بنا نهاده شده است کـه بـا برداشـتن مقدارهای ویژه‌ای کوچک از ماتریس واژه-سند ، رابطه‌ی معنـایی میـان واژه‌هـای درون سـندها آشکار خواهد شد. برای نمونه اگر در نوشـته‌ای واژه‌ی car بـه کـار رفتـه باشـد و در نوشـته‌ی دیگری واژه‌ی auto آنگاه با به کار گرفتن روش LSI و به کمک دیگر واژه‌هـای همسـان در دو نوشته مانند driving ، gasoline ، highway رابطه‌ی میان دو واژه‌ی car و auto نشان داده می‌شود. اکنون چه در درخواست کاربر واژه‌ی car به کار گرفته شود و چه واژه‌ی auto ، سیستم بازیابی اطلاعات پاسخ یکسانی به هر دو واژه خواهد داد [Belew,2000].

فصل سوم

دشواريهاي بازيابي اطلاعاتِ نوشتاريِ فارسي

3- دشواريهاي بازيابي اطلاعاتِ نوشتاريِ فارسي

بازیابی اطلاعات یکی از دانشهای به نسبت نوین است که امروزه تحقیقات بسیاری (چه تجـاری و چه دانشگاهی) بر روی آن انجام می‌شود.

بازیابی اطلاعات نوشتاری فارسی هنوز در نخستین مرحله‌های خود است و به همیـن دلیـل بـا دشواریهای فراوان دست به گریبان است. برخـی از ایـن دشـواریها بـرای هـر کـار رایـانه‌ای در زمینه‌ای فارسی وجود دارد. گرچه سالها است که بسیاری از پژوهشگران رایانه از ایـن دشـواریها آگاهند؛ ولی گامهای کوچکی برای برطرف کردن آنها برداشته شده است.

شاید یکی از بزرگترین دردسرها ، برای انجام این پایان نامه این پرسـش بـود کـه ، بـا وجود انبوه دشواریها از کجا باید کار را آغاز نمود؟

می‌توان گفت که پاسخ به این پرسش زمان زیادی از انجام این پایان نامه را گرفـت. در دسترس نبودن پایه‌ای‌ترین بخشهای مورد نیـاز بـرای بازیـابی اطلاعـات نوشـتاری فارسـی بـه سردرگمی بیشتر می‌انجامید. همچنین دشواریهای این پایان نامه بسیار در هم تنیده و بـه هـم وابسته بودند. بزرگترین دلگرمی برای ادامه‌ی کار، برخورد شایسته‌ای بزرگوارانی بـود کـه بـرای یافتن پاسخ پاره‌ای از پرسشها از آنها کمک گرفته شد. کوشش گردید که در بخش سپاسگزاری نام همه‌ی این بزرگواران آورده شود.

3.1- دشواريهاي نوشتاري فارسي

زبان فارسی دربردارنده گنجینه‌ای بزرگی از زیباترین سروده‌ها و داسـتانها اسـت. زبـان فارسـی یکی از پربارترین زبانهای دنیا است. کتابهایی چون مثنـوی معنـوی ، دیـوان حـافظ ، رباعیّـات خیام و ... به زبانهای گوناگون گیتی برگردانده شده و بارها چاپ شـده‌اند. برتریـن ویژگـی ایـن

نوشته‌ها ، انسانی بودن آنها است بگونه‌ای که همه‌ی انسانها گرایشی درونی به این نوشته‌ها دارند.

متأسفانه این درخت تنومند امروزه نیاز به توجه بیشتری دارد زیرا برای دنیای نوین آماده نشده است. پیرایش و ویرایش بر روی دیگر زبانهای دنیا خیلی پیشتر از این آغاز شده است. ساده کردن قاعده‌ها ، کم کردن قاعده‌های پیچیده و استثناها در زبان روزمره (نه زبان ادبی)، یکسان کردن گفتار و نوشتار روزمره ، به کارگیری تعداد کمی واژه و اصطلاح ، گسترش استانداردهای آماده شده برای زبان از کارهایی است که بر روی بسیاری از زبانها انجام شده است. استادان زبان انگلیسی و زبان‌شناسان ، بسیاری از قاعده‌های این زبان را پیراسته‌اند و یادگیری و به کارگیری این زبان را ساده نموده‌اند. برای نمونه در نوشتار امروزی انگلیسی کمتر حرفها به هم چسبیده نوشته می‌شوند و واژه‌ها و اصطلاحهای کمی ، بویژه در نوشته‌های علمی ، به کار گرفته می‌شود. ویراشهای انجام شده در زبان انگلیسی بسیار بر کارهای رایانه‌ای ، که بر پایه‌ای زبان انگلیسی هستند ، اثر داشته است و به پیشرفت نرم افزارهای رایانه‌ای کمک نموده است. پیرایشهایی که در زبان انگلیسی انجام شده است ، بسیاری از پیچیدگیهای ساخت نرم افزارهایی برای این زبان را کاسته است و به نوبه‌ی خود ساخت نرم افزار رایانه‌ای گسترش استاندارد آن زبان را در پی داشته است.

بیشتر نمونه‌هایی که در این بند برای نشان دادن دشواریهای نوشتار فارسی آورده شده است ، برگرفته از کتابهای درسی آموزش و پرورش و یا کتابهایی از استادان و دبیران آیین نگارش فارسی است. چگونگی نگارش در این کتابها را می‌توان استاندارد نگارشِ فارسی قلمداد کرد. نشان دادن پیچیدگیها و ناهماهنگیها در این کتابها نمایانگر دشواریهای پایه‌ای نوشتار فارسی است. اگر بنا بود برای یافتن دشواریهای نوشتار فارسی از نوشته‌های روزمره و یا از سایتهای فارسی در شبکه‌ای جهانی نمونه آورده شود ، آنگاه با انبوهی از دشواریهای بسیار پیچیده‌تر برخورد می‌شد. کافی است به چند سایت فارسی در شبکه‌ای جهانی سر زده شود تا بخوبی درستی این گفته روشن شود. بنابراین دشواریهایی که در این بند آورده می‌شوند ، دشواریهای پایه‌ای نوشتار فارسی هستند. می‌توان گفت که هر کار رایانه‌ای در زمینه‌ای نوشتار فارسی با این دشواریها روبرو است. خوشبختانه روز به روز نوشته‌های فارسی ساده‌تر و خواناتر از گذشته نوشته می‌شوند و استانداردهای نگارش فارسی بیشتر رعایت می‌شود. به این ترتیب از دامنه‌ی دشواریها کاسته می‌شود. شاید بتوان گفت که وجود و هجوم روزافزون واژه‌های زبانهای دیگر در زبان فارسی از آن دست دشواریهایی است که برخلاف دیگر دشواریهای زبان فارسی روند رو به افزایشی دارد. در بند ریشه‌یابی به بخش کوچکی از این دشواری پرداخته شده است.

3.1.1- دستور زبان فارسي

«دستور زبان فارسی قواعدی است که بدان درست گفتن و درست نوشتن را بیاموزند آنچه بدان مقاصد خود را بیان کنند کلام (و سخن) نامند و کلام مرکب از کلمات و کلمه مرکب از حروف باشد.» [رشید یاسمی،۱۳۶۸]

تعریف بالا از دستور زبان برگرفته از (شاید) نخستین کتاب رسمی درباره‌ی دستور زبان است. این کتاب پایه‌ی آموزش دستور زبان فارسی برای نزدیک به پنجاه سال بوده است و بارها چاپ شده است. ولی امروزه دستور زبان فارسی را بیشتر از دیدگاه زبان شناسی می‌نگرند. بیشتر آنچه که هست ارزش دارد و نه آنچه که باید باشد.

«کوشش شده است که دستور زبان دوره‌ی دبیرستان ، بر پایه‌ی نظریه‌ی ساخت گرایی نوشته شود و در مواردی نیز نظریه‌های دیگر همچون نقش گرایی و گشتاری که در تحلیل قضیه‌های دستوری بهتر بوده‌اند آن نظریّه‌ها به کار گرفته شود و همچنین از دستور زبان سنتی نیز دور نشود.» [حق شناس،۱۳۷۷]

برای دیدن اثر این دگرگونیها در دستور زبان فارسی ، برای نمونه در زیر تعریف فعل از چند کتاب از سالهای گوناگون آورده شده است. دقت فرمایید که تغییر در تعریفِ فعل ، نوع نگرش به فعل را تغییر می‌دهد و این تغییر به نوبه‌ی خود اثر بسزایی در کارهای رایانه‌ای دارد.

«یکی از اقسام سخن فعل (کُنش) است. فعل در اصطلاح کلمه‌ای است که دلالت کند بر حصول معنی مصدر و حدوث آن از ذاتی در یکی از زمانهای سه‌گانه : گذشته ، حال ، استقبال.» [بهار،۱۳۴۲]

«در هر گزاره یک جزء اصلی وجود دارد که اگر آن را برداریم جمله معنی و مفهومی نخواهد داشت این جزء اصلی را فعل می‌نامیم. فعل چهار مفهوم دارد: ۱- مفهوم کار یا حالت ۲- مفهوم شخص ۳- مفهوم جمع و اِفراد ۴- مفهوم زمان» [مولوی،۱۳۵۹]

«فعل کلمه‌ای است که بر انجام دادن کار و روی دادن امری یا داشتن حالت و صفتی در یکی از زمانهای (ماضی = گذشته) ، (مضارع = حال) و (مستقبل = آینده) دلالت کند.» [رزمجو،۱۳۷۲]

«در دستور ساختاری توصیف بر پایه ساخت و صورت است ؛ برای نمونه در تعریف فعل به جای اینکه بگوییم : فعل کلمه‌ای است که بر انجام دادن کاری یا روی دادن و پذیرفتن حالتی در زمان دلالت کند ، گفته می‌شود : فعل کلمه‌ای است که دارای شناسه باشد.» [سمیعی گیلانی،۱۳۸۰]

3.1.2- تغییر در استاندارد نگارش

یکی از تغییرهایی که در این چند سال در نگارش فارسی به وجود آمده است ، تغییر شکل کسره‌ی اضافه پس از «ه» است. در گذشته با گذاشتن همزه (ٔ) بر سرِ «ه» این کار انجام می‌شد ولی امروزه برای نشان دادن کسره‌ی اضافه پس از «ه» ، «ی» به کار گرفته می‌شود. در نوشته‌های زیر که از چندین کتاب درسی از سالهای گوناگون آورده شده است این دگرگونی بخوبی دیده می‌شود.

نوشته‌ی زیر از خواجه عبدالله انصاری است که نگارش آن برگرفته از [رزمجو،۱۳۷۲] می‌باشد.

«ای عزیز! در رعایت دلها کوش و عیب کسان می‌پوش و دل به دنیا مفروش. بدان که خدای تعالی در ظاهر ، کعبه بنا کرده است که او از سنگ و گل است و در باطن کعبه ای ساخته که از جان و دل است . آن کعبه ، ساخته ابراهیم است و این کعبه نظرگاه ربّ جلیل است. آن کعبه منظور نظر مؤمنان است و این کعبه نظرگاه خداوند رحمان. آن کعبهٔ حجاز است و این کعبهٔ راز است.»

نوشته‌ی زیر درباره «پاسداری از زبان فارسی» از ناصر ایرانی است که نگارش آن برگرفته از [شریعت،۱۳۸۲] است.

«زبان فارسی به اندیشه‌ی ما شکل داده است و دستِ کم ، با برخی از ظریف‌ترین و ژرف‌ترین جنبه‌های رمزی و معنوی آن یکی شده است . از این روست که اگر زبده‌ترین و خلّاق‌ترین فارسی زبانان و عربی دانان یا فارسی زبانان و انگلیسی دانان جهان گرد آیند و همه‌ی توانایی‌های زبانی خویش را روی هم بریزند ، نخواهند توانست محتوای اندیشه‌ی اندیشمندانی چون حافظ یا مولانا را به درستی در ظرف زبانی دیگر بریزند.»

هنوز در برخی از کتابهای درسی کسره‌ی اضافه پس از «ه» به شکل همزه نوشته می‌شود. برای نمونه نوشته‌ی زیر گوشه‌ای از زندگی جبار باغچه‌بان آموزگار دلسوز است. نگارش آن برگرفته از [گروه مؤلفان،۱۳۷۹] است. احتمالاً این کتاب تغییر نکرده است که همان شکل گذشته در آن رعایت شده است و البته از برخی جنبه‌ها جای سپاسگزاری است که هنوز برخی از تغییرهای نوین (نه گونه‌ی نگارشی آن) به این کتاب راه نیافته است و کتاب هنوز زیبایی (معنایی و ...) ویژه‌ای دارد.

«آن روز وقتی که پسرک کر ولال را در باغچهٔ اطفال نگه داشتم ، اندیشیدم : چگونه می‌توان به کودکی که نه می‌شنود و نه حرف می‌زند ، خواندن و نوشتن آموخت؟ شنیده بودم

که کسی در اروپا الفبایی اختراع کرده است که با آن ، کودکان کر و لال را باسواد می‌کنند و به آنان حرف زدن می‌آموزند. اگر دیگران چنین کاری کرده‌اند ، چرا من نتوانم بکنم. مگر نـه ایـن است که انسان هر کاری را که بخواهد می‌تواند با کوشش فـراوان بـه انجـام رسـاند. از آن پـس شبها و روزهای بسیاری را در کار ابداع الفبای کر و لالها گذراندم تا به مقصـود رسـیدم و چنـد کودک کر و لال دیگر را نیز در باغچهٔ اطفال پذیرفتم.»

3.1.3- ناهماهنگي در نگارش

ناهماهنگی در نگارش واژه‌های فارسی (عربی ، انگلیسی ، ترکی ، مغولی و فرانسوی یا واژه‌هـای دیگر زبانهای بیگانه‌ی آورده شده به زبان فارسی) گرچه ، بر اثر همسانی نگارش کتابهای درسی و مرکزهای نشر و مجله‌های علمی و ادبی تا حد زیادی کـاهش یـافته اسـت ولـی در برخـی از نوشته‌ها و کتابهای غیر درسی این ناهماهنگی دیده می‌شود [احمـدی بیرجنـدی،۱۳۶۸]. رواج صنعت چاپ و ماشینهای تحریر و رایانه‌ها و شتاب زدگی در نوشـتن ، بی‌دقـتی را در درسـت و زیبا نوشتن دامن زده است. همچنین داده‌ها و صفحه‌های روی شبکه‌ی جهانی که برخـی بایـد روزانه تغییر کنند ، این دردسر را چندین برابر کرده‌اند.

3.1.3.1- رعايت نكردن قاعده‌هاي آيين نگارش

سالهاست که برخی قاعده‌ها در آیین نگارش سفارش می‌شود ولی هنوز بسیاری ، ایـن قاعـده‌ها را رعایت نمی‌کنند. در این بند چند نمونه از رعایت نکردن قاعده‌های آیین نگارشی آورده شـده است.

الف- تنوین در واژه‌های عربی به کار می‌رود و قانونهای ویـژه صـرفی و نحـوی آن زبـان بـر آن جاری است. به کار بردن آن در واژه‌های فارسی و غیر عربی به هیـچ وجـه درسـت نیسـت . بـه واژه‌های زیر توجه کنید.

ناچاراً(بناچار) ، جاناً(جانی) ، زباناً(زبانی) ، ... .

اگر کوشش شود به جای کلمه‌های تنوین‌دار عربی، برابرهای فارسی آنها ـ تا آنجـا کـه ممکن است و به شیوایی گفتار لطمه نزند ـ به کار برده شود، بهتر و نوشته ، به فارسی نزدیکـتر خواهد بود [شکری،۱۳۷۲].

ب- استقلال هر واژه به خاطر معنی و نقش دستوری جداگانه‌ای که در جمله دارد ، در نوشتن باید حفظ شود ، یعنی هر واژه جدا از واژه‌های دیگر نوشته شود تا ایجاد اشتباه نکند [دفتر تحقیقات و برنامه ریزی تربیت معلم،۱۳۶۶]. بنابراین چسباندن «به» پیش از اسم یا «می» پیش از فعل نادرست است. برای نمونه «بتهران نمیروم» باید به شکل «به تهران نمی‌روم» نوشته شود. دو نمونه از رعایت نکردن قاعده‌های آیین نگارش نوین از دو کتاب به نسبت قدیمی که خود درباره‌ی آیین نگارش هستند ، در زیر آورده شده است.

نگارش نوشته‌ی زیر برگرفته از [شریعت،۱۳۷۰] است که از داستان باباکوهی این کتاب برگزیده شده است. در این نوشته زیرِ واژه‌هایی که بر پایه‌ی آیین نگارش نوین نادرست هستند ، خط کشیده‌ام.

«خانم بزرگ اختر را سرزنش کرد و همینکه فهمید پای منهم در این کار رفته ، فریاد و فغانش بلند شد که این آخوند باید از خانه ما برود! عوض اینکه بیچه‌ها درس پرهیز و عصمت بدهد ، برایشان کاغذ عاشق و معشوقی مینویسد البته اختر عروس من است اما زنیکه پیش از عقد ، بشوهرش کاغذ خاطر خواهی بنویسد ، بچه درد میخورد ، این طفل معصوم را آخوند خراب کرده!»

نوشته‌ی زیر گوشه‌ای از نامه‌ی چارلی چاپلین به دخترش می‌باشد که نگارش آن برگرفته از [سخاوت،۱۳۴۴] است. در این نوشته نیز زیرِ واژه‌هایی که بر پایه‌ی آیین نگارش نوین نادرست هستند ، خط کشیده‌ام.

«هنر پیش از آن که دو بال دور پرواز بانسان بدهد ، اغلب دو پای او را نیز می‌شکند. وقتی بآنجا رسیدی که یک لحظه ، خود را برتر از تماشاگران رقص خویش بدانی همان لحظه صحنه را ترک کن ... .»

3.1.3.2- حرکت گذاری در نوشتار فارسی

در خط فارسی افزون بر حرفهای الفبا نُه نشانه خطی دیگر نیز به کار می‌رود. این نشانه‌ها ـَ ـِ ـُ ـً ـٍ ـٌ آ هستند. کاربرد این نشانه‌ها کم است؛ زیرا در خط فارسی حرکت گذاری به کار برده نمی‌شود. در نوشتن واژه‌ها از میان نشانه‌های نه‌گانه بالا مد، تشدید، تنوین نصب (ـاً) بیشتر کاربرد دارند . تنوین رفع و جر (ـٍ) تنها در واژه‌های عربی رایج در فارسی به کار می‌رود و دیگر نشانه‌ها را در جاهایی به کار می‌بریم که رعایت نکردن آنها ابهام و بدفهمی به وجود می‌آورد [حق شناس،۱۳۷۷].

این نشانه‌ها نیز دشواری دیگری در ناهماهنگی در نگارش فارسی بـه وجـود آورده‌انـد، مانند «رَفت» و «رُفت» که تنها فرق آنها در ـَ و ـُ است که بر سر «ر» گذاشته شده است . ولـی رُفتگر را بیشتر در نگارش بدون ـُ می‌گذارند و این کار ریشه یابی را سختتر می‌کند.

خوشبختانه امروز با گسترش رایانه‌ها و در دسترس بودن نرم افزارهای توانمند نگـارش و ویرایش و دامنه‌ی بزرگی از نویسه‌ها که این نرم افزارها پشـتیبانی می‌کننـد از ایـن دشـواری کمی کاسته شده است و گذاشتن این نشانه‌ها نیز ساده‌تر گشته است. گرچه هنوز نمی‌توان بـه درستی گفت که کجا باید این نشانه‌ها رعایت شوند. چنین می‌نمایـد کـه امـروزه گـاهی در بـه کاربردن این نشانه‌ها زیاده روی می‌شود. بـرای نمـونه بخشـی از یـک غـزل حـافظ برگرفتـه از [شورای تحقیقات دانشگاه پیام نور،۱۳۷۸] در زیر آورده شده است.

دوش دیدم که ملایک درِ میخانه زدند

گِلِ آدم بسرشتند و به پیمانه زدند

ساکنانِ حرم ستر و عفافِ ملکوت

با منِ راه نشین بادهٔ مستانه زدند

آسمان بارِ امانت نتوانست کشید

قرعهٔ کار به نامِ منِ دیوانه زدند

آتش آن نیست که از شعلهٔ او خندد شمع

آتش آن است که در خرمنِ پروانه زدند

در شعر بالا به نظر می‌رسد که در نگارش کسره (ِ) بیش از اندازه بـه کـار گرفتـه شـده است.

متأسفانه به اندازه‌ی کافی زمان بـرای برداشـتن ایـن دشـواری وجـود نداشـت و درون واژه‌ها دستکاری نشد. برداشتن آنها نیز بدون برنـامه ریـزی، کـار درسـتی نبـود و دشـواریهای بیشتری را به وجود می‌آورد.

3.1.3.3- واژه‌هاي تركيبي

واژه‌های ترکیبی بیشترین دردسر را برای فارسی به وجـود آورده‌انـد. واژه‌هـایی چـون «خـوش خیال» یا «همینطور» که می‌توانند به دو گونه نوشته شـوند [صـنعتی،۱۳۷۱]. گرچـه بـه نظـر

می‌رسد که به کمک قاعده‌های آیین نگارش این دشواری کم شده است ولی برخی از دگرگونیها در نگارش کتابهای درسی نوین شگفت انگیز است.

دو نمونه‌ی زیر از دو کتاب درسی امروزی (همین دو سه سال) برداشته شده است. چگونگی نگارش هفتصد در این دو یکسان نیست.

نوشته‌ی زیر از کتاب «فاطمه فاطمه است» ، نوشته‌ی زنده یاد دکتر علی شریعتی است که نگارش آن برگرفته از [وحیدیان کامکار،۱۳۷۸] می‌باشد.

«هزار و هفتصد سال است که همه‌ی هنرمندان ، چهره نگاران ، پیکره سازان بشر در نشان دادن سیما و حالات مریم هنرمندی‌های اعجازگر کرده‌اند.»

نوشته‌ی زیر از زنده یاد دکتر محمود صناعی است که نگارش آن برگرفته از [سمیعی، ۱۳۷۸] است.

«با آن که بیش از هفت‌صد سال از حمله‌ی مغول می‌گذرد ، چنین می‌نماید که هنوز صفات اخلاقی قومی که مغول بر او حکومت می‌کرده در آداب و اخلاق برخی افراد دیده می‌شود. ستم کشیدن و تحمّل کردن ، چاپلوسی و تملّق گفتن ، خدعه و دروغ و تزویر و ریا ، وحشت از اینکه راست راه بروند و آن چه در دل دارند ، با صراحت بر زبان آورند. همه صفات مذموم اخلاقی هستند که اگر بخواهیم جامعه‌ی بهتری بسازیم ، باید مثل آبله و مالاریا با آن‌ها مبارزه کنیم.»

در هنگام افزودن و پیوند واژه‌ها یا پیشوند واژه یا واژه پسوند نیز تغییرهای گوناگونی رخ می‌دهد. در زیر نمونه‌هایی از این تغییرها نشان داده شده است.

«زنده»	+«ان»		←	«زندگان»
«زنده»	+«م»		←	«زنده‌ام»
«زنده»	+«ها»		←	«زنده‌ها»
«ن»	+	«افتاد»	←	«نیفتد»
«ن»	+	«آزما»	←	«نیازما»
«گو»	+ «م»		←	«گویم»

کوشش بسیاری برای پوشش دادن همه‌ی این حالتها نیاز است.

قاعده‌های واژه‌های ترکیبی در چند زبان دیگر ساده‌تر از فارسی می‌باشند که در اینجا دو نمونه آورده می‌شود .

واژه‌های ترکیبی در زبان انگلیسی به سه صورت نوشته می‌شوند به صورت یک واژه یکپارچه (یعنی واژه‌ی بلندی که میان واژه‌های تشکیل دهنده‌ی آن فاصله نباشد)، به صورت واژه‌ای که میان واژه‌های تشکیل دهنده‌اش خط تیره باشد و بصورت واژه‌ای که میان واژه‌های تشکیل دهنده‌اش فاصله پدیدار گردد. برای نمونه bathroom ، ape_man ، high jump همه واژه‌ی ترکیبی هستند. در زبان انگلیسی قاعده بر این است که ترکیبهای چند واژه‌ای همانند واژه‌ای community center finance committe به صورت یک واژه نوشته نمی‌شود [غلامعلی‌زاده،۱۳۷۵]. در قراردادهای نوشتار آلمانی ترکیبهای دو یا چند واژه‌ای به صورت یک واژه نوشته می‌شوند. مانند Versicherungsgesellschaft که به معنای «شرکت بیمه» است و واژه‌ای feuerundlebensversicherungsgesellschaft که به معنای «شرکت بیمه عمر و آتش» است [یارمحمدی،۱۳۷۴].

3.2- دشواریهاي نوشتار رایانه‌اي فارسي

در این بند به دشواریهایی پرداخته می‌شود که نوشتارهای رایانه‌ای فارسی با آن روبرو بوده و هستند. کوشش شده است که در هر یک از بندهای زیر بخشی از این دشواریها بخوبی بازگو شوند.

3.2.1- استاندارد خط

یکی از مهمترین دشواریهای زبان فارسی (خط فارسی) وجود نداشتن الگوی استاندارد نوشتاری برای زبان فارسی (یا رعایت نکردن قانونهای آماده شده تا کنون برای آن) است [صنعتی، ۱۳۷۱].

روند فارسی سازی و استاندارد نمودن زبان فارسی برای رایانه فراز و نشیبهای زیادی داشته است. در اینجا گوشه‌ای از آن آورده می‌شود.

3.2.1.1- نخستین کارها

کُد ascii استاندارد ۱۲۷ نویسه دارد و چون این کُد در یک بایت (۸ بیت) گذاشته می‌شود ۱۲۸ حالت دیگر از ۲۵۶ حالت آن باقی می‌ماند. در آغاز هر نویسه‌ای فارسی به یکی از این جاهای باقی مانده نگاشته می‌شد. این روش پیش از آن برای دیگر زبانها در دیگر کشورها به کار گرفته شده بود. در این روش هر برنامه‌نویس به دلخواه این نگاشت را با برنامه‌نویسی فراهم می‌کرد. کم‌کم برنامه‌هایی چون vegaf ، سایه ، سپند ، ساحر ، ... نوشته شد که این نگاشت را انجام می‌دادند و برخی تابعهای کتابخانه‌ای نیز برای برنامه‌نویسی فراهم می‌کردند. برخی برنامه‌نویسان این برنامه‌ها را برای فارسی سازی در نرم‌افزارهای خود به کار می‌بردند. هنوز نیز بسیاری از سایتهای فارسی نگاشت ویژه‌ای خود را به کار می‌برند و فونت (font) ویژه‌ای خود را دارند.

3.2.1.2- کُد cp1256

سپس شرکتهای نرم‌افزاری بزرگ دنیا که علاقه‌مند به فروش نرم‌افزارهای خود در کشورهای دیگر بودند به کمک همین روش (۱۲۸ حالت آزاد) کدی را ساختند که همه‌گیر شد. یکی از بزرگترین دردسرهای این کد برای فارسی رعایت نکردن ترتیب چهار نویسه‌ی ویژه‌ی فارسی (پ،چ،ژ،گ) بود. زیرا که این شرکتها پایه‌ی کار خود را بر زبان عربی گذاشته بودند و سپس فارسی را بدان افزوده بودند. برای هر زبان یا مجموعه‌ی زبان نزدیک به هم که با این روش کدگذاری شده بودند نام ویژه‌ای برگزیده شد [Scherer,2000]. زیرا باید بگونه‌ای می‌شد فهمید که ۱۲۸ حالت آزاد را برای چه زبانی به کار برده‌اند. کد عربی با این روش cp1256 (یا windows1256) نامیده شد. این کد در ایران آنقدر همه‌گیر شد که آن را unicode می‌خوانند که نامگذاری نادرستی است. یکی از قاعده‌هایی که در این کُد رعایت شد ، تغییر شکل نویسه با توجه به نویسه‌های پس و پیش آن است. به این ترتیب برای هر حرف فارسی تنها یک نویسه به کار گرفته می‌شود و تنها شکل نویسه هنگام در کنار هم گذاشتن نویسه‌ها در یک واژه تغییر می‌یابد.

3.2.1.3- یونی‌کد (unicode)

یونی‌کد یک کد جهانی است که چند شرکت بزرگ نرم‌افزاری دنیا به همراهی سازمان جهانی استاندارد (ISOJTC1/SC2) آن را به وجود آورده‌اند [Matthew,2000]. با چشم پوشی از فراز و نشیبهای این روش بسادگی می‌توان گفت که در این روش به جای یک بایت ، دو بایت به کار برده می‌شود. این کدگذاری رویه‌ی پایه‌ای چند زبانی[1] نامیده می‌شود. این کد ۶۵۵۳۵ نویسه دارد.

یونی‌کد از مجموعه نویسه‌های محدود ۸ بیتی بسیار فراتر رفته و با ظرفیت بیش از یک میلیون نویسه ، امکان کدگذاری همه‌ی زبانهای نوشتاری دنیا را فراهم می‌کند. افزون بر این برای برگزیدن خط و زبان نوشته ، به کدهای کنترلی نیاز ندارد. یونی‌کد رفتار یکسانی را با نویسه‌های الفبایی ، نویسه‌های اندیشه‌نگار و نمادها و نشانه‌ها دارد که امکان به کار بردن آمیخته‌ای آنها را با یکدیگر فراهم می‌کند. یونی‌کد افزون بر تعیین کد عددی و نام برای هر نویسه که در استانداردهای همانند، به کار گرفته شده است ؛ اطلاعات یونی‌کدی بیشتری را نیز برای پردازش و نمایش نوشته‌ها فراهم می‌کند؛ که از آن میان می‌توان به جهت نویسه و ویژگی‌های الفبایی اشاره کرد [علمدار میلانی،۱۳۸۱].

در این کد نیز زبانهایی که نزدیک به هم هستند در یک گروه گذاشته شده‌اند و باز چهار حرف «پ»،«چ»،«ژ» و «گ» ترتیب درستی ندارند. در این کد چند نویسه همانند با شماره‌های گوناگون وجود دارند؛ مانند چند «ی» و دو «ک» و ... . این دشواری دیگری است که به دشواریهای پیشین افزوده شده است. با اینکه این استاندارد به روشنی زبانهایی را که این نویسه‌های همانند در آنها کاربرد دارند ، نشان می‌دهد ولی برخی از سیستمها (مانند win2000 و winxp) این قانونها را رعایت نمی‌کنند و در هنگام برنامه‌نویسی سردرگمی ایجاد می‌کنند. بهترین نرم‌افزاری که یونی‌کد را پشتیبانی می‌کند unipad می‌باشد که بر پایه‌ای این استاندارد بنا شده است. سیستم عامل linux نیز این استاندارد را رعایت می‌کند و در نرم‌افزار gedit در نسخه linux mandrake 9.1 بخوبی فارسی استاندارد پشتیبانی شده است و صفحه‌ی کلید استاندارد ایران نیز رعایت شده است(بر خلاف winxp که صفحه‌ی کلید ویژه خود را داراست که بیشتر عربی است تا فارسی).

توضیحهای بیشتر درباره‌ی الگوریتم دو جهته[2] برای زبانهایی چون فارسی و عربی در [Davis,2003] گذاشته شده است. برگردان فارسی بخش فارسی این استاندارد در [علمدار

[1] 1_ Basic Multilingual Plane BMP

[2] 2_ Bidirectional Algorithm

میلانی،۱۳۸۱] گذاشته شده است. اگر چه همواره بهترین و به‌روزترین جا برای دیـدن آخریـن نسخه استانداردِ یونی‌کد در www.unicode.org است.

3.2.1.4- کد با طول متغیّر utf-8

کد ascii سالها استاندارد رایانه‌ها بوده است بسیاری از سخت‌افزارها ، شـبکه‌ای جهـانی و ... بـر پایه‌ی آن ساخته شده‌اند. بنابراین عوض کردن این کد(یک بایتی) اگـر نـاممکن نباشـد بسـیار پرهزینه و زمانگیر خواهد بود. چاره‌ای که اندیشیده شد به کار بردن کدی با تعداد بـایت متغیّـر بود که با یونی‌کد نگاشت یک به یک داشته باشد و همچنیـن ۱۲۸ نویسـه ascii را بـه همـان شکل یک بایتی باقی گذارد [Kuhn,2003].

برای نگاشت دیگر نویسه‌های یونی‌کـد ، ۱۲۸ حـالت بـاقی مانـده از نخسـتین بـایت و احتمالاً بایتهای پس از آن به کار گرفته می‌شود. نخستین بایت نشانه‌ای است که تعداد بایتهای پس از خود و رسته یونی‌کد را نشان می‌دهد در حالتی که نویسـه یـک نویسـه‌ای اسـکی باشـد همان یک بایت کافی خواهد بود [Burger,2001].

نگاشت UTF-8 به یونی‌کد از [Burger,2001] در زیر نشان داده شده است.

UTF-8 : Unicode

0x00000000 - 0x0000007F: 0xxxxxxx

0x00000080 - 0x000007FF: 110xxxxx 10xxxxxx

0x00000800 - 0x0000FFFF: 1110xxxx 10xxxxxx 10xxxxxx

0x00010000 - 0x001FFFFF: 11110xxx 10xxxxxx 10xxxxxx 10xxxxxx

0x00200000 - 0x03FFFFFF: 111110xx 10xxxxxx 10xxxxxx 10xxxxxx 10xxxxxx

0x04000000 - 0x7FFFFFFF: 1111110x 10xxxxxx 10xxxxxx 10xxxxxx 10xxxxxx 10xxxxxx

واین کد به شکل شانزده شانزدهی (octal) زیر می‌باشد.

xC20xA90 = 1010100111000010

این سیستم روزبه‌روز در حال گسترش است و بصورت استاندارد جهانی در آمده اسـت. در [Kuhn,2003] توضیح شایسته‌ای دربـاره‌ی utf-8 نوشـته شـده اسـت. در ایـران نیـز ایـن استاندارد پذیرفته شده است ولی هنوز همه‌گیر نشده است. در این پایان‌نـامه ایـن کـد بـه کـار گرفته شده است.

3.2.2- واژه‌هاي تركيبي

هنگام نوشتن واژه‌ی «خانه‌ها» باید «ها» در کنار «خانه» باشد و همزمان بـه آن نچسـبد و بـه شکل «خانهها» درنیاید. بنابراین باید فاصله مجازی بین این دو نویسه گذارده شود. نویسه‌ای به نام فاصله مجازی[1] در cp1256 و یونی‌کد وجود دارد که همین کار را انجام می‌دهد. این نویسه در بسیاری از واژه‌ها مانند «می‌روم» ، «گفته‌ام» ، ... به کار برده می‌شود. هنوز به درستی روشن نیست که این نویسه باید در چه ترکیبهایی به کار رود و کجا باید دو بخش واژه به هم چسبانده شوند و چه جاهایی باید فاصله معمولی (space) میان دو بخش گذاشته شود.

3.2.3- قالب پرونده

امروزه پرونده‌های نوشتاری فارسی با نرم‌افزارهای گوناگون نوشته می‌شوند. قالب ایـن پرونـده‌ها اغلب همخوانی کمی با هم دارنـد. نرم‌افزارهـایی همچـون pe2 ، زرنگـار ، کِلـک ، نشـر الـف ، microsoft word (که نسخه‌های گوناگون آن فارسی را به یک شکل پشـتیبانی نمی‌کننـد) ، pdf ، ... هر کدام قالب پرونده‌ی ویژه‌ای دارند. کسانی نیـز ایـن نوشـته‌ها را بـه تصـویر تبـدیل می‌کنند تا خواننده بتواند به سادگی آنهـا را بـر روی هـر رایـانه‌ای بخوانـد. آمـاده کـردن یـک برنامه‌ای رایانه‌ای که همه‌ی این قالبها را بخواند اگر ناممکن نباشـد بسـیار سـخت خواهـد بـود. پرونده‌های با قالب html (صفحه‌های شبکه جهانی) در این پایان‌نامه به کار گرفته شده‌اند. ایـن پرونده‌ها قالب استاندارد و ویژه‌ای دارند که به سادگی می‌توان واژه‌های درون آنهـا را بـا برنـامه خواند.

3.2.4- جدا سازي واژه‌هاي فارسي

شناسایی مرز میان واژه‌ها و جداسازی واژه‌ها یکی دیگر از دشواریهای زبان فارسی اسـت. بـرای نمونه در بسیاری از نوشته‌ها «و» در کنار واژه‌های پس و پیشِ خود نوشته می‌شـود. همچنیـن

[1] 1_ Zero Width Non-Joiner (ZWNJ)

برخی از واژه‌های دیگر نیز در نگارش (بویژه نگارش قدیم) در کنار یکدیگر نوشته می‌شوند و فاصله‌ای میان واژه‌های مستقل گذاشته نمی‌شود. گاهی نیز بخشهایی از یک واژه که باید در کنار هم نوشته شوند ، جدا نوشته می‌شوند. در حالی که یک بخش جداگانه‌ی آن می‌تواند ، خود واژه‌ی دیگری باشد که معنای متفاوتی دارد. برای نمونه دو واژه‌ی «نامه‌ای» و «نامهای» بسیار به هم نزدیک هستند و برخی دیگر از واژه‌ها نیز در زبان فارسی به همین گونه هستند. در [Megerdoomian,2000] این دشواری بخوبی توضیح داده شده است.

زمانی که نتوان مرز میان واژه‌ها را به روشنی شناخت ، انجام نخستین بخش سیستم بازیابی اطلاعات امکان پذیر نخواهد بود. برای به کار گرفتن روش پیشنهاد شده در [Megerdoomian,2000] برای برطرف کردن این دشواری ، نیاز به در دست داشتن لغت‌نامه‌ی کاملی بود تا بتوان به طور تقریبی واژه‌های در کنار هم را از هم جدا نمود. برای نمونه «محمدوزید» می‌تواند به دو گونه خوانده شود «محمد و زید» یا به شکل «محمد وزید» ، ولی چون می‌دانیم که محمد نام کسی است و نمی‌تواند بوزد ، بنابراین برداشت یکم آن درست است. دقت فرمایید که برای همین نمونه‌ی کوچک شما به چه دامنه‌ای بزرگی از دانش و تجربه برای جدا نمودن واژه‌ها نیاز دارید.

این دشواری به تنهایی زمانی بیش از سه ماه از کار بر روی پایان نامه را گرفت. جا دارد در اینجا بار دیگر از منصور رستگار فسایی استاد بخش ادبیات فارسی دانشگاه شیراز سپاسگزاری کنم که با راهنمایی ایشان توانستم این دشواری را از دیگر بخشهای کار جدا کنم زیرا تا پیش از آن ، دیگر بخشها را نیز به آن وابسته می‌دیدم. پس از آن این بخش کنار گذاشته شد و برخلاف پندار نخستین ، چندان لطمه‌ای به انجام کار نزد ؛ زیرا امروزه در نوشتار نوین فارسی اغلب (بویژه گونه‌ای نوشته شده با رایانه) واژه‌های مستقل ، جداگانه نوشته می‌شوند و ساده‌تر می‌توان جدا سازی را انجام داد. گرچه در نوشتار دستی فارسی این دشواری بیشتر دیده می‌شود.

3.3- واژه‌هاي سراسري

همانگونه که پیش از این گفته شد ، واژه‌های سراسری واژه‌هایی هستند که ارزش کمتری در نوشته دارند و در بردارنده‌ی معنا و مفهوم نوشتار نیستند. این واژه‌ها باید در بازیابی اطلاعات نوشتاری شناسایی شده و دور ریخته شوند. برای شناسایی واژه‌های سراسری دو روش وجود دارد:

الف- شناسایی واژه‌های سراسری بر پایه‌ی فراوانی

بر طبق نظریه‌ی لان ، واژه‌هایی که بیشترین تکـرار را دارنـد ، واژه‌هـای سراسـری هسـتند کـه ارزش به کار گیـری آنهـا کـم اسـت [Luhn,1957]. ایـن روش در فارسـی پاسـخ شایسـته‌ای نمی‌دهد. زیرا بسیاری از واژه‌های سراسری در زبان فارسی فراوانی چندانی در نوشـته‌ها ندارنـد. حتی با گردآوری بیش از یک GB صفحه‌های فارسی از شبکه‌ی جهانی و به کار گیری فراوانی واژه‌های درون آنها ، باز پاسخ گرفته شده بر پایه‌ی این روش بـه هیـچ وجـه کامـل نیسـت. بـه کارگیری این روش نمی‌تواند بسیاری از واژه‌های سراسری فارسی را بردارد.

ب- شناسایی واژه‌های سراسری به کمک فهرست آماده

در این روش یک فهرست از واژه‌های سراسری آماده می‌شود تا این واژه‌ها از نوشـته‌ها برداشـته شوند. این روش نیز نمی‌تواند همه‌ی واژه‌های سراسری فارسی را با یک فهرست کوچک درآورد. بزرگترین فهرست در [فرهودی،۱۳۸۰] دیده شد که باز کامل نبود. بـه خـاطر ویژگـی ترکیـبی واژه‌های فارسی ، فهرست آماده شده باید بزرگتر از اندازه‌های کنـونیِ خـود باشـد. بـرای نمـونه حالتهای گوناگون واژه‌ی «برای» در زیر نوشته شده است:

«برایم» ، «برایت» ، «برایشان» ، «ازبرایشان» ، ....

باید به گونه‌ای بتوان همه‌ی حالتهای ترکیبی را پوشش داد. دقت فرمایید که میان هیچ یـک از بخشهای واژه‌ها ، فاصله‌ی معمولی گذاشته نشده است.

3.4- ریشه‌یابی

یکی از پردردسرترین بخشهای این پایان‌نامه ، ریشه‌یابی واژه‌های فارسی بود. ریشه‌یابی واژه‌های فارسی یکی از پروژه‌های دردسر آفرین برای بسیاری از کارهای رایانه‌ای اسـت. بـا تـوجه بـه در دسترس نبودن ساده‌ترین امکانات ، برای این پایان نـامه ، دشـواری ایـن بخـش چنـدین برابـر گردید.

3.4.1- ناکارآمدي روشهاي ريشه‌يابي خودکار

برای یافتن کارهای دیگران در زمینه‌ای ریشه‌یابی ، به جاهـای زیـادی سـر زده شـد و پاسـخی شایسته گرفته نشد. دو روش ریشه‌یابی خودکار در دسترس بـر روی شـبکه‌ای جهـانی بررسـی گردید. که در زیر فشرده‌ای از هر کدام آورده شده است.

3.4.1.1- ريشه‌يابي به کمک روش porter

این روش یکی از توانمندترین و قدیمیترین روشهای ریشـه‌یابی در زبـان انگلیسـی اسـت. ایـن روش بر پایه‌ی زبان‌شناسی و دسته‌بندی واژه‌ها به کمک واجها و هجاها بنـا نهـاده شـده اسـت. پس از آن وندهای واژه‌های درون گردایه به طـور خودکـار برداشـته می‌شـوند. ایـن روش در [porter,1980] آورده شـده اسـت. الگـوریتم پیـاده سـازی آن بـه زبـان snowball در http://snowball.tartarus.org/porter/stemmer.html آورده شده است. ولی پیاده‌سـازی این روش در فارسی به نظر ناممکن می‌نمود (یا شاید به دانش زبان‌شناسی بیشتری نیـاز بـود). دسته‌هایی از آن گونه که در این روش برای زبان انگلیسی به کار رفتـه در زبـان فارسـی یـافت نشد.

3.4.1.2- روش آماري

در این روش (یا دسته از روشها) یک گردایه‌ی بزرگ از واژه‌ها با ساختهای گونـاگون گـردآوری می‌شود. هر چه این گردایه بزرگتر و کاملتر باشد این ریشه‌یابها بهتر کار می‌کنند. در ایـن روش تحلیل آماری به کار گرفته می‌شود . بـا روش آمـاری ونـدهایی کـه در واژه‌هـا تکـرار شـده‌اند، شناسایی می‌گردند. ایـن روش در [Manning,1999] آورده شـده اسـت. ایـن روش بـه زبـان بستگی ندارند و این بزرگترین خوبی این روش می‌باشد. اگر این روش بتواند برای زبان انگلیسی پاسخ شایسته‌ای بدهد ؛ گسترش آن به دیگران زبانهای دسته‌ای هند و اروپایی ساده خواهد بود. این روش سه دشواری بزرگ دارد :

الف ــ در این روش به یک گردایه‌ی بزرگ از واژه‌ها نیاز است و این گردایه باید کامـل باشـد و واژه‌های درون آن نیز درست باشند. وجود واژه‌های نادرست در گردایه بر کارآیی این ریشـه‌یاب

اثر بسیار بد می‌گذارد و آن را گمراه می‌کند. گردآوری گردایه‌ی بزرگی از واژه‌های صد در صد درستِ فارسی نیز، ناممکن می‌نمود.

ب ــ هنوز این روشها در حال آزمایش هستند و کارآیی آنها چشمگیر نیست.

ج ــ این روشها نیاز به رایانه‌های با سرعت زیاد و حافظه بزرگ دارند و اجرای برنامه‌های نوشته شده بر پایه‌ای این روشها بسیار زمانبر است. برای اجرای این روشها با رایانه‌های در دسترس باید تعدادی از آنها با هم موازی می‌شدند و شاید برای یک بار اجرا ، چند روز زمان گرفته می‌شد. گرچه در پیاده سازی این روشها بهتر می‌توان به نیازهای آنها پی برد.

3.4.2- ریشه‌یابي به کمک قاعده‌هاي فارسي

گرچه دو روشی که برای ریشه‌یابی خودکار در بالا آمد ؛ کنار گذاشته شد ولی ایده‌هایی از آنها گرفته شد که در ساخت ریشه‌یاب فارسی کمک نمود.

آخرین چاره ، ریشه‌یابی به کمک قاعده‌های زبان فارسی بود. باید کاری که یک فارسی زبان برای ریشه‌یابی انجام می‌دهد، همانند سازی شود. برای نمونه می‌دانیم که «گفت» ریشه‌ی گذشته یک فعل است و ریشه غیرگذشته آن «گو» می‌باشد ؛ بنابراین با نگاهی به دستور زبان در می‌یابیم که ریشه همه‌ی واژه‌های زیر «گفت» می‌باشد.

«گفتم» ، «می‌گفتم» ، «نگفتم» ، «گفته‌ام» ، «می‌گوید» ، «گوینده» ، «گفتار» و ...

همچنین می‌دانیم که «خوب» یک صفت ساده است بنابراین «خوبتر» ، «خوبترین» ، «خوبی» و ... همه از «خوب» ساخته شده‌اند. پس باید قاعده‌های دستور زبان به کمک برنامه‌نویسی پیاده سازی می‌شد.

3.4.2.1- در دسترس نبودن لغت‌نامه

برای به کار گرفتن قاعده‌های دستور زبان لغت‌نامه‌ای نیاز بود که همه‌ی واژه‌های بنیادی فارسی را دربرداشته باشد. این لغت‌نامه باید رسته‌ی واژه‌ها را نیز به همراه واژه داشته باشد تا روشن نماید که یک واژه فعل یا صفت یا اسم یا ... است. به کمک این لغت‌نامه می‌شد قاعده‌های دستور زبان فارسی را برای ریشه‌یابی به کار گرفت. همچنین به کمک این لغت‌نامه واژه‌هایی که فارسی نیستند یا نادرست نوشته شده‌اند ، شناسایی شده و دور ریخته می‌شدند.

باید یادآوری شود که نمی‌توان یک قاعده‌ی فارسی را روی همه‌ی واژه‌های یک رده به کار بـرد. برای نمونه «پرورد»+«گار» ← «پروردگار» یک واژه‌ی درست است در حالی که «گفت»+«گار» ← «گفتگار» واژه‌ی درستی نیست.

پس از ساخته شدن واژه‌های ترکیبی باید درستی آنها آزموده می‌شد. بنابراین باز در اینجـا نیـز به یک لغت‌نامه‌ای کامل نیاز بود. به دو دلیل چنین لغت‌نامه‌ای یافت نشد.

الف- پایگاه واژه‌های یک لغت‌نامه‌ی فارسی در دسترس نبود. بـه کـارگیری پایگـاه واژه‌ی نـرم افزارهای در دسترس (یا روشنتر ، شکستن پایگاه داده و برداشـت واژه) نیـز بـه دلیـل تجـاری نادرست و غیر قانونی بود. خرید پایگاه واژه‌ی این نرم افزارها نیز احتمالاً هزینه‌ی زیادی می‌برد. برای خرید لغت‌نامه‌های آماده با قالب مشخص (برای نمونه با قالب نرم افزار Excel) از شبکه‌ی جهانی نیز بودجه‌ای فراهم نبود.

ب- می‌توان گفت که در هیچ لغت‌نامه‌ای نمی‌توان واژه‌هایی ترکیبی چون «برایم» ، «برایت» ، ... و یا «گفتم» ، «گفتید» ، ... را یافت. با توجه به کارکرد لغت‌نامه‌ها ، نیازی هم نیست که این واژه‌ها در لغت‌نامه وجود داشته باشند. در حالی که این واژه‌ها در روزمره به کاربرده می‌شوند.

نوشتن (تایپ) یک لغت‌نامه‌ی کوچک نیز در دسـتور کـار قـرار گرفـت ولـی بـه پایـان رسانیدن آن (آن هم یک نفری) ناممکن و بسیار زمانگیر می‌نمود. پس باید روش نوینی سـاخته می‌شد تا بتواند بر این دشواریها چیره شود.

3.4.2.2- واژه‌هاي زبانهاي ديگر در زبان فارسي

اگر بنا باشد که روش خودکار دیگری نیز برای ریشـه‌یابی بـه کـار رود ، وجـود انبـوه واژه‌هـای زبانهای دیگر در زبان فارسی دشواریهای دیگری را برای ریشه‌یابِ خودکار ایجاد می‌کند. در زیر یک نمونه از این دشواریها آورده شده است.

برخی از واژه‌های زبانهای بیگانه‌ی آورده شده به زبان فارسی به ساختهای دستور زبـان فارسی نزدیک هستند و این واژه‌ها کار ریشه‌یابی را سخت‌تر می‌کنند. گاهی یـک واژه‌ی بیگـانه بصورت ترکیب یک «وند»(پیشوند یا پسـوند) و یـک واژه‌ی فارسـی درمی‌آیـد و واژه‌ی فارسـی ریشه آن در نظر گرفته می‌شود . برعکس گاهی واژه‌ای بیگـانه ، ریشـه‌ی واژه‌ای فارسـی ـــ بـه دلیل نزدیکی آن دو و به کار رفتن یک قاعده‌ای فارسی ـ در نظر گرفته می‌شود.

3.5- زبان برنامه‌نویسي

برای شناسایی واژه‌های سراسری ترکیبی و ریشه‌یابی فعلهای فارسی باید یک زبان برنامه‌نویسی شایسته برگزیده می‌شد. هنگام برنامه‌نویسی بخشهای فارسی این پایان نامه ، باید حالتهای گوناگونی که هنگام ترکیب واژه‌های فارسی پیش می‌آید ، پوشش داده می‌شد. همچنین ویرایشهای پیچیده و بسیار زیادی بر روی قاعده‌هایی از دستور زبان فارسی که به کار گرفته شده بود ، باید انجام می‌شد یا به عبارت دیگر در هنگام برنامه‌نویسی ، پیاپی بسیاری از بخشها دگرگون می‌شد. به هیچ روی این امکان وجود نداشت که یک روند خطی برای طراحی و پیاده سازی نرم افزار[1] در نظر گرفته شود. با توجه به این دشواریها یک زبان برنامه‌نویسی بسیار ساده و انعطاف پذیر و توانمند نیاز بود. این زبان باید استاندارد یونی‌کد را به خوبی و سادگی پشتیبانی می‌کرد. در این زبان باید کار با رشته‌های یونی‌کدی نیز بسیار ساده می‌بود.

زبانهای برنامه‌نویسی فراوانی آزموده شد. که از آن میان می‌توان cpp و java و python را نام برد. با این که این زبانها ، زبانهای بسیار توانمندی هستند ولی به کمک این زبانها (و بسیاری از زبانهای آشنای دیگر) نیازهای این نرم‌افزار برآورده نمی‌شد. در فصل چهارم بیشتر به این بخش پرداخته می‌شود.

3.6- گردایه‌ي ویژه‌ي بازیابي اطلاعات

همان گونه که پیش از این گفته شد ، برای ارزیابی یک سیستم بازیابی اطلاعات یک گردایه‌ی ویژه نیاز است. چنین گردایه‌ی آماده‌ای برای زبان فارسی یافت نشد و ساخت چنین گردایه‌ای که با دقت فراهم شده باشد ، بسادگی امکان‌پذیر نیست. برای زبان انگلیسی چنین گردایه‌های آماده‌ای وجود دارد. اغلب سیستمهای بازیابی اطلاعات انگلیسی به کمک این گردایه‌های آماده ارزیابی می‌شوند. گردآوری چنین گردایه‌ای بیشتر با پشتیبانی دولتها انجام می‌شود.

[1] 1_ top down design

فصل چهارم

پیاده سازیِ سیستم پیشنهادي

4- پیاده سازیِ سیستم پیشنهادي

در فصلهای پیشین به پایه‌های بازیابی اطلاعات و دشواریهای بازیابی اطلاعات نوشتار فارسی پرداخته شد. در این فصل به راهکارهایی برای برطرف کردن دشواریهای کارهای پیش‌پردازشی در فارسی و همچنین پیاده‌سازی یک سیستم بازیابی اطلاعات فارسی پرداخته شده است. در بندهای آغازین این فصل به روش پیشنهادی برای برداشتن واژه‌های سراسری و ریشه‌یابی پرداخته می‌شود.

سپس چگونگی جمع‌آوردی یک گردایه‌ی ویژه‌ی بازیابی اطلاعات آورده شده است. پس از انجام پیش‌پردازش لازم بر روی این گردایه ، ماتریس سند-واژه ساخته شده است. سپس به کمک پرس و جوهای درون‌گردایه پاسخهای سیستم فراهم شده است. با کمک نگاشت میان پرس و جوها و سندها در این گردایه نمودار یادآوری دقت این سیستم کشیده شده است. این سیستم یک بار بدون ریشه‌یابی و برداشتن واژه‌های سراسری ارزیابی شده است. بار دیگر کارآیی سیستم با ریشه‌یابی و برداشتن واژه‌های سراسری سنجیده شده است. سپس نمایه‌گذاری معانی پنهان پیاده‌سازی شده است.

4.1- زبان برنامه‌نویسي

همانگونه که در فصل پیش آورده شد، بخش پیش پردازشی بازیابی اطلاعات نوشتاری فارسی با دشواریهای فراوانی روبرو شده بود. با برگزیدن یک زبان برنامه‌نویسی شایسته از این دشواریها کاسته می‌شد. زبانهای برنامه‌نویسی فراوانی برای انجام بخش پیش‌پردازشی آزموده شد. یافتن زبان برنامه‌نویسی شایسته، خود دشواری دیگری شده بود.

ساخت یک شبه‌زبان برنامه‌نویسی نوین، به جای به کارگیری زبانهای برنامه‌نویسی در دسترس بهترین چاره‌ی این دشواری دیده شد. با این شبه‌زبان برنامه‌نویسی باید همه‌ی

ویژگیهای مورد نیاز بخش پیش‌پردازشی برآورده می‌شد. تجربه‌ی پروژه‌ی کارشناسی (ساخت مفسر REXX) و آشنایی با بسیاری از زبانهای برنامه‌نویسی به ساخت این شبه‌زبان نوین بسیار کمک کرد، ولی باز دشواریهای زبان فارسی طراحی و ساخت این شبه‌زبان نوین را بسیار سنگین می‌نمود. با توجه به انبوه کارهای باقی‌مانده و زمان بسیار کم باقی‌مانده برای انجام همه‌ی بخشهای این پایان‌نامه، آماده کردن برنامه‌ی اجرا کننده‌ی دستورهای این شبه‌زبان نوین ناممکن شده بود. دوست بزرگوار آقای مهدی احمدی زحمت ساخت اجراکننده‌ی این شبه‌زبان را کشیدند. در اینجا باز از همکاری گرم ایشان سپاسگزاری می‌کنم.

4.1.1- شبه‌زبان برنامه‌نویسي جدید

امروزه با وجود همه‌ی زبانهای برنامه‌نویسی در دسترس، شرکتهای بزرگ رایانه‌ای به این نتیجه رسیده‌اند که برای برخی از کارهای بسیار ویژه و گسترده‌ی خود یک زبان ویژه برای آن کار طراحی کنند. گرچه ساخت یک زبان برنامه‌نویسی جدید زمان و نیروی زیادی می‌گیرد ولی برای کارهای بسیار ویژه که نیاز به ویرایشهای تخصصی فراوانی در آینده در آنها هست؛ این هزینه‌ای گزاف در هنگام به کار بردن ویرایشها بخوبی جبران می‌شود.

این شبه‌زبان بسیار به قالبی که دستور نویسان و زبان‌شناسان برای نوشتن دستور زبان فارسی به کار می‌برند نزدیک است. سادگی این شبه‌زبان بگونه‌ای است که بدون دانستن هیچ زبان برنامه‌نویسی دیگر می‌توان این شبه‌زبان را به کار گرفت. انگیزه‌ی این کار این بوده که شکاف کنونی میان متخصصان زبان فارسی و متخصصان رایانه کاسته شود. این چنین می‌نماید که برای چیره شدن بر پیچیدگیهای فراوان کارهای رایانه‌ای درباره‌ی زبان فارسی نیاز به همکاری بیشتر میان متخصصان این دو رشته است. نحو[1] این شبه‌زبان برنامه‌نویسی و علامتها و عملگرهای تعریف‌شده در آن ، بر پایه‌ای شیوه‌ی نگارش در]رحیمیان، 1381[است.

این شبه‌زبان اکنون گردایه‌ی بسیار کوچکی از عملگرها را داراست. دلیل نیفزودن امکانات بیشتر به این شبه‌زبان بی‌نیازی به آنها (تا این مرحله) بوده است. هنگامی که امکانات یا دستورهایی نیاز شد آن امکانات به ساده‌ترین و بهترین روش به این شبه‌زبان افزوده خواهد شد. افزودن امکانات اضافی تنها به پیچیده‌تر شدن این زبان می‌انجامد، بنابراین از افزودن امکاناتی که نیاز نبود (حتی دستورهای پایه همچون if else)، پرهیز شد. برای آشنایی با این شبه‌زبان یک خط نوشته شده با آن در زیر آورده شده است.

[1] 1_ Syntax

«کجا» + شناسه گذشته/ «ست» / نشانه_ها/∅ .

برای نوشتن برنـامه بـا ایـن شـبه‌زبان ویرایشـگر gedit در linux mandrake9.1 و ویرایشگر unipad در windows به کار گرفته شد. این دو ویرایشگر بدرستی اسـتاندارد یـونی کد را پشتیبانی می‌کنند و صفحه کلید فارسی به کار گرفته شده در آنها نیز برپایه‌ای اسـتاندارد ایران است. دستورهای این زبان همانند فارسی از راست به چـپ نوشـته می‌شـوند. نویسـه‌های فارسی یا انگلیسی یا (-) در نامگذاری متغیرهای این زبان به کار گرفتـه می‌شـود. بـرای نمـونه «ضمیر_متصل_مفعولی» می‌تواند نام یـک متغیـر باشـد. مقـدارهای ثـابت میـان گیـومه («») گذاشته می‌شوند.

∅ : در]رحیمیان، 1381[به کار گرفته شده است و به معنای تهی یا هیچ نویسه‌ای است.

چون برخی از دستورهای نوشته شده به این زبان بسیار بزرگ می‌شد، بهتر بـود بـرای خوانایی برنامه‌ها، یک دستور در چند خط نوشته شود. بنابراین برای نشان دادن پایان دسـتورها (همانند جمله‌های فارسی) نقطه به کار گرفته شد. همانند برخی از زبانهای برنامه‌نویسـی میـان دو نشانه /× ×/ توضیحها[1] گذاشـته می‌شـود. ماننـد همه‌ی زبانهـای برنامه‌نویسـی پرانـتز بـرای دسته‌بندی و تغییر اولویت تعریف شده است.

عملگر / (یا or) دو عملوند می‌گیرد که به معنای عملوند یک یـا عملونـد دو یـا هـر دو (ولی نه در کنار هم) است. این عملگر نیز از]رحیمیان، 1381[گرفته شده اسـت. (=) عملگـر انتساب است.

(+) عملگر اتصال است که دو عملوند خود را (دو رشته را) در کنار هـم می‌گـذارد و یـک واژه‌ای جدید یا بخشی از یک واژه را می‌سازد. همانگونه که پیش از این گفته شـد، قاعـده‌های اتصـال فارسی حالتهای ویژه‌ای دارند. در پیاده‌سازی این عملگر کوشیده شد تا همه‌ی حالتهـا در نظـر گرفته شوند و پیاده‌سازی گردند.

در دستور زبان فارسی برخی از پسوندها نام ویژه‌ای دارند. به جای تعریف کردن آنها به صورت واژه‌ای کلیدی در شبه‌زبان با نوشتن دستورهایی (به شبه‌زبان) این واژه‌ها به شـکل یـک متغیر تعریف شدند. در زیر چند دستور از دستورهایی که این متغیرهـا را مقـداردهی می‌کننـد آورده شده است.

ضمیر_متصل_مفعولی= «م» / «ت» / «ش» / «مان» / «تان» / «شان».

نشانه_جمع = «یم» / «ید» / «ند».

[1] 1_ Comments

برخی از وندها و واژه‌هایی که در زبان فارسی بسیار به کار می‌روند؛ نیز به صورت متغیر تعریف شدند.

وابسته_ان_جمع= «یست» / «ست» / «ی» + «که» /∅ .

نشانه_ان_جمع= «ان» + شناسه_جمع / ضمیر_متصل_مفعولی / وابسته_ان_جمع/∅.

4.2- ساخت واژه‌هاي سراسري تركيبي

به کمک این شبه‌زبان واژه‌های سراسری ترکیبی به سادگی ساخته شد. برای نمونه حالتهای گوناگون واژه «برای» به کمک دستور زیر ساخته می‌شود:

«از» / ∅ + «برای» + ضمیر_متصل_مفعولی/∅.

با اجرای این دستور ، واژه‌های «برایم»، «از برایت»، ... همه یک جا ساخته می‌شوند.

برای افزودن دیگر واژه‌های سراسری به همراه پیشوند یا پسوندهایی که ممکن است به همراه آنها بیاید تنها کافی است که دستوری به برنامه افزوده شود؛ برنامه‌ای به شبه‌زبان جدید نوشته شد. که همه‌ی دستورهای ساخت واژه‌های سراسری در آن آورده شده است. اجرا کننده‌ای[1] این شبه‌زبان با اجرای دستورها فهرست بزرگی از واژه‌های سراسری ترکیبی را ساخت. به خاطر نادرستی برخی از ترکیبها ، بسیاری از واژه‌های این فهرست بزرگ نادرست بودند. باید به گونه‌ای خودکار واژه‌های درست این فهرست شناسایی می‌شد.

4.3- آزمون درستي واژه

برای آزمودن درستی یک واژه‌ای ساخته شده نیاز بود تا لغت‌نامه‌ی کاملی در دسترس باشد که همه‌ی واژه‌های فارسی روزمره را دربرداشته باشند. در فصل پیش به دشواری لغت‌نامه پرداخته شد.

بنابراین باید بگونه‌ای فهرست کاملی از واژه‌های فارسی روزمره گردآوری می‌شد. پس از کوششهای فراوان بهترین منبع برای گردآوری واژه‌های فارسی روزمره، روزنامه‌های ایرانی در

[1] 1_ interpreter

دسترس روی شبکه جهانی تشخیص داده شد. (شاید نقل نوشته‌ای از روزنـامه‌ای بـدون آوردن نام منبع غیرقانونی باشد ولی احتمالاً به کارگیری واژه‌های درون آنها کار نادرستی نیسـت). بـه کمک نرم‌افزار رایگان httrack هزاران صفحه‌ای روزنامه گـردآوری شـد (نزدیـک بـه یـک GB نوشته)، کُد نوشتاری این صفحه‌ها CP1256 (یا Windows1256) بود. واژه‌های فارسی درون این نوشته‌ها و فراوانی آنها درآورده شد. چون واژه‌هایی که فراوانی بسیار کمی دارند به احتمـال زیاد نادرست (اشکال تایپی یا نگارشی) هستند یا واژه‌هـای بسـیار تخصصـی هسـتند؛ بنـابراین واژه‌های با فراوانی بسیار کم دور ریخته شدند.

سپس کُد این واژه‌ها به utf-8 تبدیل شدند تا با کُد واژه‌های سـاخته شـده بـه کمـک شبه‌زبان یکسان شوند. بدین سان فهرست واژه‌هـای روزمـره گـردآوری شـد و آزمـون درسـتی واژه‌های سراسری ساخته شده به کمک آن انجام شد. این گردایه‌ی بـزرگ از واژه‌هـا در بخـش شناسایی خودکار واژه‌ها و ریشه‌یابی نیز به کار گرفته شد.

4.4- ریشه‌یابی

روشهای گوناگونی برای ریشه یابی فارسی آزموده شد. به دلیل دشواریهای فراوانی که در فصـل پیش آورده شد، پاسخ شایسته‌ای از این روشها گرفته نمی‌شـد. بـه دلیـل ناکارآمـدی روشـهای خودکار ریشه‌یابی ، به کار گرفتن قاعده‌های دستور زبان فارسی چاره‌ی کار دیده شد. بـرای بـه کار گرفتن قاعده‌های دستور زبان نیاز به یک لغت‌نامه است. همان گونه که پیش از ایـن گفتـه شد ، چنین لغت‌نامه‌ای در دسترس نبود. در زبان فارسی شناسایی رده‌ی یک واژه (فعل ، اسم ، ...) بسادگی امکانپذیر نیست. برای نمونه به کارگرفتن گزاره‌ی ساده‌ی «فعل کلمه‌ای است کـه شناسه می‌پذیرد.» ، به هیچ روی نمی‌تواند فعلهـا را شناسـایی کنـد. زیـرا بسـیاری از اسـمها و صفتها نیز شناسه می‌گیرند. برای نمونه اغلب «خوبیم» به جای «خوب هستیم» و «عبارتند» به جای «عبارت هستند» به کار گرفته می‌شوند. فعلها در زبان فارسی رده‌ی بسیار بزرگی هسـتند که به کمک آنها بسیاری از اسمها و صفتها ، ... ساخته می‌شوند. بنابراین با شناسایی ایـن رده ، می‌تـوان بسـیاری از واژه‌هـای فارسـی را ریشـه‌یابی نمـود. در [احمـدی گیـوی،1380] روش شایسته‌ای برای شناسایی فعلهای فارسی و بُن گذشته و بُن غیر گذشته‌ی آنهـا پیشـنهاد شـده است. فشرده‌ی این روش در زیر آورده شده است.

« در فعلهای ساده پس از حذف « ـَن » از مصدر ، بُن ماضی باقی می‌ماند و از جهت تغییری که از بُن ماضی به بُن مضارع انجام می‌گیرد ، آنها را می‌توان در هشت گروه جای داد. شکل زیر تغییرهای فعلها را در گروههای هشت‌گانه نشان می‌دهد. »]احمدی گیوی،1380[

شماره‌ی گروهها	حروف پایانی بن‌ماضی+پسوند	حروف پایانی بن‌ماضی	بن مضارع	مثال	
1	یدن	ـ ید	پس از حذف «ـید» باقی‌مانده بن مضارع است.	نال+ـید+ـَن	نالیدن نال / نالید
2	دن	د	پس از حذف «د» باقی‌مانده بُن مضارع است.	خور+د+ـَن	خوردن خور / خورد
3	ودن	ـود	پس از حذف «و» ، «و» به «ا» تبدیل می‌شود.	آزمو+د+ـَن	آزمودن آزما / آزمود
4	ادن	ـاد	پس از حذف «اد» باقی‌مانده بن مضارع است.	افت+اد+ـَن	افتادن افت / افتاد
5	ختن	ـخت	«ت» حذف و «خ» تبدیل به «ز» می‌شود.	ساخ+ت+ـَن	ساختن ساز /ساخت
6	ستن	ـست	پس از حذف «ست» باقی‌مانده بن مضارع است.	آرا+ست+ـَن	آراستن آرا / آراست
7	شتن	ـشت	پس از حذف «ت» ، «ش» تبدیل به «ر» می‌شود.	کاش+ت+ـَن	کاشتن کار / کاشت
8	فتن	ـفت	پس از حذف «ت» ، «ف» تبدیل به «ب» می‌شود.	تاف+ت+ـَن	تافتن تاب / تافت

شکل 4-1- هشت گروه فعلهای فارسی

تعداد کمی از فعلهای فارسی از قاعده‌های این هشت گروه پیروی نمی‌کنند. بُن گذشته و غیر گذشته‌ی این فعلها جداگانه نوشته شدند.

این روش بر روی گردایه‌ی بزرگ واژه‌ها به کار گرفته شد. بُن گذشته و غیر گذشته‌ی همه‌ی فعلهای ساده‌ی فارسی بخوبی شناسایی شدند. درصد بسیار پایین خطا در شناسایی فعلها به کمک این روش با وجود انبوه واژه‌های زبانهای بیگانه در گردایه‌ی واژه‌ها ، شگفت انگیز بود.

به جای روش سر راست برای ریشه‌یابی ، روشی وارون به کار گرفته شد. در ادامه به این روش پرداخته می‌شود.

دو واژه‌ی دیگر به شبه زبان افزوده شد:

الف- «بُن_گذشته» که نشان دهنده‌ی بُنهای گذشته‌ی فعلهای شناسایی شده است. به عبارت دیگر در هنگام اجرای برنامه به شبه زبان ، به جای این واژه‌ی کلیدی ، بُنهای گذشته‌ی همه‌ی فعلهای شناسایی شده گذاشته می‌شود.

ب- «بُن_غیر_گذشته» که نشان دهنده‌ی فهرست بُنهای غیر گذشته‌ی فعلهای شناسایی شده است.

اکنون به کمک این گسترش در شبه زبان و با بـه کـارگیری قاعـده‌های دسـتور زبـان فارسی مشتقهای همه‌ی فعلهای فارسی ساخته شدند. همزمان نگاشـت میـان ریشـه‌ی فعـل و واژه‌های مشتق ساخته شده ، نگهداری شـد. بـرای نمـونه دو دسـتور از دسـتورهایی کـه بـرای ساخت مشتقها نوشته شده است ، در زیر آورده شده است.

بُـن_گذشـته + «گـار» + نشـانه_ان_جمـع/نشـانه_یسـتکه/نشـانه_هـا /شناسـه_گذشـته /ضمیر_متصل_مفعولی/«ا»/∅ .

×/ صفت لیاقت و فاعلی می‌سازد و برخی مفهوم اسمی دارند .

آفریدگار ، کردگار ، پروردگار ، رستگار ، ماندگار /×

بُن_غیر_گذشته + «گار» + نشانه_یستکه/نشانه_ان_جمع/نشانه_ها/نشانه_تر/نشانه_ترین/«ا»/ شناسه_گذشته/∅ .

×/ آموزگار ، سازگار ، آمرزگار /×

پس تا اینجا فهرست همه‌ی مشتقهای فعلهای فارسی شناسایی شد. ولی در اینجـا نیـز (مانند واژه‌های سراسری ساخته شده) بسیاری از واژه‌های ساخته شده ، نادرست هستند. بـرای نمونه در زبان فارسی «پروردگار» (بُن_گذشته + «گار») وجود دارد ، در حـالی کـه «گفتگـار» وجود ندارد و به عبارت دیگر قانونهای زبان فارسی استثناهای فراوان دارند. بنابراین باز درسـتی واژه‌های ساخته شده به کمک گردایه‌ی واژه‌ها آزموده شد. مشتقهای ساخته شـده‌ای کـه درون این گردایه‌ی بزرگ از واژه‌ها وجود نداشتند ، دور ریخته شدند.

پس فهرست مشتقهای فعلهای فارسی و نگاشت آنها به ریشه‌ی فعل آماده شد. پـس از مرتب سازی این فهرست ، روند ریشه‌یابی در کمترین زمان ممکن انجام می‌شود.

همانگونه که دیده می‌شود فهرست ساخته شده برای کمـک بـه ریشـه‌یابی ، جـدای از سیستمی که نیاز به ریشه‌یابی دارد ، آماده شد[1]. سپس در هنگام نیاز به ریشـه‌یابی بـرای یـک واژه در سیستم پایانی ، تنها یک جستجوی دودویی[2] انجام می‌شود. در صورت وجود آن واژه در مشتقها ، به کمک نگاشت به ریشه‌ی آن واژه ، ریشه‌ی واژه برگردانده می‌شود و در صورتی کـه آن واژه در فهرست یافت نشود ، خود واژه برگردانده می‌شود.

[1] 1_ offline

[2] 1_ binary search

4.5- گردایه‌ي ویژه‌ي بازیابي اطلاعات نوشتاري فارسي

همانگونه که در فصل پیش آورده شد، چنین گردایه‌ی آماده‌ای برای فارسی در دسترس نبود. پس از کندوکاوهای بسیار بهترین منبع برای گردآوری سندهای چنین گردایه‌ای چکیده‌ی پایان نامه‌های در دسترس بر روی www.irandoc.ac.ir تشخیص داده شد. چکیده‌ی پایان نامه‌های روی این سایت در قالب html و با کد CP1256 هستند. چون زمان بسیار کوتاهی برای انجام این کار باقی‌مانده بود، زمان چندانی برای خواندن چکیده‌ها (دستِ کم تعداد کمی از آنها) وجود نداشت.

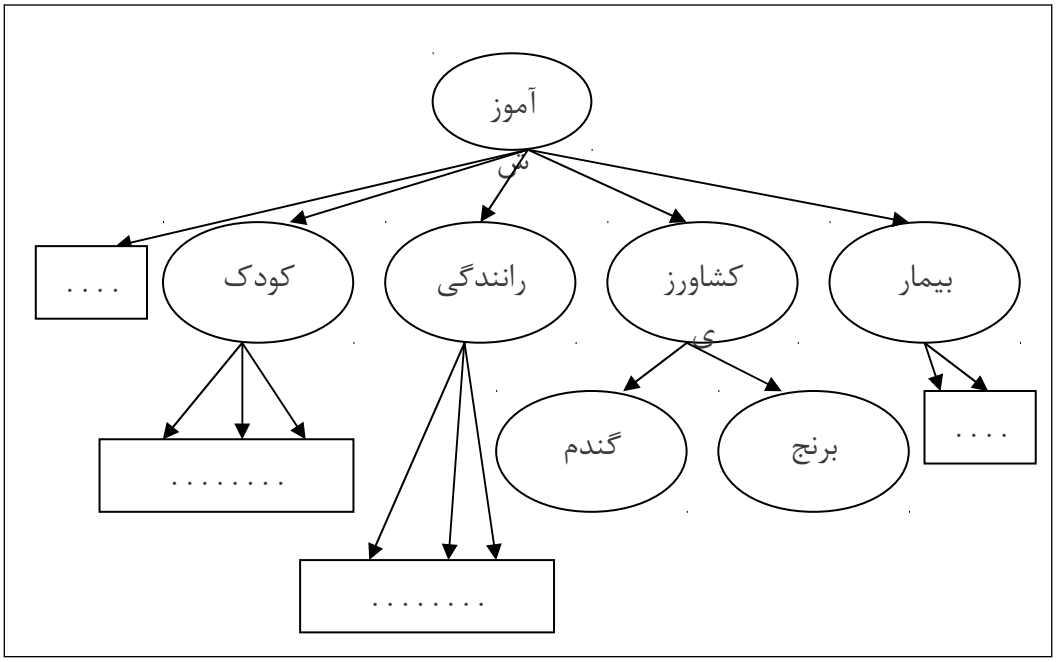
بنابراین به کمک جستجویی که در همان سایت بود، کار برگزیدن سندهای وابسته به پرس و جو انجام شد. جستجوی درون این سایت به صورت جستجوی واژه به واژه[1] است. در این گونه از جستجو هیچ کار پیش پردازشی بر روی واژه انجام نمی‌شود. برای اینکه شاید تا اندازه‌ای این گردایه به استاندارد نزدیک باشد، در پرس و جوهایی که از سایت می‌شد، واژه‌های هم‌خانواده به همراه هم در یک جستجو گذاشته شدند. به عبارت دیگر در هنگام جستجو واژه‌های هم‌خانواده با هم یای منطقی[2] شدند. همانند بیشتر سایتها در این سایت نیز باید«+» برای یای منطقی کردن واژه‌ها در یک پرس و جو در میان واژه‌ها گذاشته می‌شود. با انجام این کار، سایت irandoc در پاسخ به پرس و جو، همه‌ی چکیده‌ی پایان نامه‌هایی را که هر کدام از این واژه‌ها در درون آن قرار داشتند، نشان می‌داد. برای نمونه اگر بنا بود در چند تا از پرس و جوهای پایانی سیستم بازیابی ساخته شده در این پایان نامه، واژه‌ی آموزش به کار رفته باشد، آنگاه پرس و جویی به صورت زیر به سایت irandoc داده می شد.

آموخت + آموختم + ...+ می آموزم + می آموزی + ...+ آموزگار + ...

یکی از امکاناتی که درون جستجوی سایت irandoc وجود دارد ، جستجو بر روی پاسخهای داده شده کنونی است. پس هر بار باید واژه‌ی دیگری همراه با هم‌خانواده‌های آن برگزیده می‌شد، تا آن واژه (به همراه هم‌خانواده‌های آن) در این سایت جستجو شود. این کار به صورت درختی انجام شد. در زیر بخشی از نمودار به کار گرفته شده برای جستجو در این سایت، نشان داده شده است.

[1] 2_ exact search

[2] 3_ or



شکل 4-2- بخشی از نمودار به کار گرفته شده برای جستجو در سایت

با پیمایش یک راه از ریشه‌ی درخت تا یک برگ، یک پرس و جو ساخته می‌شد و سندهای پایانی در برگ، سندهای وابسته به آن پرس و جو بودند. یکی از شرطهای گردایه‌ی ویژه‌ی بازیابی اطلاعات این است که باید سندهای درون گردایه در یک یا چند موضوع نزدیک به هم باشند. به کمک روش به کار گرفته شده در جستجو از سایت irandoc برای گردآوری سندها و پرس و جوهای پایانی، این شرط تا اندازه‌ای برآورده شد. در پایان با این روش 120 سند و 40 پرس و جو آماده شد.

این روش گردآوری به هیچ روی نمی‌تواند یک گردایه‌ی شایسته فراهم کند. ولی با توجه به زمان بسیار کمی که در اختیار بود، شاید این روش بهترین روش شدنی برای انجام کار بود.

خواندن سندهای بسیار و ساخت پرس و جوهای شایسته و ساخت نگاشت میان پرس و جوها و سندها بسیار زمانبر است و دست کم به متخصصانی از رشته‌ی کتابداری نیز نیاز بود. باید توجه داشت که این گردایه بر روی کارآیی دیگر بخشهای سیستم اثر بسزایی دارد. بنابراین به نتیجه‌های ارزیابی بر پایه‌ی این گردایه نیز نمی‌توان چندان اعتماد نمود.

4.6- بازیابي اطلاعات

پس از پیاده‌سازی کارهای پیش پردازشی و آماده شدن گردایه‌ی ویژه‌ی بازیابی اطلاعات، سیستم بازیابی اطلاعات بر پایه‌ی نظریه‌های فصل اول طراحی و پیاده‌سازی شد. این سیستم به کمک زبان برنامه نویسی python و بسته‌ی عددی آن (numarray package) آماده شد. در آغاز همه‌ی واژه‌های فارسی درون صفحه‌های html در گردایه‌ی بازیابی اطلاعات به کمک برنامه‌نویسی بیرون کشیده شدند. سپس دو ماتریس سندـواژه ساخته شد.

الف- ماتریس سندـواژه بدون به کارگرفتن بخش پیش پردازشی (برداشتن واژه‌های سراسری و ریشه‌یابی).

ب- ماتریس سندـواژه به همراه بخش پیش پردازشی.

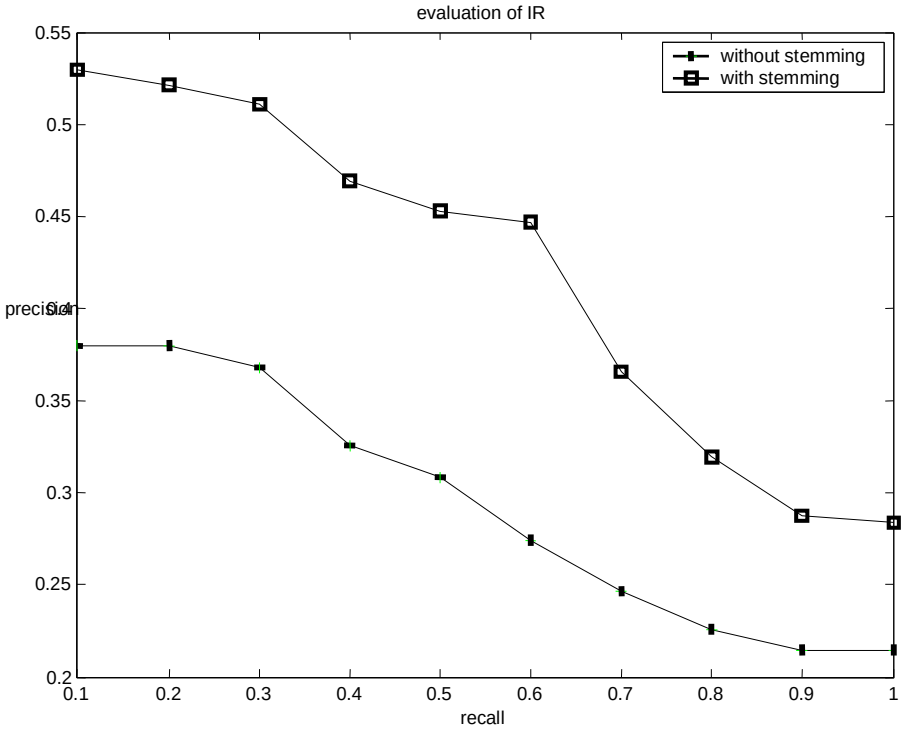
با این دو ماتریس امکان سنجش بخش پیش پردازشی فراهم شد. برای وزن دهی به واژه‌ها روش ساده‌ی فراوانی واژه به کار گرفته شد. با به کارگیری روشهای دیگر وزن دهی از کارآیی سیستم کاسته می‌شد. دلیل احتمالی این رفتار سیستم، استاندارد نبودن گردایه‌ی بازیابی اطلاعات است. دو ماتریس پرس و جوـواژه نیز آماده شد. همانند دو ماتریس سندـواژه، تنها فرق این دو ماتریس در بخش پیش پردازشی است.

به کمک تابع شباهت، اندازه‌ی نزدیکی هر سند به هر پرس و جو روشن شد. سپس برای هر پرس و جو تعداد 25 سند با بیشترین نزدیکی به آن پرس و جو برگزیده شد (DCV=25). از سیستم دو دسته خروجی گرفته شد.

الف- خروجی بدون کمک گرفتن از بخش پیش پردازشی

ب- خروجی به کمک بخش پیش پردازشی

به کمک نگاشت میان سندها و پرس و جوها (آماده شده در گردایه‌ی بازیابی اطلاعات) خروجیهای سیستم ارزیابی شد. ارزیابی دو خروجی سیستم در نمودار زیر نشان داده شده است.



شکل 4-3- ارزیابی دو سیستم بازیابی اطلاعات ساخته شده

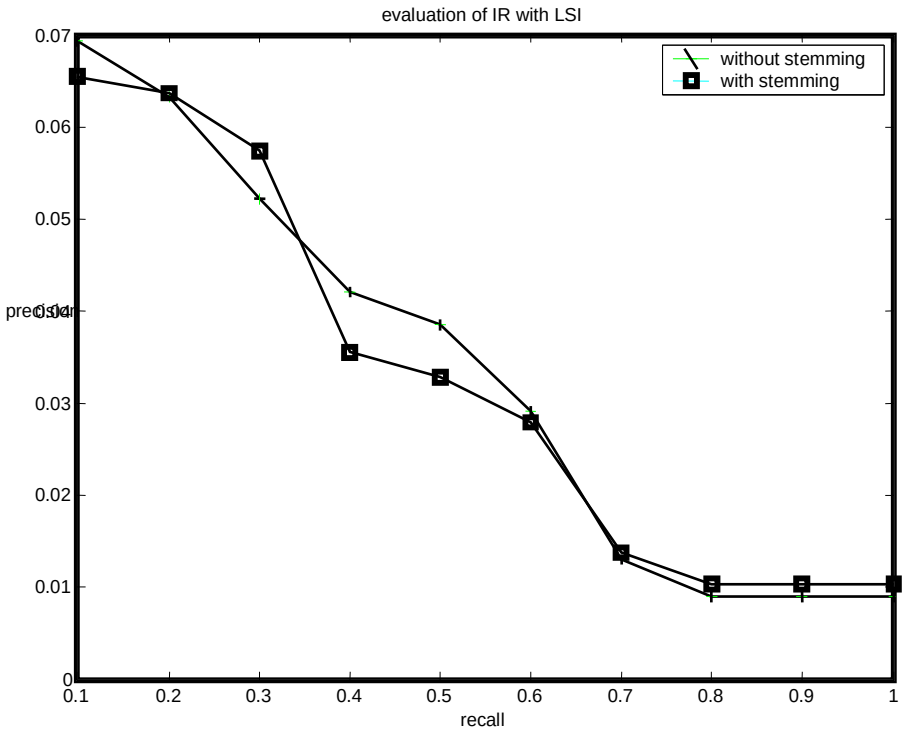
همانگونه که نمودار نشان می دهد، کارآیی سیستم به کمک ریشه‌یابی و برداشتن واژه‌های سراسری یک بهبود کلی را نشان می دهد.

4.7- نمایه گذاري معاني پنهان

برای پیاده‌سازی این بخش نیز زبان python به کار گرفته شد. برای ساخت ماتریس SVD از تابع Singular_value_decomposition از تابع‌های کمکی numarray در بسته‌ی Linear_algebra کمک گرفته شد. به کمک روش گفته شده در فصل دوم این بخش پیاده‌سازی گردید. از ماتریس سندـواژه، SVD گرفته شد. ماتریس پایانی تنها با برداشتن صفرهای بر روی قطرِ ماتریس $\sum$ به دست آمد.

با برداشتن مقدارهای ویژه‌ای کوچک از ماتریس $\sum$ ، به شدت از کارآیی سیستم کاسته می‌شد. سپس ماتریس $\hat{A}$ ساخته شد. ماتریس $\hat{A}$ به دست آمده به جای ماتریس

سندـواژه در بازیابی اطلاعات به کار گرفته شد. همانگونه که در بند پیش گفته شد، دو ماتریس سندـواژه آماده گردید. یکی از ماتریسها همراه بـا ریشـه‌یابی و برداشـتن واژه‌هـای سراسـری و دیگری بدون آن، آماده شده است. نمایه گذاری معانی پنهان جداگانه بر روی هـر کـدام از ایـن دو ماتریس (ترانهاده‌ی آن دو) به کار گرفته شد. در شکل زیر ارزیابی نمایه گذاری معانی پنهان برای این دو ماتریس نشان داده شده است.



شکل 4-4- ارزیابی سیستم بازیابی اطلاعات به همراه نمایه گذاری معانی پنهان

دلیل این رفتار سیستم، استاندارد نبودن گـردایه‌ی بازیـابی اطلاعـات اسـت. در آمـاده شدن گردایه به معنای نوشته‌ها و گزینش پرس و جوهایی بر پایه‌ی معنا پرداخته نشـده اسـت. برای همین هر روشی که بخواهد به معنای نوشته‌ها نزدیک شود، پاسخ شایسته‌ای بر روی ایـن گردایه نخواهد داد.

فصل پنجم

نتيجه گيري و پيشنهادها

5- نتیجه گیری و پیشنهادها

در این پایان نامه به کارهای گوناگونی پرداخته شد. به دشواریهای زبان فارسی و دشواریهای رایانه‌ای آن پرداخته شد ؛ امید است با همکاری استادان زبان و ادبیات فارسی و زبان‌شناسی به دشواریهای آورده شده در این پایان نامه بیشتر پرداخته شود. همچنین از همه‌ی دست اندرکاران کارهای رایانه‌ای (از دانشجویان، سازندگان صفحه‌های فارسی بر روی شبکه‌ی جهانی typist_ها و . . .) خواهشمند است برای پیشبرد زبان فارسی، قراردادهای گذاشته شده تا کنون را رعایت کنند.

در این پایان نامه یک شبه‌زبان نوین برنامه‌نویسی طراحی و پیاده سازی شد. بزرگترین ویژگی این شبه‌زبان ساده بودن آن است. این ویژگی کمک می‌کند تا متخصصان زبان فارسی بتوانند در انجام پروژه‌های از این دست ، همکاری بیشتری با متخصصان رایانه داشته باشند. با به کارگیری و گسترش این شبه زبان، انجام بسیاری از پروژه‌های رایانه‌ای فارسی ساده‌تر می‌شود. همان گونه که با به کارگیری این شبه زبان بخشهایی از این پایان نامه بسیار ساده‌تر انجام شد. به کمک این شبه‌زبان فهرست واژه‌های سراسری کامل گشت. با افزودن واژه‌های سراسری ساده به فهرست دستورهایی که برای ساخت واژه‌های سراسری ترکیبی با این شبه زبان نوشته شده است ، می توان فهرست کاملتری از واژه‌های سراسری ترکیبی را به دست آورد. گردایه‌ای از واژه‌ها از شبکه‌ی جهانی گردآوری شد. به کمک این واژه‌ها، درستی واژه‌های ترکیبی ساخته شده به کمک شبه‌زبان تا اندازه‌ای بررسی شد. دقت فرمایید که همانند بسیاری از کارهای مهندسی در اینجا نیز تقریب کردن و نپرداختن به برخی از دشواریها راهگشای بسیاری از کارها است. به همین خاطر در همه‌ی سیستمها، همواره کم کردن از درصد خطای سیستم هدف است و هیچگاه خطای صفر نداریم. مطلق نبودن دانش بشری یکی از ویژگیهای ذاتی دانش بشری است.

به کمک روشی که از کتابهای دستور زبان برداشت شد، راه شایسته‌ای برای شناسایی خودکار واژه‌های سراسری پیاده سازی شد. به کمک قاعده‌های دستور زبان مشتقهای همه‌ی فعلهای فارسی به دست آمد. باز به کمک گردایه‌ی واژه‌های گرفته شده از شبکه‌ی جهانی

درستی این مشتقها بررسی گردید. سر آخر فهرستی از مشتقها و نگاشت آنها به ریشه‌شان آماده گردید. بدین گونه روشی برای ریشه‌یابی پیاده سازی شد. این روش بر خلاف روشهای متداول ریشه‌یابی جهتی وارون برای ریشه یابی پیشنهاد می‌دهد.

نتیجه‌ی به دست آمده از این روش بسیار خشنود کننده بود. در حالی که روشهای سر راست آزموده شده بخوبی روشن کنونی نبودند و درصد خطای بیشتری داشتند. به کارگیری روش وارون در یافتن پاسخ برای مسأله‌های ریاضی و مهندسی متداول است.

با شناسایی واژه‌های سراسری و ریشه‌یابی، بخش بزرگی از کارهای پیش پردازشی آماده شد. گردایه‌ای از سندها و پرس و جوها آماده شد. بدین گونه داده‌های ورودی سیستم بازیابی آماده شدند به کمک نگاشت فراهم شده میان سندها و پرس و جوها در این گردایه کارآیی سیستم آزموده شد. به دلیل کمبود زمان و نیاز به همکاری متخصصانی از رشته‌ی کتابداری و رشته‌های همانند آن، گردایه‌ای که یک نفری آماده شد، نمی‌توانست کارآیی سیستم بازیابی اطلاعات را بخوبی ارزیابی کند.

به کمک روش فضای برداری یک سیستم بازیابی اطلاعات نوشتاری فارسی پیاده‌سازی شد. می توان این گونه پنداشت که دو سیستم بازیابی پیاده‌سازی شد.

الف- سیستم بازیابی اطلاعات نوشتاری بدون کمک گرفتن از بخش پیش پردازشی.

ب- سیستم بازیابی به کمک ریشه‌یابی و برداشتن واژه‌های سراسری.

کارآیی این دو سیستم ارزیابی شد. پاسخ سیستم «ب» یک بهبود کلی را نسبت به سیستم «الف» نشان می‌داد. گرچه به دلیل فراهم نبودن گردایه‌ای شایسته از سندها و پرس و جوها نمی‌توان چندان به نتیجه‌ی به دست آمده خوشبین بود. سپس نمایه گذاری معانی پنهان به سیستم بازیابی افزوده شد. نتیجه‌ی ارزیابی به دست آمده از دو سیستم بازیابی «الف» و «ب» همراه با نمایه گذاری معانی پنهان بخوبی نشان دهنده‌ی خوب نبودن گردایه‌ی سندها و پرس و جوها بود. بنابراین شاید یکی از پایه‌ای‌ترین کارها در زمینه‌ی بازیابی اطلاعات نوشتاری فارسی آماده نمودن گردایه‌ای ویژه‌ی بازیابی اطلاعات باشد، بگونه ای که بتوان بخوبی سیستمهای بازیابی اطلاعات نوشتاری فارسی را ارزیابی نمود.

در زیر پیشنهادهایی برای ادامه‌ی کار داده شده است. گرچه گستردگی این پایان نامه این اجازه را می‌دهد که پیشنهادهای فراوانی نوشته شود، ولی در اینجا تنها پاره‌ای از آنها نوشته شده است.

1- آماده کردن یک گردایه‌ی ویژه‌ی شایسته برای بازیابی اطلاعات نوشتاری فارسی. همانگونه که گفته شد، برای انجام این کار نیاز به همکاری گروههایی از متخصصان

رشته‌های گوناگون است. انجام این کار پایه‌ی دیگر پیشنهادها است. زیرا دیگر پیشنهادها وابستگی بسیاری به این پیشنهاد دارند.

2- به کار گرفتن دیگر روشهای وزن دهی واژه در روش فضای برداری برای زبان فارسی و ارزیابی سیستمهای آماده شده با هر کدام از این روشهای وزن گذاری واژه و نشان دادن بهترین روش وزن دهی برای بازیابی اطلاعات نوشتاری فارسی.

3- کمک گرفتن از بازخورد ارتباطی برای بهبود کارآیی سیستم بازیابی اطلاعات فارسی، روشهایی هوشمند برای بهبود بازخورد ارتباطی وجود دارد.

4- به کارگرفتن تابعهای شباهت دیگر، برای بازیابی اطلاعات فارسی همچنین آزمون روشهای هوشمند برای ساخت خودکار تابع شباهت بهینه.

5- به کارگیری مدلهای دیگر برای بازیابی اطلاعات نوشتاری فارسی برای هر مدل دستِ کم باید یک پایان نامه تعریف شود. برای نمونه مدل بر پایه‌ی نظریه‌ی احتمال از دسته مدلهای پایه‌ی بازیابی اطلاعات است.

6- با آماده شدن گردایه‌ی ویژه‌ی بازیابی اطلاعات، روش نمایه گذاری معانی پنهان و روشهای همانند آن بر روی سیستم بازیابی اطلاعات فارسی به کار گرفته شود و کارآیی آنها ارزیابی شود.

با گذراندن دوره‌ی کارشناسی ارشد در گرایش هوش مصنوعی برایم روشن شد که درجه‌ی هوش بسیار پایینی بر پایه‌ی تعریف درجه‌ی هوش دارم. همانگونه که می‌دانید بر پایه‌ی تعریف کنونی از درجه‌ی هوش، درجه‌ی هوش در انسان و در هوش مصنوعی رابطه‌ی مستقیمی با اندازه‌ی سازگاری موجود هوشمند با پیرامون خود دارد. سازگار شدن و کنار آمدن با پیرامون، در واقع به معنای پذیرفتن هر آنچه در پیرامون موجود هوشمند وجود دارد، است. همان گونه که می‌دانیم سازگاری انسان با همه‌ی کُنشهای پیرامون خود چندان شایسته نیست. یک انسان به همان اندازه که خوبیهای پیرامون را یاد می‌گیرد، از کُنشهای نادرست، پیرامون خود می‌پرهیزد. پایه‌ی درستی یا نادرستی کُنشهای پیرامون، اغلب چیزهایی ورای سود انسان است، نه یک هدف روشن و مشخص. به نظر می‌رسد آنچه بهتر است در آغاز برای شناخت درستی یا نادرستی کُنشی تصمیم گیری کند، احساس است و نه عقل حسابگر. گرچه برای انجام کار یا چگونگی پذیرفتن یا نپذیرفتن کُنشی از پیرامون، عقل می تواند یاور خوبی در کنار احساس باشد. به نظر می‌رسد هوش زیاد (بر پایه‌ی تعریف کنونی از آن) به تنهایی زندگی زیبایی را برای انسانها فراهم نمی‌کند. همزمان مقدار زیادی از عنصرهای بنیادی انسانی برای زندگی نیاز است.

فهرست منابع فارسی

- آریا ، غلامعلی ، و منصور استخری (۱۳۶۵). برگزیده‌ی ادب فارسی سال اول مراکز تربیت معلم. تهران: شرکت چاپ و نشر ایران . صفحه‌ی ۱۵.
- احمدی بیرجندی ، احمد (۱۳۶۸). شیوه‌ی آموزش املای فارسی و نگارش . تهران: انتشارات وزارت فرهنگ و ارشاد اسلامی . چاپ اول بهار ۱۳۶۸.
- احمدی گیوی ، دکتر حسن ، و دکتر حسن انوری(۱۳۸۰). دستور زبان فارسی ۲. تهران: انتشارات فاطمی ، چاپ بیست و یکم ۱۳۸۰.
- بهار ، محمد تقی ، و غلامرضا رشید یاسمی ، بدیع الزمان فروزانفر ، عبدالعظیم قریب ، جلال الدین همایی (۱۳۴۲) . دستور زبان فارسی دوره اول و دوم و سوم دبیرستانها .تهران: شرکت سهامی کتابهای درسی ایران . فصل چهارم ، فعل ، صفحه ۱۱۱.
- جباری‌فر ، معصومه (۱۳۸۰). «بررسی پارامترهای ارزیابی و لیست دسته بندی شده جویشگرها». تهران: مرکز تحقیقات مخابرات ایران ، پاییز ۱۳۸۰. از سایت http://www.itrc.ac.ir/Reports/j-info/1380/8032141/5.doc
- حق شناس ، محمد علی ، و دیگران (۱۳۷۷) . زبان فارسی (۱) و (۲) ، سال اول نظام جدید آموزش متوسطه . تهران: شرکت نشر و چاپ کتابهای درسی ایران.
- خاتون آبادی ، سید احمد (۱۳۸۰). استخراج اطلاعات خاص از اسناد نیمه ساخت یافته با استفاده از یادگیری ماشین . شیراز: پایان نامه‌ی کارشناسی ارشد کامپیوتر دانشگاه شیراز. آذر ماه ۱۳۸۰.
- دستغیب ، محمد باقر (۱۳۸۰). بازیابی اطلاعات متنی با استفاده از منطق فازی. شیراز: پایان نامه‌ی کارشناسی ارشد کامپیوتر دانشگاه شیراز. شهریور ماه ۱۳۸۰.

- دفتر تحقیقات و برنامه ریزی تربیت معلم (بدون نام نویسنده و نویسندگان ۱۳۶۶). آیین نگارش و خلاصه دستور زبان فارسی سال دوم تربیت معلم . تهران: شرکت چاپ و نشر کتابهای درسی ایران . صفحه ۲۵.
- رحیمیان، دکتر جلال(۱۳۸۱). دستور زبان فارسی: صورت نقش معنا. شیراز:مرکز نشر دانشگاه شیراز ، چاپ یکم ، ۱۳۸۱.
- رزمجو ، حسین ، و محمد مهدی رکنی ، احمد احمدی بیرجندی ، ابراهیم عقدایی ، حسن انوری (۱۳۷۲) . فارسی سال چهارم دبیرستان . تهران: شرکت چاپ و نشر کتابهای درسی ایران.
- رشید یاسمی ، غلامرضا ، و ملک الشعرا بهار ، بدیع الزمان فروزانفر ، عبدالعظیم قریب ، جلال الدین همایی (۱۳۶۸) . دستور زبان فارسی پنج استاد . تهران: انتشارات کتابفروشی مرکزی . چاپ هشتم ، ۱۳۶۸، صفحه‌ی ۴.
- زارع بیدکی ، علی محمد ، و مازیار صالحی. «ارائه‌ی RFP برای یک جویشگر دو زبانه فارسی/ انگلیسی». تهران: مرکز تحقیقات مخابرات ایران ، گروه جامعهٔ اطلاعاتی ، پروژه‌ی فارسی سازی ،زمستان ۱۳۸۰.

  از سایت http://www.itrc.ac.ir/Reports/j-info/1380/8032141/8.doc
- سخاوت ، عباس ، و احمد آزادی اردکانی (۱۳۴۴). انشاء دو دبیر . اصفهان: انتشارات کتابفروشی شهسواری . چاپ اول ، صفحه‌ی ۲۱۹.
- سمیعی ، دکتر احمد ، و دیگران (۱۳۷۸). ادبیات فارسی سال سوم دبیرستان(نظام جدید). تهران: مرکز چاپ و نشر کتابهای درسی ایران.
- سمیعی گیلانی ، احمد ، و دیگران (۱۳۸۰). زبان فارسی (۳) سال سوم آموزش متوسطه . تهران: شرکت چاپ و نشر کتابهای درسی ایران . چاپ سوم ۱۳۸۰ ، صفحه‌ای ۲.
- شریعت ، محمدجواد (۱۳۷۰). آیین نگارش. تهران: انتشارات اساطیر. چاپ پنجم ۱۳۷۰ ، صفحه ۱۹۷.
- شریعت ، محمد جواد ، و سمیعی گیلانی احمد ، و دیگران (۱۳۸۲) . فارسی سال سوم دوره‌ی راهنمایی تحصیلی . تهران: شرکت چاپ ونشر کتابهای درسی ایران . چاپ هشتم ۱۳۸۲.

- شکری ، دکتر یدالله ، و دیگران (۱۳۷۲). زبان و آیین نگارش فارسی . تهـران: سـازمان مطالعه و تدوین کتابهای علوم انسانی (سمت) ، چاپ ششم زمستان ۱۳۷۲.
- شورای تحقیقات دانشگاه پیـام نـور (۱۳۷۸). فارسـی عمـومی. تهـران: مرکـز چـاپ و انتشارات دانشگاه پیام نور. چاپ پنجم آبان ماه ۱۳۷۸.
- صنعتی ، دکتر محمد (۱۳۷۱). دشواری های زبان فارسی با کامپیوتر. تهـران: ماهنـامه صنعت چاپ ، مرداد - شهریور ۱۳۷۱.
- علمدار میلانی ، امید ، و دیگران(۱۳۸۱). «فناوری اطلاعات ، تبادل و شـیوه‌ی نمـایش اطلاعات فارسی بر اساس یونی‌کد». از سایت http://farsiweb.info/standard
- غلامعلی زاده ، دکتر خسرو (مترجم) ، نویسندگان: آندرین آکماجین و ریچارد آدامـز و ان ک فارمر (۱۳۷۵). زبان شناسی (درآمدی بر زبان و ارتبـاط). کرمانشـاه: انتشـارات طاق بستان ، دانشگاه رازی ۱۳۷۵.
- فرهودی ، مژگان (۱۳۸۰). "طراحی و پیاده سازی نرم افزاری بـرای تشـخیص اعتبـار و نوع کلمات فارسی و انگلیسی در جویشگر.

  از سایت http://www.itrc.ac.ir/Reports/j-info/1380/8032141/2.doc
- گروه مؤلفان (بدون نام نویسنده یا نویسندگان ۱۳۷۹) . فارسی پنجم دبسـتان. تهـران: شرکت چاپ و نشر کتابهای درسی ایران . صفحه‌ی ۳۵.
- مولوی ، عباسعلی (۱۳۵۹) . دستور زبان فارسی . تهران: چاپخانه اتحاد.
- وحیدیان کامکار ، تقی ، و دیگران (۱۳۷۸). زبان فارسی (۳) و (۴) نظام جدید آموزش متوسطه. تهران: شرکت چاپ و نشر کتابهای درسی ایران . صفحه‌ی ۵۶.
- یارمحمدی ، دکتر لطف الله (۱۳۷۴). پانزده مقـاله در زبانشناسـی مقـابله‌ای و سـاخت زبان فارسی: دستور،متن و گفتمان. تهران: انتشارات رهنما.

فهرست منابع انگلیسی

- Arms , William Y.(2002). "Latent Semantic Indexing". URL : http://courses.cs.cornell.edu/cs430/2002fa/slides/lecture11.ppt.

- Baeza-Yates, Richardo, Berthier Ribeiro-Neto. modern information retrieval. New York, ACM Press, Addison-Wesley, 1999.

- Belew, Richard(2000). Finding Out About: Search Engine Technology From A Cognitive Perspective. Cambridge University Press.

- Berry, Michael(1992). "Large scale singular value computations". international journal of supercomputer Applications, Vol.6, pp.13-49.

- Berry, Michael, Susan T. Dumais, Gavin W. O'Brien(1995). "using linear algebra for intelligent information retrieval". SIAM Rev. Vol 37, PP 573-595.

- Berry, Michael, Susan T. Dumais(1996). "Computational models for intelligent information access". proceeding of supercomputing, San Diego, CA December 1996.

- Burger, Thomas Wolfgang(2001). "How to incorporate and utilize Unicode for foreign language support". URL: http://www-106.ibm.com/developerworks/linux/library/l-linuni.html .

- Davis, Mark(2003). "The Bidirectional Algorithm". URL: http://www.unicode.org/reports/tr9/ .

- Dumais, Susan T.(1991). “Improving the retrieval of information from external sources”, Behavior Research Methods, Instruments, and Computers, 23: 229-236.

- Golub, Gene, Charlie Van Loan(1996). Matrix Computations. John hopkins university press.

- Honglan, JIN, WONG Kam Fai William(2002). "A Chinese Dictionary Construction Algorithm for Information Retrieval", ACM Transactions on Asian Language Information Processing vol.1 no.4, pp.281-296. 2002.12.

- Kowalski, Gerald. Information retrieval systems: Theory and Implementation. kluwer publisher, 1998.

- Kuhn, Markus(2003). "UTF-8 and Unicode FAQ for Unix/Linux". URL: http://www.cl.cam.ac.uk/~mgk25/unicode.html.

- Luhn, Hans Peter(1957). "A statistical approach to mechanized information encoding and searching of literary information". IBM journal of research and development, vol .1, no 4, page 309.

- Manning, Chris, Hinrich Schütze(1999). Foundations of Statistical Natural Language Processing. MIT Press. Cambridge, MA: May 1999.

- Matthew, Neil, Richard Stones, others(2000). Professional Linux Programming. Wrox Press Ltd.

- Megerdoomian, Karine, Rémi Zajac (2000). "Processing Persian Text:Tokenization in the Shiraz Project". URL: http://www.crl.nmsu.edu/shiraz/ .

- Nakos, George, david joyner(1998). linear algebra with applications. Brooks/cole publisshing company.

- Porter, M.F.(1980). "An algorithm for suffix stripping". Program, volume 14, number 3, 130-137.

- Salehi, Maziar(2000). An Intelligent Agent-Oriented Structure for Text Information Retrieval, MSC thesis,Shiraz university.

- Salton, Gerard, Michael J. McGill(1983). An introduction to Information retrieval. McGraw Hill.

- Scherer, Markus(2000). "A brief introduction to code pages and Unicode". URL: http://www.ibm.com/developerworks/library/codepages.html

- Tang , Chunqiang, Zhichen Xu, Sandhya Dwarkdas, Mallik Mahalingamzhichen(2003). "Distributed Content-based Search on Structured Peer-to-Peer Overlay Networks". URL: http://www.hpl.hp.com/news/events/csc/2003/zhichen_slides.ppt

Abstract

A FARSI TEXT INFORMATION RETRIEVAL SYSTEM
BASED ON LATENT SEMANTIC INDEXING

BY:

AHMAD YOUSEFAN NAJAFABADI

As the amount of digital information in the form of text is increasing on a daily basis, the need for accurate and efficient retrieval systems is increasing.

In this thesis, a text retrieval system for Farsi language based on Latent Semantic Indexing is designed and implemented. For the purpose of finding a list of Farsi stop words and stemming, a simple programming language is developed. Based on Farsi grammar, a new method for automatic recognition of verbs is proposed.

The performance of the system with and without stemming is evaluated, using a small corpus prepared for this purpose and using precision and recall as evaluation parameters.

IN THE NAME OF GOD

A FARSI TEXT INFORMATION RETRIEVAL SYSTEM
BASED ON LATENT SEMANTIC INDEXING

BY:

AHMAD YOUSEFAN

THESIS

SUBMITTED TO THE OFFICE OF VICE-CHANCELLOR FOR GRADUATE STUDIES IN PARTIAL FULLFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF MASTER OF SCIENCE (Ms.)

IN

COMPUTER ENGINEERING - ARTIFICIAL INTELLIGENCE

SHIRAZ UNIVERSITY

SHIRAZ, IRAN

EVALUATED AND APPROVED BY THE THESIS COMMITTEE AS: <u>EXCELLENT</u>

……..M.ZOLGHADRI JAHROMI,Ph.D.,ASSISTANT PROF. OF COMPUTER ENGINEERING

…….. H.EGHBLI JAHROMI, Ph.D., ASSOCIATE PROF. OF COMPUTER ENGINEERING

…….. J.RAHEMIAN, Ph.D., ASSOCIATE PROF. OF FOREIGN

LANGUAGE

March 2004