امکان‌سنجی نمایه‌سازی ماشینی مدارک زبان فارسی در مرکز اطلاع‌رسانی جهاد کشاورزی

نوشته: شهرزاد نیاکان *

چکیده

هدف از انجام این پژوهش بررسی فرایند نمایه سازی ماشینی و سنجش امکانات لازم برای استفاده از نمایه سازی ماشینی در مرکز اطلاع رسانی جهاد کشاورزی می باشد تا الگویی مناسب جهت استفاده از نمایه‌سازی ماشینی به زبان فارسی در ایران ارائه گردد.

گردآوری اطلاعات به شیوه مصاحبه و استناد به مدارک موجود انجام گرفته است. از روش طراحی سیستم استفاده شده است؛ به طوریکه پس از مقایسه وضعیت کنونی نمایه سازی در مرکز مورد مطالعه، حداقل امکانات لازم جهت نمایه سازی ماشینی، پیش بینی چگونگی استقرار امکانات لازم و دلایل و عوامل مؤثر برای نمایه سازی ماشینی صورت گرفته است.

امکانات ذکر شده در این تحقیق برای انجام نمایه‌سازی ماشینی در مرکز مزبور شامل استفاده از نمایه‌سازان و متخصصین کامپیوتر و کشاورزی در مرکز و بهره‌گیری از مشاوره زبان‌شناسان و فراهم‌آوری تجهیزاتی جهت ارتقاء نرم‌افزارهای کنونی یا طراحی نرم‌افزاری جدید جهت ذخیره سازی، تشخیص ، تفکیک، مقایسه، طبقه بندی و انتخاب اطلاعات موجود در مدارک می‌باشد که مورد ارزیابی قرار گرفته است و مواردی که برای ذخیره اطلاعات مدارک به زبان فارسی باید مد نظر

* کارشناس ارشد کتابداری و اطلاع‌رسانی، کتابخانه دانشکده صنعت آب و برق (شهید عباسپور)

قرار گیرد بررسی شده و مراحل ساخت پیشنهادی اصطلاحنامه الکترونیکی فارسی در زمینه کشاورزی تشریح شده است. در پایان مراحل پیشنهادی انجام نمایه‌سازی ماشینی در مرکز با استفاده از امکانات موجود برای ذخیره سازی، تشخیص ، تفکیک، مقایسه، طبقه بندی و انتخاب اطلاعات موجود در مدارک با استفاده از روشهای مختلف نمایه‌سازی ماشینی توصیف شده است.

کلیدواژه‌ها: نمایه‌سازی ماشینی، امکان‌سنجی، مدارک فارسی مرکز اطلاع‌رسانی جهاد کشاورزی

مقدمه‌ای بر نمایه‌سازی ماشینی

از سال‌های ۶۰ علاقه به استفاده از مدل‌های برداری جهت بازیابی خودکار اطلاعات آغاز گردید. به گونه‌ای که در این مدل‌های برداری در برابر سئوال مرجع، با کلیدواژه‌ها قابل بازیابی باشند. همچنین تحقیقاتی در پروژه AIR/X توسط لوستیگ[1] و فانگ مایر[2] صورت گرفت که منجر به بین‌المللی شدن این حرکت‌ها در سال‌های ۱۹۶۸ و ۱۹۶۹ گردید. به موازات آن تحقیقاتی در ارتباط با مشکلات نمایه‌سازی براساس روش‌های زبان‌شناسی انجام گرفت.

در آلمان از آغاز سال‌های ۷۰ مطالعات ریخت‌شناسی بر روی سیستم‌های ترجمه با هدف بازیابی اطلاعات آغاز گردید. طرح‌های مختلفی به اجرا در آمد که مهمترین آنها طرحی به نام PASSAT بود که اساساً در زمینهٔ مطالعهٔ زبان‌های کامپیوتری اجرا گردید و داستان جداگانه‌ای در سال ۱۹۶۹ به وجود آورد.

در سال‌های ۸۰ علاقه به سیستم‌های هوشمندی که بسیاری از دانشمندان در جهت تحلیل محتوا به آن اعتماد داشتند، ایجاد شد. از نمونه‌های آن سیستم دانشگاه ییل و نیز برلین در آلمان می‌توان نام برد.

در هر صورت مسلم است که مطالعات در مورد اطلاعات با عملکرد سیستمهای متن آزاد شکل گرفته و به سوی تحقیق دربارهٔ بازیابی اطلاعات یا با عناوین جدید تحلیل هوشمندانه متن، فرا رسانه‌ها، تحلیل سیستم‌های اطلاعاتی روی کرده است. سال‌های اخیر علاقهٔ روزافزون و نیاز فزاینده به نمایه‌سازی ماشینی را تشکیل داده است (کنورتس، ۱۹۹۴ ،ص۳).

طرح‌های موفقی که در اجرای نمایه‌سازی ماشینی ارائه گردیده عبارتند از:

۱- طرح AIMS از کتابخانه ملی پزشکی آمریکا
۲- طرح AIR/X در دانشگاه فنی دارمشتات با عنوان تحلیل عنوان و چکیده مدارک با موضوع فیزیک برای پایگاه اطلاعاتی فیزیک
۳- طرح LISA و طرح COPSY توسط زیمنس اجرا شد .
٤- طرح Indexing Aidsystem در کتابخانه ملی پزشکی آمریکا، سیستمی برای نمایه‌سازی نشریات پزشکی
٥- MAI در مرکز دکومانتاسیون آمریکا، تحلیل عنوان و چکیده مدارک علمی برای نمایه‌سازی
٦- PASSAT توسط زیمنس، سیستم نمایه‌سازی برای زبان آلمانی و دیگر زبان‌ها
۷- SAPHIR در دانشکده پزشکی هاروارد، سیستمی برای یک سیستم بازیابی هوشمند اطلاعات در مدارک بیولوژی و پزشکی بر پایه روش‌های آماری و زبان‌شناسی
۸- SPECIALIST توسط کتابخانه ملی پزشکی در آمریکا، سیستمی برای تحلیل اطلاعات بیولوژیکی و پزشکی بر مبنای تکنیک‌های استفاده از زبان طبیعی
۹- TCS توسط Carnegie group در آمریکا، ابزاری برای نمایه‌سازی تحلیلی و فهرست‌نویسی بر مبنای ارتقای سیستم CONSTRVE/TIS (کنورتس، ۱۹۹۴، ص ۳).

در ایران مطالعاتی به تازگی در این زمینه انجام گرفته است که از جمله می‌توان به مقاله «اصول و روش‌های نمایه‌سازی رایانه‌ای» از آقای احمد یوسفی در فصلنامه کتاب تابستان ۷۷ نام برد که در آن روش‌های نمایه‌سازی رایانه‌ای توضیح داده شده است و پیش‌بینی‌هایی در مورد آیندهٔ نمایه‌سازی ماشینی انجام گرفته است.

تعریف نمایه‌سازی ماشینی

بر اساس استاندارد نمایه‌سازی(BS3700:1988)[3] نمایه مجموعه‌ای منظم از شناسه‌های نشانه‌گذاری شده می‌باشد برای اینکه کاربران را قادر سازد اطلاعاتی که محل آنها مشخص شده را در مدرک پیدا کنند. فرایند ساخت نمایه، نمایه‌سازی نامیده می‌شود و شخصیکه به اینکار مبادرت می‌نماید نمایه‌ساز نامیده می‌شود(امریکن سوسایتی آو ایندکسرز، ۱۹۹۴).

در تعریفی دیگر نمایه مجموعه شناسگرهایی است که توصیفگر مدرکی و چکیده‌ای یا پاره اطلاعاتی دیگری هستند. این شناسگرها بر اساس نیاز اطلاعاتی کاربران به شیوه‌های گوناگون از قبیل الفبایی و رده‌ای و غیره منظم میشوند و نمایه‌سازی یعنی فرایند تهیه نمایه؛ به عبارت دیگر فرایند تحلیل محتوای اطلاعاتی اثر و بیان کردن آن با زبان ویژه نظام نمایه‌سازی است(کینن،۱۳۸۰،ص ۱۳).

نمایه‌سازی فعالیت اصلی هر سیستم اطلاعاتی می‌باشد. در حین نمایه‌سازی، جهت انتخاب توصیفگرهایی که نماینده محتوای یک مدرک هستند، مدرک را تجزیه و تحلیل می‌نمائیم. هر توصیفگر، نشانگر ارزش معنایی یک بخش از مدرک جهت بازیابی است در نمایه سازی مراحل زیر باید در نظر گرفته شود.

- کنترل اصطلاحات تخصصی
- تعیین توصیفگرها
- هماهنگی و تناسب توصیفگرها

هنگامی که توصیفگرها توسط نمایه‌ساز انتخاب می‌شوند از نمایه‌سازی دستی سخن به میان می‌آید و اگر این کار را کامپیوتر انجام دهد، نمایه‌سازی خودکار یا ماشینی نامیده می‌شود و در صورتی که این کار توسط انسان و کامپیوتر مشترکاً انجام گیرد، نمایه‌سازی به کمک کامپیوتر می‌نامیم. به عنوان مثال هنگامی که نمایه‌ساز از میان توصیفگرهایی که توسط کامپیوتر پیشنهاد شده انتخاب می‌کند.(وایس، ۲۰۰۱،ص۲)

جهت استفاده از روش نمایه سازی ماشینی باید داده‌ها به صورت ماشین‌خوان درآیند. اینکه واژه‌ها از کدام محدوده از متن انتخاب شوند، بستگی به نرم‌افزار دارد. شرکت زافتکس[4] نوعی برنامهٔ کامپیوتری با عنوان IDX طراحی کرده که به ارائه خدمات نمایه‌سازی ماشینی می‌پردازد. این نرم‌افزار، اصولی برای آزمایش بازیابی ساخته است که در مقاله گرومان کاملاً تشریح گردیده است. این نرم‌افزار از روش استفاده از واژه‌نامه بهره می‌برد. تغییرات واژگانِ متن توسط مطابقت دادن با واژه‌نامه‌های مختلف الکترونیکی اعمال می‌شود. همچنین این نرم‌افزار امکان تبدیل واژگان به ریشهٔ آنها را جهت بازیابی بعدی فراهم می‌کند و علامتگذاری و محدودکردن واژه‌های ناخواسته[5] را انجام می‌دهد. شکستن واژه‌های مرکب و ترجمه و انجام عمل ارجاع و مترادف‌سازی و ساخت عبارات را نیز انجام می‌دهد (گرومان، ۲۰۰۰، ص۲۹۹).

بنابراین در صورتی که توصیفگرها به صورت خودکار از مدارک استخراج شده یا تولید گردند به آن نمایه‌سازی ماشینی گفته می‌شود (کایزر، ۱۹۹۳، ص۱۹).

در صورتی که در نمایه‌سازی دستی، مراحل نمایه‌سازی توسط شخصِ نمایه‌ساز انجام می‌پذیرد. شخص نمایه‌ساز توصیفگرهای هر مدرک را که محتوای مدارک را به خوبی توصیف می‌کنند، طبقه‌بندی می‌کند.

بررسی وضعیت نمایه‌سازی مدارک در مرکز اطلاع‌رسانی جهاد کشاورزی

کلیه مدارکی که تا خرداد ۱۳۸۱ در این مرکز نگهداری می شود ۲۸۷۳۱ فقره و شامل موارد زیر می باشد:

ردیف	مدارک	تعداد
۱	اسناد لاتین	۸۲۹۱ عنوان
۲	اسناد فارسی	۱۰۶۰ عنوان
۳	اسناد سازماندهی نشده فارسی و لاتین	۱۹۰۰۰ عنوان
۴	مقالات لاتین	۴۶۳۶ عنوان
۵	مقالات فارسی	۱۲۷۷۱ عنوان
۶	مقالات فارسی و لاتین سازماندهی نشده	۱۲۰۰۰ عنوان
۷	نشریات لاتین	۷۷۱ عنوان
۸	نشریات فارسی	۱۰۶۰ عنوان
۹	پایان‌نامه‌های لاتین	۱۰ عنوان
۱۰	پایان‌نامه‌های فارسی	۲۰۲ عنوان
۱۱	دیسک‌های فشرده	۳۰۰ عنوان
۱۲	فیلم‌های برگزاری جلسات	۱۰۰ عنوان

۱۳	نوارهای کاست آموزشی	۱۵ عنوان
۱۴	پروانه‌های ثبت اختراع	۲۵۰۰ عنوان

اسناد فارسی شامل گزارش‌های دولتی و تحقیقاتی می باشد که در زمینه کشاورزی نوشته شده است . تعدادی نقشه و پوستر نیز در مرکز موجود است که هنوز مشخصات کتابشناختی آنها به پایگاه اصلی منتقل نشده است.

همچنین تعدادی بانک اطلاعاتی در مرکز موجود است که از طریق شبکه به صورت تمام متن قابل دسترسی است.

نتایج حاصل از بررسی وضعیت نمایه سازی در مرکز حاکی از آن است که بسیاری از اسناد موجود در مرکز مزبور هنوز نمایه‌سازی نشده است و برای برخی از اسنادی که نمایه سازی شده‌اند از نمایه‌سازی دستی استفاده می‌شود. در ذخیره و بازیابی اطلاعاتِ مدارکِ نمایه سازی شده، از کامپیوتر استفاده می شود. از دو نرم افزار نوسا و پارس آذرخش که هردو تحت سیستم عامل داس کار می‌کنند در ذخیره سازی و بازیابی اطلاعات استفاده می‌شود. از شیوهٔ پیش همارا در برخی از مدارک و از پس همارا دربرخی دیگر از مدارک استفاده میگردد. از اصطلاحنامهٔ CAB,Agrovoc برای مدارک انگلیسی و از بانک واژگان موجود در نرم افزار و سرعنوانهای موضوعی کتابخانهٔ ملی برای مدارک فارسی استفاده می شود. از یک کارشناس ارشد کتابداری و یک دیپلمه کار نمایه سازی را انجام می دهند. نمایه سازان درفضایی مناسب و با ابزاری مناسب به کار مشغولند. مشکلات آنها عدم وجود اصطلاحنامهٔ فارسی در زمینهٔ کشاورزی و فیلدهای نامناسب برای ذخیره سازی محتوای اسناد میباشد. بودجهٔ خاصی به این بخش اختصاص داده نشده است.

بررسی امکانات لازم جهت نمایه‌سازی ماشینی در مرکز اطلاع‌رسانی جهاد کشاورزی

نیروی انسانی

استفاده از متخصص کامپیوتر جهت برنامه‌نویسی ضروری است و استفاده از زبانشناس در شناخت ساختار زبان مدرک جهت تهیه نرم‌افزار لازم و در صورت استفاده از روشهای زبانشناسی امری مسلم است و متخصصین در زمینه کشاورزی در جهت دقت نظر درکنترل واژگان تخصصی در تهیه نرم‌افزار و همچنین جهت ساخت اصطلاحنامه لازم است. همچنین استفاده از متخصصین علوم کتابداری و اطلاع‌رسانی در ایجاد ارتباط مناسب بین متخصصین کامپیوتر ، زبان‌شناسی و متخصصین موضوعی ضروری است؛ چرا که متخصصین کامپیوتر با هدایت کتابداران می‌توانند برنامه جامع و کاملی ارائه دهند و متخصصین موضوعی اگر در زمینه استفاده بهینه از اصطلاحنامه توسط کتابداران آموزش ندیده باشند نمی‌توانند بازده مناسب داشته باشند.

فضا

آنچه از بررسی منابع یافت نگردید فضای مناسب جهت استفاده از نمایه‌سازی ماشینی است گویا مقدار فضا برای این امر بسته به شرایط تغییر می‌کند و فضای خاصی در این مورد در نظر نمی‌باشد.

تجهیزات

تجهیزات لازم جهت نمایه‌سازی ماشینی شامل موارد زیر است:

- سیاهه بازدارنده

- واژه‌نامه ریشه کلمات یا اصطلاحنامه تخصصی که برای ساخت آن لازم است مجموعه بزرگی از مدارک که قبلاً نمایه‌سازی دستی در مورد آنها انجام گرفته است و نمایه‌ای از واژگان موجود در مدارک نمایه‌سازی شده در دسترس باشد.

- نرم‌افزاری با قابلیت محدودیت سازی در سیاهه باز دارنده، تعیین شکل اصلی لغت، تفکیک لغات تشکیل دهنده کلمات مرکب، تبدیل کلمات صرف شده و صفات و ضمایر به صورت اسم، شناسایی کلمات مرکب و عبارات، ترجمه کلمات به زبانهای مختلف، توانایی برقراری ارتباط بین کلمات مترادف(ارجاع دهی) (نور، ۱۹۹۹) و علاوه بر شناسایی انواع لغات، اسامی و ریشه لغات، پیشوندها و پسوندها، اسامی صرف‌نظر شده را می‌شناسد و به تجزیه کلمات مرکب و عبارات و ترجمهٔ مصدرها، افعال صرف‌شده

و تبدیل آنها به حالت اسم مصدر می‌پردازد(بریل مایر، ۱۹۹۷).

بودجه

بودجه مورد نیاز جهت برنامه‌نویسی، تهیه نرم‌افزارها و سخت‌افزارهای لازم در وهله اول و روزآمدسازی واژه‌نامه و سیاهه بازدارنده در وهله دوم همچنین مشاوره با زبانشناسان و آموزش نمایه‌سازان جهت کنترل باید تأمین گردد.

بررسی امکانات موجود در مرکز برای نمایه‌سازی ماشینی

با توجه به بررسی انجام شده در مورد وضعیت و امکانات موجود در مرکز اطلاع‌رسانی جهاد کشاورزی، امکانات موجود در مرکز جهت انجام نمایه‌سازی ماشینی به طور خلاصه به شرح زیر می‌باشد:

نیروی انسانی

- یک نفر نیروی انسانی متخصص در نمایه‌سازی از لحاظ آکادمیک و تجربی
- یک نفر نیروی انسانی متبحر در نمایه‌سازی از لحاظ تجربی
- متخصصین کامپیوتر شاغل در بخش مهندسی اطلاعات (با توجه به ساختار تشکیلاتی مرکز) که در امر برنامه‌نویسی کامپیوتری می‌توانند کمک مؤثری باشند
- متخصصین کشاورزی شاغل در مرکز که در بررسی اصطلاحات مناسب در زمینهٔ کشاورزی جهت ساخت اصطلاحنامه می‌توانند کمک مؤثری باشند

تجهیزات

نرم‌افزار کتابخانه‌ای نوسا
نرم‌افزار کتابخانه‌ای پارس آذرخش
نرم‌افزار OCR

لازم به ذکر است که در مرکز تعداد زیادی مدرک نمایه‌سازی شده به روش دستی وجود دارد و بانک واژگانی که شامل نمایه موضوعات و اصطلاحات استفاده‌شده در نمایه‌سازی به روش دستی که توسط نرم‌افزارهای کتابخانه‌ای موجود ساخته شده و سیاهه واژگان موجود در نرم‌افزار که باید تکمیل شود در دسترس می‌باشد.

فضا

فضای مناسب جهت استفاده از سخت‌افزار و لوازم مربوط به نرم‌افزار لازم برای نمایه‌سازی ماشینی در بخش سازماندهی در مرکز اطلاع‌رسانی جهاد کشاورزی وجود دارد. همچنین فضای لازم برای کار نمایه‌سازان و نگهداری مدارک و آماده سازی آنها برای نمایه‌سازی ماشینی موجود است.

بررسی کمبودهای موجود در مرکز برای انجام نمایه‌سازی ماشینی

نیروی انسانی

کمبود زبان‌شناس جهت بررسی ساختار زبان فارسی و تطبیق آن با نمایه‌سازی ماشینی.

بودجه

بودجه خاصی برای بخش نمایه‌سازی جهت استفاده از تکنولوژی در نمایه‌سازی در نظر گرفته نشده است.

تجهیزات

- کمبود اصطلاحنامه تدوین‌شده کشاورزی به زبان فارسی چه به صورت چاپی و چه به صورت الکترونیکی
- عدم وجود نرم‌افزار خاص نمایه‌سازی ماشینی

راهکارهای پیشنهادی جهت رفع کمبودها و استفاده از امکانات موجود

نیروی انسانی

می‌توان از نیروی انسانی موجود در مرکز (نمایه‌سازان) با توجه به تجربه کار و تخصص مسئول بخش نمایه‌سازی از لحاظ آکادمیک به عنوان کنترل‌کنندگان نمایه‌سازی ماشینی در تمام مراحل استفاده نمود خصوصاً در ذخیرهٔ اطلاعات- با توجه به نقائص نرم‌افزارهای موجود در ذخیره و بازیابی الکترونیکی اطلاعات به طور تمام اتوماتیک-و چنانچه از روش زبانشناسی استفاده گردد در تهیه اصطلاحنامه فارسی کشاورزی از آنها استفاده کرد. همچنین کنترل انتخاب و

طبقه‌بندی توصیفگرها به ترتیب نیاز چنانچه از روشهای آماری استفاده شود توسط نمایه‌سازان مرکز ضروری به نظر می‌رسد (بدلیل نقص روش در امکان نادیده گرفتن اصطلاحات مهم بخاطر توجه به بسامد حضور واژگان در مدرک).

از متخصصین کامپیوتر مرکز جهت برنامه‌نویسی در کلیه مراحل می‌توان بهره گرفت. همچنین نباید از مشاوره زبانشناسان در مرحله ذخیره اطلاعات و ساخت اصطلاحنامه بی‌نصیب ماند؛ بدلیل نبودن زبانشناس در مرکز، مشاوره زبانشناس در خارج از مرکزضروری است.

فضا

فضای مناسب جهت کار نمایه‌سازان و متخصصین کامپیوتر و زبانشناسان باید در نظر گرفته شود و برای سخت‌افزارهای مورد نیاز فضاسازی مناسب باید صورت گیرد.

بودجه

بودجه لازم برای پیاده‌سازی نمایه‌سازی ماشینی باید از طریق طرحهای تحقیقاتی که البته معمولاِ راهی طولانی را جهت تصویب می‌پیماید تأمین شود.

تجهیزات

تجهیزات سخت‌افزاری و نرم‌افزاری جهت ذخیره سازی، تشخیص محتوا، تفکیک محتوا به پاره‌های اطلاعاتی، مقایسه پاره‌های اطلاعاتی با یکدیگر یا با اصطلاحنامه، طبقه‌بندی پاره‌های اطلاعاتی برحسب اولویتِ نیاز و انتخاب توصیفگرها از میان پاره‌های اطلاعاتی لازم است. همچنین ابزاری جهت ساخت اصطلاحنامه یا تصحیح بانک واژگان و سیاهه بازدارنده در صورت استفاده از روشهای زبانشناسی مورد نیاز است.

به همین منظور می‌توان به ارتقاء نرم‌افزارهای موجود-مثلاِ OCR مورد استفاده در تهیه فهرست مندرجات نشریات و نرم‌افزارهای کتابخانه‌ای موجود- یا طراحی نرم‌افزار جدید اقدام نمود. همچنین از سیاهه بازدارنده و بانک واژگان نرم‌افزارهای کتابخانه‌ای موجود در مرکز جهت تهیه اصطلاحنامه تخصصی کشاورزی الکترونیکی به زبان فارسی استفاده کرد. از آنجا که نرم‌افزارهای مورد استفاده جهت ذخیره‌سازی اطلاعات به صورت الکترونیکی در حال حاضر قابل استفاده به صورت تمام اتوماتیک نیستند و حتماِ باید از نیروی انسانی جهت کنترل آن استفاده کرد چنانچه گفته شد می‌توان با پیروی از قواعد خاص و مدون، نمایه‌سازی ماشینی با زبان فارسی را آغاز نمود و در انتظار سیاستگزاری فرهنگستان زبان فارسی نماند.

بررسی مراحل پیشنهادی انجام نمایه‌سازی ماشینی در مرکز اطلاع‌رسانی جهاد کشاورزی

در این بخش با توجه به نتایج حاصل از بررسی فرایند نمایه‌سازی ماشینی در مبانی نظری و با توجه به بررسی امکانات موجود در مرکز مورد مطالعه و کمبودها پیشنهاداتی در زمینه انجام نمایه‌سازی ماشینی و مراحل اجرای آن مطرح می‌شود.

نتایج حاصل از بررسی مراحل نمایه‌سازی ماشینی

مراحل نمایه‌سازی ماشینی را می‌توان به طور خلاصه چنین تقسیم‌بندی نمود:

- ذخیره محتوا
- تشخیص محتوای مدارک
- تفکیک اطلاعات به پاره‌های اطلاعاتی
- مقایسه پاره‌های اطلاعاتی
- طبقه‌بندی
- انتخاب

در نمایه‌سازی ماشینی، اساس کار، ذخیرهٔ محتوای مدارک به صورت الکترونیکی و تشخیص محتوا- بسته به توانایی سیستم جهت این تشخیص- و سپس تفکیک به پاره‌های اطلاعاتی توسط کامپیوتر می‌باشد.

منظور از ذخیرهٔ محتوا به صورت الکترونیکی، ذخیرهٔ محتوای بخشی از مدرک است که جهت نمایه‌سازی از آن استفاده می‌شود و شامل عنوان یا چکیده یا فهرست مندرجات یا حتی کل مدرک می‌باشد. هنگام ذخیره اطلاعات به صورت الکترونیکی باید براساس ویژگی‌های

خاص هر زبان، از قواعد خاصی پیروی کرد تا تشخیص محتوا به گونه‌ای صحیح انجام گیرد.

در روش‌های آماری و استفاده از احتمالات جز در مرحله ذخیره محتوا، نوع زبان نقش مهمی ندارد.

منظور از تشخیص محتوای مدارک، تشخیص کلمات و جملاتی است که به صورت الکترونیکی ذخیره شده است و بستگی به توانایی سیستم از لحاظ سخت‌افزاری و نرم‌افزاری دارد.

تفکیک محتوا به پاره‌های اطلاعاتی به گونه‌ای انجام می‌پذیرد که کل محتوای یک مدرک که ذخیره شده است به جملات یا کلمات (بسته به آنکه از چه روشی استفاده شود) شکسته شود .

مرحله بعد مقایسه پاره‌های اطلاعاتی است.

مقایسه پاره‌های اطلاعاتی همان مقایسه جملات یا کلماتی است که در مرحله قبل شکسته شده است.

در روش‌های آماری مقایسه پاره‌های اطلاعاتی مدرک با یکدیگر صورت می‌گیرد.

در روش‌های مبتنی بر احتمالات نخست سئوال کاربر تجزیه و تحلیل و به پاره‌های اطلاعاتی تفکیک می‌شود و با پاره‌های اطلاعاتی مدرک مقایسه می‌گردد و در روش‌های بهره‌گیرنده از زبان‌شناسی مقایسه براساس استفاده از ابزاری مانند اصطلاحنامه و یا بانک واژگانی است که در اثر نمایه‌سازی تعدادی مدرک به وجود آمده است، صورت می‌گیرد. بنابراین از دو روش ساخت اصطلاحنامه و نمایه‌سازی دستی به عنوان مبنای کار جهت تولید بانک واژگان استفاده می‌شود و پاره‌های اطلاعاتی با اصطلاحنامه یا بانک واژگان مقایسه می‌شود.

مرحله بعد طبقه‌بندی پاره‌های اطلاعاتی است.

طبقه‌بندی در مورد پاره‌های اطلاعاتی جهت اولویت‌بندی پاره‌های اطلاعاتی بر حسب نیاز و با توجه به برنامه‌ای که قبلاً به سیستم داده شده است انجام می‌گیرد.

در روش‌های آماری طبقه‌بندی براساس بسامد حضور نوعی از پاره‌های اطلاعاتی به نام کلیدواژه در مدرک صورت می‌گیرد و در روش‌های مبتنی بر احتمالات طبقه‌بندی بر مبنای میزان تطابق پاره‌های اطلاعاتی مدرک با پاره‌های اطلاعاتی سئوال کاربر و گاه بر اساس قوانین ریاضی و احتمال حضور پاره‌های اطلاعاتی در مدارک انجام می‌پذیرد و در روش‌های زبان‌شناسی طبقه‌بندی براساس میزان شباهت پاره‌های اطلاعاتی مدرک با واژگان موجود در اصطلاحنامه یا بانک واژگان انجام می‌شود و بدین ترتیب اولویت‌بندی پاره‌های اطلاعاتی صورت می‌گیرد.

انتخاب، آخرین مرحله است و در این مرحله از قواعد ریاضی و اصول برنامه‌نویسی استفاده می‌شود.

در این مرحله از نمایه‌سازی ماشینی پس از اولویت‌بندی پاره‌های اطلاعاتی آندسته از اصطلاحاتی که در اولویت قرار گرفته‌اند به عنوان توصیفگر انتخاب می‌شوند و اصطلاحات نامناسب حذف می‌شوند.

مراحل نمایه‌سازی ماشینی با استفاده از روشهای آماری و احتمالات در مرکز اطلاع‌رسانی جهاد کشاورزی

اولین مرحله تهیه سیاهه بازدارنده سپس محدود سازی متن بوسیله آن می‌باشد و سپس تعیین قواعد تشخیص اصطلاحات متن توسط کامپیوتر می‌باشد. بهترین قاعده، تعیین اصطلاحات است به صورت اسم و در حالت مفرد. سپس ارزشگذاری هر یک از اصطلاحات طبق این فرمول انجام می‌گیرد:

$$\text{ارزش اصطلاح } (t,d) = \frac{\text{تعداد t در d}}{\text{تعداد کل مدارک}}$$

t تعداد اصطلاح در مدرک میباشد

منظور از d مدرکی است که به نمایه‌سازی آن می‌پردازیم.

بعد مقایسه ارزش اصطلاحات در یک جدول و سپس تعیین ارزش نهایی انجام می‌شود که معمولاً ۰/۵ انتخاب می‌شود و اصطلاحاتی که ارزش آنها کمتر از آن باشد حذف می شوند. بعد اصطلاحاتی که دارای بالاترین ارزش می‌باشند انتخاب می‌شوند. هر چه ارزش اصطلاح بالاتر باشد درصد اطمینان برای انتخاب توصیفگر بیشتر است. همچنین بهترین توصیفگرها، یک بسامد بالا در یک

مدرک و بسامد پایین در مجموعه مدرک را دارا می‌باشند (نیاکان، ۱۳۸۱، ص۱۷۴-۱۸۴).

روش دیگر

مرحله اول نمایه‌سازی تعدادی مدارک به صورت دستی است. از آنجا که این کار قبلا در مرکز انجام گرفته است (هرچند با دو روش متفاوت پیش همارا و پس همارا) می‌توان آنرا مبنای کار قرار داد. بعد استفاده از بانک واژگان حاصل از نمایه‌سازی دستی است و سپس محاسبه تعداد حضور اصطلاحاتِ انتخاب شده از مدارکی که نمایه‌سازی دستی در مورد آنها صورت گرفته به منظور مشخص شدن رواج این واژه در بین متخصصین.

در مرحله بعد محاسبه احتمال حضور اصطلاحات در مدارکی است که قصد داریم به صورت ماشینی آنها را نمایه‌سازی کنیم.

$$z(t,s,U)=\frac{h(t,s,U)}{f(t,U)}$$

U: تعداد مجموع اشکال حضور اصطلاح در مدرک است.

f(t,U) تعداد مدارکی است که به صورت دستی نمایه‌سازی شده‌اند و در آنها اصطلاح t به صورت یکی از اشکال U حضور پیدا می‌کند (تعداد مدارک که اصطلاح t در آنها وجود دارد).

h(t,s,U) تعداد مدارکی که در آنها توصیفگر S به صورت دستی انتخاب شده است (تعداد t و s با هم در نظر گرفته می‌شود)

هنگامی که h(t,s,U) و یا z(t,s,U) بسیار کوچک است از لحاظ احتمالات احتمال انتخاب آن اصطلاح بسیار کم است(نور، ۱۹۹۹).

مراحل نمایه‌سازی ماشینی با استفاده از روش زبان‌شناسی در مرکز اطلاع‌رسانی جهاد کشاورزی

نمایه‌سازی شامل دو مرحله می‌باشد:

- مرحله توصیف:

در اولین قدم اصطلاحات موجود در متن مدرک تعیین می‌شوند. کار تحلیل و شناسایی، توسط نرم‌افزارِ طراحی شده برای نمایه‌سازی ماشینی، صورت می‌گیرد و پاراگراف‌هایی که لغات موجود در آنها بار معنایی دارند معرفی می‌شوند. سپس اصطلاحات با نرم‌افزار از لحاظ ریشه‌شناسی تجزیه و تحلیل می‌شوند و با شکل ریشه‌ای که در اصطلاحنامه وجود دارند مقایسه می‌شوند. سپس بررسی می‌شود که دربارهٔ کدام اصطلاحِ موجود در متن، در اصطلاحنامهٔ نمایه‌سازی توصیه‌ای مبنی بر اینکه با توصیفگرهای اصطلاحنامه رابطه‌ای می‌تواند داشته باشد، وجود دارد.

- مرحلهٔ انتخاب:

براساس نتایج حاصل از مرحله قبل مرحلهٔ اصلی نمایه‌سازی آغاز می‌شود، یعنی ارزشگذاری اصطلاحات انجام می‌شود. مراحل مختلف نمایه‌سازی براساس معادله‌ای خاص انجام می‌شود و از یک الگوی کلی $a_0+a_1*x_1+a_2*x_2+...a_n*x_n$ پیروی می‌کند.

a ضرایب مختلف از x است.

متغیر x ارزش عناصر مطابق با نتایج حاصل از مرحلهٔ توصیف می‌باشد.

سپس عناصر مختلف براساس ارزششان اولویت‌بندی می‌شوند.

ارزش x نمادین می‌باشد. از آنجا که a_0 تغییرپذیر نیست، برای x_0 ارزش ۱ به صورت استاندارد در نظر گرفته شده است (البته این عدد برگرفته از نتایج مرحلهٔ توصیف نیست.)

برخی مفاهیم عناصر محاسبه‌شده بدین شرح می‌باشد:

- TermsInGF تعداد اصطلاح در مدرک به صورت اسم مطابق با فرم ریشه‌ای آن در اصطلاحنامه
- enTermsInGF تعداد اصطلاح در مدرک که نوعی اسم خاص می‌باشد و در اصطلاحنامه موجود است
- nrOfUseHints تعداد اصطلاح در مدرک که درzاصطلاحنامه رابطهٔ "بکار ببرید" را دارد
- nrOfZHints تعداد اصطلاح در مدرک که در اصطلاحنامه وجود ندارد و از طریق فرمول محاسبه می‌شود.

(بریل مایر، ۱۹۹۷)

مراحل ساخت اصطلاحنامه الکترونیکی در مرکز مورد مطالعه

در این تحقیق مبنای پیشنهاد اصطلاحنامه جهت استفاده در نمایه‌سازی ماشینی در مرکز اطلاع‌رسانی وزارت جهاد کشاورزی، که مستلزم ساخت اصطلاحنامه‌ای با رعایت اصولی جهت کامپیوتری شدن می‌باشد، اصطلاحنامه شیمی طراحی شده در مرکز اطلاعات و مدارک علمی ایران و استفاده از تجربه ساخت آن می‌باشد.

اخیراً فعالیت‌هایی در مرکز اطلاعات و مدارک علمی ایران در واحد خدمات ماشینی در جهت استفاده از سیستم‌های بازیابی هوشمند اطلاعات صورت گرفته است. در سیستم مورد نظر از اصطلاحنامهٔ نما استفاده شده است. کامپیوتری نمودن این اصطلاحات و تعبیه آن در سایت مرکز اطلاعات و مدارک علمی جهت بازیابی اطلاعات پدیده‌ای است برای بالابردن جامعیت بازیابی. همچنین فعالیت‌هایی در زمینه یک طرح پژوهشی توسط گروه اصطلاح‌شناسی در جهت تألیف اصطلاحنامه‌های علوم مختلف انجام پذیرفته و مراحل نهایی خویش را می‌گذرانند. همچنانکه اصطلاحنامه شیمی که از ساختاری خاص جهت استفاده به صورت کامپیوتری برخوردار می‌باشد به اتمام و به مرحلهٔ بهره‌برداری رسیده است.

بررسی مراحل ساخت این اصطلاحنامه که به زبان فارسی تدوین شده و پیش‌بینی‌هایی که جهت الکترونیکی کردن آن صورت گرفته است می‌تواند ره‌گشای استفاده از اصطلاحنامه‌ای الکترونیکی در نمایه‌سازی ماشینی در زبان فارسی باشد.

همچنین بر اساس تحقیقاتی که در زمینه مقایسه دو اصطلاحنامه CAB و Agrovoc انجام گرفته و انتخاب اصطلاحنامه CAB به عنوان مبنای تدوین اصطلاحنامه در زمینه کشاورزی به زبان فارسی که توسط آقای جلالی دیزجی انجام گرفته پیشنهاد می شود در تهیه اصطلاحنامه کشاورزی به زبان فارسی از CAB استفاده شود. بایستی از نظر وجود یا عدم وجود معادل انگلیسی کلید واژه، داشتن مترادف یا شبه مترادف و به ویژه در استخراج و ارائه صحیح ترین شبکه سلسله مراتب برای واژه مورد نظر ، هر دو اصطلاحنامه مورد مشورت قرار گیرد. در ضمن توصیه می شود برای تهیه اصطلاحنامه فارسی در سایر حوزه های علمی اصطلاحنامه های خارجی موجود شناسایی و برای انتخاب بهترین آنها ، روش مورد استفاده در این تحقیق بکار گرفته شود (جلالی دیزجی، ۱۳۷۰).

با توجه به استفاده از هر دو اصطلاحنامه در بخش نمایه‌سازی مرکز اطلاع‌رسانی جهاد کشاورزی، مراحل ساخت اصطلاحنامه الکترونیکی فارسی در زمینه کشاورزی به شرح ذیل پیشنهاد می‌شود:

باید از بانک واژگان موجود در سیستم استفاده نمود و روابط سلسله مراتبی برای مفاهیم در نظر گرفت و با توجه به تجربه نمایه سازان و متخصصین کشاورزی در مرکز و مشاوره زبان‌شناسان در خارج از مرکز الگوهایی جهت تصمیم گیری در باب انتخاب واژگان در صورت حضور در مدارکی که بعداً به صورت ماشینی نمایه سازی می شوند ساخته شود .

چنانچه ازشکل اصلی کلمات در واژه نامه استفاده شود، باید ارجاعاتی صورت گیرد مثلاً از شکل اصلی به شکل خلاصه ارجاعاتی صورت گیرد و استفاده از سیاهه بازدارنده در سیستم ضروری است. در این لیست می توان علامتهای ناخواسته را گنجانید مانند «ها» و «ات» که در هنگام جمع بستن اسامی در فارسی مورد استفاده قرار می گیرند و «می» برای علامت صرف فعل در حالت مضارع همچنین حروف ربط و حروف اضافه .

قواعدی برای انتخاب یا عدم انتخاب کلید واژه‌هایی که ممکن است در مدرک ظاهر شوند و در بانک واژگان وجود دارند در نظر گرفته شود تا هنگامی که سیستم در هنگام نمایه‌سازی با واژه ای برخورد نمود طبق قاعده تعیین شده در مورد آنها تصمیم می‌گیرد. جهت تصمیم گیری در باب انتخاب اصطلاحات مرجح برای انتخاب هر یک از آنها ارزشی در نظر گرفته می شود و برای تعیین آنها قواعدی ساخته می‌شود.

از طرف دیگر می‌توان بدون استفاده از بانک واژگان موجود در مرکز با طراحی نرم‌افزاری که قابلیت استخراج

شکل ریشه‌ای لغات را دارد با الگو پذیری از نرم‌افزار GERTWOL که نرم‌افزاری تحلیلگر می‌باشد، استفاده شود. این نرم‌افزار برای تحلیل‌های ریشه‌شناسی زبان آلمانی توسط شرکت Lingsoft[۶] طراحی شده است به استخراج ریشه کلمات اقدام کرد.

در واژه نامه ای که از ریشه کلمات در واژه نامه استفاده می شود تبدیل شکل صرف‌شده افعال به بن آنها و جداسازی حروفی که در هنگام صرف به آنها می چسبد انجام می پذیرد . اما از آنجا که ما در فارسی بن ماضی و بن ماضی داریم باید در واژه نامه ها هم بن ماضی داشته باشیم و هم بن مضارع . در مورد اسامی هم باید حتماً شکل مفرد کلمه در واژه نامه وجود داشته باشد. همچنین اجزای کلمات مرکب قابل تشخیص باشند.

جهت تعیین توصیفگرها در اصطلاحنامه با استفاده از فرمول Z همهٔ اصطلاحات موجود در بانک واژگان ارزشگذاری می‌شوند و اصطلاحات نامناسب از بانک واژگان حذف می‌شوند.

$$z(t,s,U)=\frac{h(t,s,U)}{f(t,U)}$$

محاسبه با دو نکته محدود می‌شود:

- از یک طرف با انتخاب اصطلاحاتی که دارای بار معنایی هستند: اصطلاحات با استفاده از فرمول Z ارزشگذاری می‌شوند، شامل اسامی، صفات فاعلی و اسم مصدر (که توسط نرم‌افزار تجزیه و تحلیل شده‌اند).
- از طرف دیگر با نوع کاربرد آنها در نمایه‌سازی چرا که تعداد زیادی اصطلاح وجود دارد که باید توانایی این اصطلاحات در استفاده از اصطلاحنامه کاملاً بررسی شود.

همچنین تعیین یادداشت دامنه برای اصطلاحات و توضیحاتی برای چگونگی استفاده از اصطلاحات اصطلاحنامه و سپس تعیین میزان ارتباط میان اصطلاحات و کلیه مراحل ساخت و مدیریت اصطلاحنامه که در مبانی نظری به آنها اشاره گردید باید مورد توجه قرار گیرد.

در مرحله بعد باید بر اساس ساختار درختی ترسیم شده ، توسط متخصصان برنامه نویسی نرم افزاری طراحی گردد که بتوان ساختار درختی را به عنوان ورودی به آن وارد نمود و ساختارهای متفاوت اصطلاحنامه ای را به صورت خروجی از آن دریافت کرد. ضمن این که هنگام ورود اطلاعات نیز می توان ساختاردرختی را به شکل ترسیمی بروی نمایشگر مشاهده نمود . با وجود این نرم افزار ، دیگر نیازی به تبدیل ساختار درختی به ساختار خطی بر روی کاغذ وجود ندارد ، بلکه می توان ساختار درختی (خورشیدی) را مستقیماً وارد نرم افزار نمود (نیاکان، ۱۳۸۱، ص۱۸۸).

کنترل کیفیت سیستم

جهت آزمایش سیستم می‌توان از دو روش استفاده کرد:

جهت آزمایش سیستم ۲۰ مدرک انتخاب می‌شود و به دو روش نمایه‌سازی دستی و ماشینی نمایه‌سازی می‌شوند. سپس برای هر مدرک که به دو روش فوق نمایه‌سازی گردیده است، از فرمول زیر جهت مقایسه دو روش استفاده می‌شود:

$$K=\frac{S\cap I}{S\cup I}$$

S: تعداد اصطلاحات حاصل از نتیجه نمایه‌سازی ماشینی است.

I: تعداد اصطلاحات حاصل از نتیجه نمایه‌سازی دستی است.

$\cap$: تعداد متوسط S و I می‌باشد.

$\cup$: مجموع S و I می‌باشد.

این مقایسه این امکان را فراهم می‌کند که اصطلاحات انتخاب‌شده به صورت خودکار با ارزش احتمالی این اصطلاحات در هنگام انتخاب به روش دستی مقایسه شود. بدین ترتیب ارزش‌های تعیین‌شده در یک جدول جمع‌آوری می‌شود و برحسب ارزش نهایی ۰/۵ طبقه‌بندی می‌شوند و پنج اصطلاح اول که در اولویت قرار دارند، انتخاب می‌شوند(بریل مایر، ص۱۹۹۷).

- جهت کنترل کیفیت نمایه‌سازی می‌توان از یک آزمایش بازیابی اطلاعات استفاده نمود. بدین صورت

که می‌توان مانند طرح AIR/PHYS (طرح تحقیقاتی که در دانشکده فنی شهر دارمشتات به منظور اجرای نمایه سازی ماشینی اجرا شد) به تناسب هر ۱۵۰۰۰ مدرک ۳۰۰ جستجو انجام داد. بدین ترتیب جامعیت و مانعیت را برای نمایه‌سازی دستی محاسبه و با یکدیگر مقایسه نمود (نیاکان، ۱۳۸۱، ص۱۹۱).

مشکلاتی که ممکن است در حین نمایه سازی ماشینی در مرکز با آن مواجه شویم (متغیرهای دخیل):

- تعدد اصطلاحنامه‌های مورد نیاز و اختلاف در ساختار و انتخاب توصیفگرها در آنها
- سطح غیر تخصصیِ زبان مدرک

مدارک بسیار قدیمی با لغات منسوخ موجود در آنها

- هزینه ساخت اصطلاحنامه و روزآمدسازی آن
- مهمترین مشکل، تطبیق ساختار زبان فارسی با مراحل انجام نمایه‌سازی ماشینی با استفاده از ویژگی‌های خاص زبان فارسی از جمله رسم‌الخط می‌باشد (نور، ۱۹۹۹).

پی‌نوشت‌ها:

1. Gorhar Lusting
2. Fangmeyer
3. British Indexing Standard
4. http://www.softex.de
5. Stopp words
6. www.lingsoft.fi

منابع:

جلالی دیزجی، علی (۱۳۷۰). ***"بررسی تطبیقی دو اصطلاحنامه کشاورزی CAB , Agrovoc"***. پایان‌نامه کارشناسی ارشد کتابداری و اطلاع رسانی،دانشکده علوم تربیتی، دانشگاه تهران.

کینن، استلا(۱۳۸۰). ***فرهنگ فشرده علوم کتابداری و اطلاع‌رسانی***. ترجمه و تدوین دکترفاطمه اسدی‌کرگانی. تهران: نشر کتابدار.

نیاکان، شهرزاد (۱۳۸۱). ***امکان‌سنجی نمایه‌سازی ماشینی در مرکز اطلاع‌رسانی جهاد کشاورزی***. پایان‌نامه کارشناسی ارشد کتابداری و اطلاع‌رسانی، دانشکده علوم تربیتی، دانشگاه تهران.

American Society of Indexers (1994***). "Frequently Asked Questions Indexing"***. Index review in Books, Ireland. [On-Line].
Available: http: //www.asindexing.org/site/indfaq.shtml.

Brilmayer, Iris, und et all (1997***). " Automatische Indexierung von dpa-Meldungen: Kleines Experiment zur Evaluierung des Darmstadter Indexierungsanalyses"***. [one-line].
Available: http://www.iud.fh-darmstadt.de/iud/wwwmeth/publ/ausarb/ ausarb1.htm

Grummann, Martin (2000).***"Sind Verfahren zur maschinellen Indexireung fur literaturbestande offentlicher Bibliotheken geeignet?"***. Bibliothek. 3: 297-318. [on-line].
Available: http: // www. Bibliothek-saur.de/ 2000-3/297-318.pdf

Knorz, Gerhard, und et al (1994***). " Automatische Indexierung ." Wissensreprasentation und Information Retrieval Universitat Posdam***, Informationswissenschaft, Modellversuch BETID, Lehrmaterialien Nr 3. Kapitel 4.,S.138-196] .on-line].
Available: http: // www.iud.fh-darmstadt.de/iud/wwwmeth /publ/skript/autind94/paper1.htm

Nohr, Holger(1999). ***" Maschinelle Indexierung: Skript zur Vorlesung im Fach Inhaltlische Erschliessung"***. Stuttgart: Hochschule fur Bibliotheks- und Informationswesen(HBI).[on-line]. Available: http:www.iuk.hdm-stuttgart.de/nohr/WS/Mi/Ml.html

Weiss, Michaela (2001). ***" Automatische Indexierung mit besonderer Berucksichtigung deutschsprachiger Texte".*** In Seminar aus Informationswissenschaft (3726)-WS 2000-2001 (Institut fur Informationsverarbeitung und Informationswissenschaft). [on-line].
Available: http: // www.wu-wien.ac.at/usr/h88/ h8807610/indexierung.