نمایه‌ساز ماشینی منابع فارسی: مدلی یکپارچه برای پژوهشگاه علوم و فناوری اطلاعات ایران[1]

عمار جلالی‌منش[*] کارشناس ارشد مهندسی صنایع؛ عضو هیئت علمی پژوهشگاه علوم و فناوری اطلاعات ایران؛ تهران

سیروس علیدوستی[2] دکتری مدیریت؛ استادیار پژوهشگاه علوم و فناوری اطلاعات ایران؛ تهران

محمود خسروجردی[3] کارشناس ارشد علم اطلاعات و دانش‌شناسی؛ عضو استعدادهای درخشان، باشگاه پژوهشگران جوان و نخبگان دانشگاه آزاد اسلامی، واحد تهران مرکزی؛تهران

دریافت: ۱۳۹۱/۰۵/۲۵ پذیرش: ۱۳۹۱/۱۰/۱۱

فصلنامهٔ علمی پژوهشی
پژوهشگاه علوم و فناوری اطلاعات ایران
شاپا (چاپی) ۸۲۲۳-۲۲۵۱
شاپا (الکترونیکی) ۸۲۳۱-۲۲۵۱
نمایه در LISA، SCOPUS و ISC
http://jimp.irandoc.ac.ir
دورهٔ ۲۹ | شمارهٔ ۲ | ص‌ص۴۲۵-۴۵۱
زمستان ۱۳۹۲

نوع مقاله: پژوهشی

چکیده: نمایه‌سازی ماشینی به گونه‌ای از نمایه‌سازی گفته می‌شود که در آن با استفاده از الگوریتم رایانه‌ای، واژه‌های کلیدی یک مدرک از عنوان یا متن، استخراج و سپس به شکل مدخل‌های نمایه مرتب و سازمان‌دهی می‌شوند. با وجود نیاز روزافزون به کاربرد رایانه در نمایه‌سازی، تاکنون برای زبان فارسی، نمایه‌ساز ماشینی ساخته نشده است. از این رو در این مقاله نتیجه بررسی و شناخت جنبه‌های علمی و عملیاتی این سیستم نمایه‌سازی در پژوهشگاه علوم و فناوری اطلاعات ایران ارائه می‌شود که بر پایه مأموریت و دانش سازمانی خود یکی از آماده‌ترین بسترها را برای ساخت این گونه نمایه‌ساز دارد. به این منظور، ابتدا به پیوند میان نمایه‌سازی و بازیابی اطلاعات، مباحث نظری و سیر تحول نمایه‌سازی ماشینی، نمایه‌سازی ماشینی از دیدگاه‌های عملیاتی و فرایندی، پروژه‌های موفق نمایه‌سازی به زبان‌های گوناگون در دنیا پرداخته شده و نیازمندی‌های نمایه‌ساز ماشینی برای زبان فارسی توصیف شده است. سپس اجزای یک نمایه‌ساز برای زبان فارسی تشریح و مدل مفهومی نمایه‌ساز ماشینی و همچنین روابط میان اجزا، جزئیات زیرسیستم‌ها، و مدل مفهومی سیستم طراحی شده‌اند. در پایان، نیازمندی‌های ساخت نمایه‌ساز ماشینی برای این پژوهشگاه به بحث گذاشته شده است و وضعیت پروژه‌های انجام‌شده و تجربه‌های موجود در این زمینه نیز تبیین شده‌اند.

کلیدواژه‌ها: نمایه‌سازی؛ نمایه‌سازی ماشینی؛ «یو.ام.ال.»؛ مدل مفهومی؛ پژوهشگاه علوم و فناوری اطلاعات ایران

۱. این مقاله از طرح پژوهشی با عنوان «طراحی مدل مفهومی یکپارچه نمایه‌ساز ماشینی برای منابع فارسی در پژوهشگاه علوم و فناوری اطلاعات ایران» که با حمایت مالی پژوهشگاه علوم و فناوری اطلاعات ایران به انجام رسیده استخراج شده است.

[*] پدیدآور رابط:
jalalimanesh@irandoc.ac.ir
2. alidousti@irandoc.ac.ir
3. mkhosro@gmail.com

۱. مقدمه

با افزایش روزافزون تولید اطلاعات، نمایه‌سازی دستی که کاری زمان‌بر است، پاسخ‌گوی نیازهای سازمان‌ها و خدمات اطلاع‌رسانی نیست. از این رو، نرم‌افزارهایی برای انجام خودکار فرایند نمایه‌سازی طراحی و ساخته شده‌اند که تا حد قابل قبولی سازمان‌دهی اطلاعات را از نمایه‌سازی دستی بی‌نیاز می‌کنند. با وجود پیشینه بلند نمایه‌سازی منابع فارسی در کشور و وجود کارشناسان بسیار در این حوزه، تنها چند مطالعه نظری در زمینه نمایه‌ساز ماشینی برای زبان فارسی انجام شده (نیاکان ۱۳۸۳؛ سنجی ۱۳۸۷؛ بهشتی ۱۳۸۲) و هنوز پروژه یا نرم‌افزار درخوری برای این کار به اجرا درنیامده است. یک نمایه‌ساز ماشینی قوی برای زبان فارسی از بلوک‌های ساختاری اصطلاحنامه‌های جامع دیجیتالی، موتورهای استنتاج نحوی، پایگاه قواعد دستوری، الگوریتم‌های وزن‌دهی گوناگون، و مانند آنها تشکیل شده است که مناسب برای زبان فارسی باشند. بر این پایه، زبان‌شناسی رایانه‌ای بنیاد چنین سیستمی را می‌سازد، ولی در این زمینه کار چندانی انجام نشده است و بلوک‌های ساختاری لازم برای ساخت نمایه‌ساز ماشینی که مانند سیستم‌هایی همچون مترجم ماشینی یا سیستم‌های تشخیص گفتار هستند، هنوز برای زبان فارسی ارائه نشده‌اند.

برای ساخت نمایه‌ساز ماشینی برای زبان فارسی به‌عنوان یک سیستم نرم‌افزاری پیچیده، باید یک مدل مفهومی جامع با روش‌شناسی درخور، طراحی و سیستم بر پایه آن در طول زمان ساخته شود. طراحی چنین سیستمی باید بر اساس نیازمندی‌های زبان فارسی و تجربه‌های موجود در کشور از سویی، و منابع علمی و تجربه‌های موفق کشورهای دیگر انجام شود. برای اینکه مدل پیشنهادی، عملی شود، باید بستری مناسب برای اجرای مدل نیز فراهم باشد. این مقاله گزارش پژوهشی را ارائه می‌کند که برای طراحی مدل مفهومی سیستم نمایه‌ساز ماشینی منابع فارسی انجام شده است. در این پژوهش، پژوهشگاه علوم و فناوری اطلاعات ایران به‌عنوان سازمانی که هم از نظر مأموریت و هم از نظر پیشینه در زمینه نمایه‌سازی منابع فارسی یکی از آماده‌ترین بسترها را برای ساخت این‌گونه نمایه‌ساز دارد، برگزیده و جنبه‌های اجرایی و عملیاتی سیستم بر پایه وضعیت آن طراحی شده است.

بر این پایه، ابتدا پیوند میان نمایه‌سازی و بازیابی اطلاعاتی به همراه بحث‌های نظری و سیر تحول موضوع نمایه‌سازی ماشینی ارائه می‌شود. سپس به نمایه‌سازی ماشینی

به‌صورت عملیاتی و از نگاه فرایندی پرداخته، و به فنون مشترک در این حوزه اشاره می‌شود و گونه‌های نمایه‌سازی ماشینی تشریح می‌شوند. در ادامه، پروژه‌های موفق در دنیا و برای زبان‌های گوناگون معرفی و ویژگی‌های آنها بیان می‌گردد. نیازمندی‌های نمایه‌ساز ماشینی در زبان فارسی، اجزای یک نمایه‌ساز برای زبان فارسی، و مدل مفهومی نمایه‌ساز ماشینی بخش‌های دیگر مقاله هستند. در پایان نیز نیازمندی‌های ساخت نمایه‌ساز ماشینی در پژوهشگاه علوم و فناوری اطلاعات در سایه پروژه‌های انجام‌شده و دانش سازمانی موجود در آن ارائه می‌شوند.

۲. بررسی نوشته‌ها

۲-۱. نمایه‌سازی و بازیابی اطلاعات

سازمان‌دهی اطلاعات فرایندی است که در آن اطلاعات در شکل‌های گوناگون برای استفاده کاربران دستکاری شده و با ابزارهای جست‌وجو قابل بازیابی می‌شوند (Cleveland and Cleveland 2001, 29 - 34). یکی از فعالیت‌های مهم در این فرایند، نمایه‌سازی یا به‌عبارتی انتساب مجموعه‌ای از کدها یا واژه‌ها به مدرک برای نمایش محتوای موضوعی آن است (Ellis, Ford, and Furner 1998). بدین‌ترتیب هدف نمایه‌سازی، راهنمایی کاربر به محتوا و مکان فیزیکی مدارک برای بازیابی است (Cisco 1998). نمایه‌سازی و بازیابی اطلاعات پیوندی دوسویه دارند. هنگامی که پرسش جست‌وجو مطرح می‌شود، سؤال یا مسئله اطلاعاتی را نمایه‌سازی می‌کنیم؛ مدارک نیز هنگام ورود به پایگاه‌های اطلاعات نمایه‌سازی می‌شوند. بدین‌ترتیب می‌توان میان پرسش جست‌وجو و مدارک با استفاده از اصطلاح‌های نمایه‌ای پیوند برقرار کرد (Cleveland and Cleveland 2001, 32-34).

۲-۲. نمایه‌سازی ماشینی

اندیشه استفاده از نمایه‌سازی ماشینی را می‌توان همزاد با علاقه به مدل‌های برداری و احتمالی بازیابی اطلاعات در دهه ۱۹۶۰ دانست (نیاکان ۱۳۸۳). مدل‌های فضای برداری بازیابی اطلاعات زیرمجموعه‌ای از فنون بازیابی اطلاعات هستند و بر این فرض استوارند که می‌توان معنای یک مدرک را از اصطلاح‌هایی به‌دست آورد که آن را تشکیل

می‌دهند. درمدل‌های احتمالی بازیابی اطلاعات از تئوری احتمالات برای مدل‌سازی ریاضی مراحل بازیابی استفاده می‌شود. در این حالت فرض بر آن است که مدارک در هنگام بازیابی بر اساس احتمال ربط ارزیابی می‌شوند (سنجی ۱۳۸۷).

نمایه‌سازی ماشینی به‌گونه‌ای از نمایه‌سازی اطلاق می‌شود که در آن با استفاده از الگوریتم رایانه‌ای، واژه‌های کلیدی یک مدرک از عنوان یا متن، استخراج و سپس در قالب مدخل‌های نمایه مرتب و سازمان‌دهی می‌شوند (ODLIS 2009). از نظر سرعت و یک‌دستی، نمایه‌سازی ماشینی بسیار کارآمدتر از نمایه‌سازی دستی است، و یافتن برخی از اشتباه‌ها برای انسان آسان‌تر است (دایرةالمعارف فارسی کتابداری و اطلاع‌رسانی، نسخه آنلاین، زیرواژه *نمایه‌سازی*).

۲-۳. فنون مشترک در نمایه‌سازی ماشینی

بیشتر روش‌های نمایه‌سازی ماشینی کنونی از اصطلاح‌های نمایه‌ای زبان طبیعی استفاده می‌کنند. اصطلاح نمایه‌ای مستخرج در این روش‌ها، شامل واژه‌های ساده و عبارت‌های چندواژه‌ای است که محتوای مدرک را باز می‌کنند. «لون»[1] نخستین کسی بود که گفت می‌توان واژه‌های معیّنی را به‌طور خودکار از متن استخراج کرد که محتوای متن را بنمایانند (Luhn 1957). به هر روی، همه واژه‌های موجود در مدارک، اصطلاح‌های نمایه‌ای مناسبی نیستند و واژه‌هایی نیز که اصطلاح‌های نمایه‌ای مناسب قلمداد می‌شوند، در تعریف محتوای یک متن نقش هم‌اندازه‌ای ندارند. برخی از فنون به شناسایی و وزن‌دهی اصطلاح‌های درست موجود در مدرک کمک می‌کنند. یک فرایندِ معمولِ گزینشِ اصطلاح‌های نمایه‌ای زبان طبیعی از متون، که همهٔ محتوای آن را بازنمایانند، دارای گام‌های زیر است که البته می‌توانند با ترتیب‌های گوناگونی به‌انجام برسند (Salton 1989, 303):

۱. شناسایی واژه‌های مفرد متن که تحلیل واژگانی[2] نامیده می‌شود؛
۲. جابه‌جایی یا حذف واژه‌های اضافی و اصطلاح‌هایی که در متن بسیار تکرار شده‌اند؛

1. Luhn
2. lexical analysis

۳. کاهش اختیاری واژه‌های باقی‌مانده به شکل ریشه‌ای آنها که فرایند ریشه‌گیری[1] نام دارد؛
۴. شکل‌دهی اختیاری عبارت‌ها به‌عنوان اصطلاح‌های نمایه‌ای؛
۵. جانشینی اختیاری واژه‌ها، ریشه‌های واژه‌ها، یا عبارت‌های مربوط به آنها؛ و
۶. محاسبه وزن هر واژه یا ریشه آن، اصطلاح‌های موجود در رده مورد نظر در اصطلاحنامه، یا اصطلاح عبارتی.

۲-٤. گونه‌های نمایه‌سازی ماشینی

دو گونه نمایه‌سازی ماشینی که در نوشته‌ها بیشتر آمده‌اند، نمایه‌سازی «کوئیک»[2] و «کواک»[3] هستند. در نمایه‌سازی «کوئیک» که همان نمایه درون‌بافتی است، همخوانی همه واژه‌های عنوان مدارک را که با برنامه‌ای به کامپیوتر داده شده‌اند، با سیاهه‌ای از واژه‌های غیرمجاز در برنامه کامپیوتر بررسی می‌شوند. واژه‌هایی که در سیاهه غیرمجاز وجود ندارند، به‌طور خودکار کلیدواژه محسوب می‌شوند. مدخلی که با هر یک از این کلیدواژه‌ها درست می‌شود به‌صورتی است که این واژه را در درون بافت خود، یعنی همراه با بقیه واژه‌های آن عنوان، در شکل و نظم طبیعی‌اش نشان می‌دهد. در برابر، نمایه «کواک» همان نمایه برون‌بافتی است. این نمایه برای رهایی از مسائل و مشکلات ناشی از ضرورت کوتاه‌کردن عنوان و نیز برای ساده‌سازی خوانش مدخل‌ها درست شده است. در این نمایه هرکلیدواژه به‌ترتیب از عنوان خود خارج می‌شود و مقدم بر دیگر اجزای عنوان قرار می‌گیرد. سپس عنوان مدرک به‌ترتیب طبیعی خود و به‌طور کامل در زیر این واژه یا به‌دنبال آن می‌آید. به این ترتیب برای هر واژه مهم یک مدخل ساخته می‌شود (دیانی ۱۳۷۹).

۲-۵. نمایه‌سازی‌با کمک رایانه

در نمایه‌سازی با کمک رایانه، نخست نمایه‌ساز با دقت متن را می‌خواند، شناسه‌های موضوعی را انتخاب و علامت‌گذاری می‌کند، و نمایه دستنویس تهیه می‌شود. آنگاه

1. stemming
2. KWIC: KeyWord In Context
3. KWOC: KeyWord Out Of Context

اپراتور مدخل‌ها را بر پایه نمایه دستنویس از راه صفحه‌کلید وارد رایانه می‌کند. بعد از ورود، رایانه مدخل‌ها را بر پایه نظم خاصی مرتب می‌کند، توصیفگرها را زیر شناسه‌ها می‌برد، جای‌نماهای هر مدخل را با هم ادغام می‌کند، و در پایان یک نمایه الفبایی ارائه می‌کند. این نوع نمایه‌سازی با دو روش گوناگون انجام می‌شود (نوروزی ۱۳۸۰). روش اول، نمایه‌سازی در محیط گرافیکی یا نمایه‌سازی درون‌کاشتی است که در آن نمایه‌ساز یا ویراستار، پس از پایان صفحه‌آرایی و مشخص‌شدن شماره صفحه‌های کتاب، صفحه‌ها را روی صفحه نمایش کامپیوتر مرور می‌کند و کلیدواژه‌ها یا واژه‌ها و عبارت‌های مهم را برمی‌گزینند. هر عبارت برگزیده از نظر سیستم یک کلیدواژه به‌شمار می‌رود. روش دوم، نمایه‌سازی در محیط دستوری[1] یا نمایه‌سازی فرمانی است که در آن، نمایه‌ساز پس از اتمام صفحه‌آرایی و مشخص‌شدن شماره صفحه‌های کتاب، صفحه‌ها را در صفحه نمایش کامپیوتر نگاه می‌کند و کلیدواژه‌ها را نشانه‌گذاری و در همان فایل، ارجاعات را نیز تعیین می‌کند. سپس یک ابزار نرم‌افزاری همه نمایه را تهیه و الفبایی می‌کند.

۲-۶. پروژه‌های موفق نمایه‌ساز و نمایه‌سازی ماشینی

تاکنون پروژه‌های بسیاری به ساخت مدل‌های نمایه‌ساز و نمایه‌سازی ماشینی پرداخته‌اند. این مدل‌ها دارای ابعاد موضوعی، جامعیت، و مانعیت گوناگونی هستند. برخی از این مدل‌ها نیز ترکیبی از مدل‌های دیگر هستند و سازندگان آنها با توجه به نیازهای زبان‌شناختی و فرهنگی کشور خویش مدل ویژه‌ای را یا ابداع یا بومی‌کرده‌اند.

در همین زمینه «ال‌بلتاگی» و «رافیا» سیستمی با نام «کی‌پی ـ ماینر»[2] معرفی کردند که برای استخراج اصطلاح‌های کلیدی از متون عربی و انگلیسی مناسب است. در برابر سیستم‌های استخراج واژه کنونی، «کی‌پی ماینر» نیازی به یادگیری درباره یک مجموعه مدرک ندارد تا بتواند کار خود را انجام دهد (El-Beltagy and Rafea 2009).

«منصور» و همکاران روشی را برای نمایه‌سازی خودکار متن‌های عربی ارائه کردند. روش آنها بر تحلیل ریخت‌شناختی و تخصیص وزن به واژه‌ها متکی است. در تحلیل ریخت‌شناختی، از شماری از قاعده‌های دستوری برای استخراج ریشه واژه‌هایی که نامزد

1. Command Line Interface
2. KP-Miner

نمایه می‌شوند، استفاده می‌شود. وزن‌دهی بر پایه چگونگی پخش یک واژه در یک مدرک و نه‌تنها بر مبنای فراوانی آن در مدرک انجام می‌شود. سپس واژه‌های نامزدِ نمایه به‌ترتیب نزولی وزنشان مرتب می‌شوند، به‌گونه‌ای که بازیابندگان اطلاعات بتوانند مهم‌ترین واژه‌های نمایه‌ای را برگزینند. این روش در آزمون‌ها، ۴۶ درصد جامعیت و ۶۴ درصد مانعیت را نشان می‌دهد (Mansour et al. 2008).

«ماتسومورا»، «اوساوا»، و «ایشیزوکا» سیستم «پای» را معرفی کردند که یک سیستم نمایه‌سازی ماشینی است. این سیستم تأکید بسیاری بر کلیدواژه‌هایی دارد که نویسنده مدرک انتخاب کرده است. به باور آنان، از آنجایی که نویسنده‌ای یک مدرک را می‌نویسد نکات مورد توجه خویش را به‌خوبی بیان می‌کند؛ اصطلاح‌های برانگیزاننده‌ای[1] که در ذهن خواننده تولید می‌شوند، می‌توانند کلیدواژگان درست را نشان دهند (Matsumura, Ohsawa, and Ishizuka 2002).

«لی» و «آن» به ارائه روشی برای نمایه‌سازی خودکار متون کره‌ای پرداختند. در زبان کره‌ای، صدها پسوند و پیشوند وجود دارند و افزون بر این، وندها می‌توانند به روش‌های گوناگونی ترکیب شوند و واژه‌ها و اصطلاح‌های مرکب را پدید آورند. این ویژگی زبان کره‌ای، ریشه‌گیری واژه‌ها را بسیار دشوار کرده است. بنابراین، آنان از روش نمایه‌سازی واژک‌مدار[2] برای حل این دشواری استفاده کردند (Lee and Ahn 1996).

«فوجی» و «کرافت» نظامی برای نمایه‌سازی متون ژاپنی ابداع کردند. آنها با استفاده از نظام «اینکوئری»[3]، نظامی برای نمایه‌سازی زبان ژاپنی با نام «جی‌اینکوئری»[4] پدید آوردند. این نظام که وابسته به زبان نیست دارای دو ماژول است. یکی از این ماژول‌ها، ماژول نمایه‌سازی برای ایجاد پایگاه داده‌ای از منابع متنی است. ماژول دوم نیز برای فرموله‌کردن یک پرسش ساختاری با استفاده از جنبه‌های دستوری ژاپنی به‌کار می‌رود (Fuji and Croft. 1993).

«وودروف» و «پلانت» سیستمی با نام «جیپسی»[5] برای نمایه‌سازی خودکار ابداع

---

1. impressive
2. Morpheme-based
3. INQUERY
4. JINQUERY
5. GIPSY

کردند. استراتژی بنیادین «جیپسی» استخراج نام‌های مکان‌ها و شاخص‌های جغرافیایی از مدارک و استفاده از این پیوندهای درونی برای شناسایی ناحیه‌ای است که مدرک به آن اشاره می‌کند (Woodruff and Plaunt 1994).

«اوساوا»، «بنسون»، و «یاکیدا» الگوریتمی برای استخراج ایده اصلی از چکیده متون با نام «کی‌گراف»[1] راه‌اندازی کردند. الگوریتم پیشنهادی، بر پایه بخش‌بخش‌سازی و تفکیک «گراف»، به بازنمایی میزان هم‌رویدادی بین اصطلاح‌های موجود در یک مدرک، به‌صورت چندخوشه می‌پردازد. هر خوشه به یک مفهوم وابسته است که نویسنده در نظر گرفته است؛ سپس با یک تحلیل آماری، اصطلاح‌هایی که بالاترین جایگاه را با توجه به اندیشه نویسنده داشته‌اند، به‌عنوان کلیدواژه گزینش می‌کند (Ohsawa, Benson, and Yachida 1998).

۲-۷. نرم‌افزارهای نمایه‌سازی ماشینی

«سیندکس»[2] یکی از برنامه‌های تهیه نمایه‌هایی خودکار برای کتاب‌ها، روزنامه‌ها، و دیگر نشریه‌های ادواری است. این برنامه ویژگی‌های نمایه‌ساز حرفه‌ای را تقلید می‌کند و کاربرد آن بسیار آسان است. از این برنامه می‌توان در ساخت اصطلاحنامه‌ها یا مستندات موضوعی نیز بهره برد که بنیانی برای نمایه‌سازی‌های خودکار کنونی تلقی می‌شوند (Indexer Research 2010).

«ماک‌رکس»[3] برنامه‌ای برای نمایه‌سازی ماشینی با پیشینه‌ای نزدیک به ۳۰ سال است (Calvert and Calvert 2010). بسیاری از نمایه‌سازان بنام در جهان از این سیستم استفاده می‌کنند. این برنامه برای هر سطح و میزان از نمایه‌سازی تعریف شده است و در حجم‌های بالا بسیار خوب کار می‌کند. از ویژگی‌های مهم این سیستم می‌توان به ساخت شناسه‌ها، فهرست مستندات، بررسی دقیق مراحل، صرفه‌جویی در زمان، استفاده از حروف ویژه، بررسی دوباره نمایه‌ها، و اصلاح آنها اشاره کرد. با این برنامه می‌توان نمایه‌های دیگر را وارد[4] سیستم یا از آن دریافت کرد.

---

1. KeyGraph
2. Windex
3. Macrex
4. import

«اِسکای ایندکس»[1] نیز یکی دیگر از این نرم‌افزارهاست که در همایش سالانه انجمن نمایه‌سازی آمریکا[2] در «پورتلند» رونمایی شد (Sky Software 2010). از جمله ویژگی‌های این سیستم می‌توان به تعداد سطح‌ها برای سرعنوان‌های موضوعی، برچسب‌های فصل و جلد، و امکان تعریف پیشوند و پسوند برای همه برچسب‌ها اشاره کرد.

«دکستر»[3] نیز برای نمایه‌سازی مستقیم در واژه‌پردازهایی مانند «مایکروسافت ورد» طراحی شده است (Editurium 2010) و محدودیت‌های برنامه‌های نمایه‌سازی دیگر و نیز برنامه‌های واژه‌پرداز را ندارد. با این برنامه، نمایه‌ساز نیاز به ورود، ویرایش، یا توجه به شناسه‌های افزوده ندارد. نشانه‌هایی که برای شناسایی گستره وسیعی از متن‌ها نیاز است به‌طور خودکار در متنِ برگزیده، افزوده می‌شوند. این سیستم شناسه‌های نمایه را همواره به‌صورت قابل ویرایش و به‌صورت جدول ترتیبی در اختیار نمایه‌ساز قرار می‌دهد.

۲-۸. زبان فارسی و نمایه‌سازی ماشینی

«نیاکان» ویژگی‌های زبان فارسی را از دیدگاه نمایه‌سازی ماشینی در دو بخش ویژگی‌هایی که به نمایه‌سازی ماشینی کمک می‌کنند و آنچه مانع آن می‌شوند، بخش‌بندی می‌کند (نیاکان ۱۳۸۳). از میان ویژگی‌های زبان فارسی که به سود نمایه‌سازی هستند، می‌توان به نبودن حرف تعریف، عامل جنسیت، و جمع شکسته اشاره کرد. از سویی نیز وجود کسره اضافه در زبان فارسی از مزایای آن به‌شمار می‌رود که شکل اضافه را بسیار ساده کرده است. صفت و موصوف و عدد و معدود هم با هم مطابقت ندارند و یک صفت می‌تواند به چندین موصوف برگردد.

از نظر وی، دشواری‌های زبان فارسی برای نمایه‌سازی ماشینی بیشتر معطوف به دستور این زبان هستند که در سه‌بخش قرار می‌گیرند. اول اینکه میان گفتار و نوشتار تفاوت‌های زیادی وجود دارد؛ یعنی مردم در گفتن واژه‌ها الگوهایی را به‌کار می‌برند که در نوشتار نمی‌آورند. دوم، اضافه‌ها که عبارت‌اند از نسبت میان دو واژه یا گروهی از واژه‌ها با هم که گونه‌ای از پیوند لفظی یا معنوی را معلوم می‌کنند و دارای سه گونه ملکی، تخصیصی، و بیانی هستند. سوم، چگونگی نگارش است که از جمله می‌توان به

1. Sky Index Professional
2. Amercian Indexing Society
3. DEXter

جدا یا پیوسته‌نویسی مانند «یکطرفه» و «یک‌طرفه»، واژه‌های ترکیبی مانند «منابع آب» و «مکانیک خاک»، شیوه دوگونه املای برخی از واژه‌ها مانند «جستجو» و «جست‌وجو»، جایگاه الفبایی واژه‌ها، و ناتوانی در نشان‌دادن صداها اشاره کرد.

«داوودآبادی» نیز به شماری از پیچیدگی‌های درک متن فارسی اشاره می‌کند که می‌توانند بر نمایه‌سازی ماشینی تأثیر بگذارند. به گفته وی، زبان فارسی از ساختار ریخت‌شناسی پیچیده‌ای برخوردار است و یک شکل ظاهری می‌تواند واحدهای معنی گوناگونی را نشان دهد. افزون بر این، بی‌ترتیب‌بودن زبان، دشواری در تعیین حدود عبارت‌ها به‌ویژه حدود گروه‌های اسمی، چندمعنایی و چندنقشی‌بودن واژه‌هایی مانند «شیر»، و حذف واژه‌ها یا عبارت‌ها به قرینه لفظی یا معنوی از جمله این پیچیدگی‌ها به‌شمار می‌روند (داوودآبادی ۱۳۸۴).

۳. روش

نمایه‌ساز ماشینی نرم‌افزاری است که طراحی آن نیازمند استفاده از روش‌های نوین طراحی نرم‌افزار است. این نرم‌افزار دارای ویژگی‌هایی نیز هست که بر گزینش روش طراحی آن تأثیر می‌گذارد. از جمله این ویژگی‌ها می‌توان به ساخت و توسعه گام‌به‌گام این سیستم اشاره کرد. افزون بر این، یک نمایه‌ساز ماشینی ممکن است از زیرسیستم‌های گوناگونی تشکیل شده باشد که برخی از آنها برای سیستم بایسته و دیگر زیرسیستم‌ها نقش توسعه‌ای دارند. به‌عنوان نمونه در «کی‌پی ـ ماینر» از ریشه‌گیری کلمات برای استخراج واژه‌ها استفاده شده بود در حالی که در «کی‌گراف» تلاش در استخراج روابط معنایی بین مفاهیم و واژه‌ها بود. بر این پایه می‌توان گفت که ممکن است یک نمایه‌ساز ماشینی در یک مرحله فقط به روابط نحوی بپردازد و در حالت بالغ‌تر سعی در استخراج روابط معنایی کند. به گفته دیگر، یک نمایه‌ساز ماشینی ممکن است از سطوح بلوغ گوناگونی برخوردار باشد و در طول زمان بالغ شود. این ویژگی باعث می‌شود که فرایند تحلیل به‌صورت یک مرحله از پروژه تعریف نشود، بلکه فرایندی پیوسته باشد که در همه مراحل بلوغ سیستم همراه مراحل دیگر است.

از این‌رو، برای طراحی نمایه‌ساز ماشینی، روش‌شناسی «آر.یو.پی.»[1] برگزیده شد که

1. Rational Unified Process: RUP

یک فرایند پیوسته است؛ یک معماری نرم‌افزاری نیز هست که به اجزای اصلی سیستم به‌خوبی می‌پردازد؛ مدیریت پروژه‌های نرم‌افزاری را تسهیل می‌کند؛ از یک زبان مدل‌سازی استاندارد به نام «یو.ام.ال.»[1] برای تجزیه و تحلیل، طراحی، و پیاده‌سازی استفاده می‌کند؛ و همچنین از معماری مدل‌رانه[2] پشتیبانی می‌کند.

در طراحی نمایه‌ساز ماشینی از دیاگرام‌های «یو.ام.ال.» بیشتر برای مدل‌کردن ماهیت رفتاری اجزای سیستم و ساخت مدل مفهومی و طراحی اسکلت سیستم استفاده و کمتر به جنبه‌های فیزیکی پرداخته شد. در معماری «آر.یو.پی.»، این مدل‌ها و دیاگرام‌ها بنیاد ساخت سیستم قرار خواهند گرفت و تحلیل‌گر و سازنده سیستم آن را مانند نقشه کار خود قرار خواهد داد.

بر این پایه، مراحل زیر برای طراحی مدل مفهومی نمایه‌ساز ماشینی منابع فارسی برای پژوهشگاه علوم و فناوری اطلاعات ایران دنبال شدند:

۱. نیازسنجی نمایه‌ساز ماشینی برای زبان فارسی؛
۲. طراحی زیرسیستم‌ها؛
۳. شناخت تأثیر ویژگی‌های زبان فارسی برای نمایه‌ساز ماشینی؛
۴. طراحی دیدگاه مورد کاربرد[3]؛
۵. طراحی دیدگاه منطقی[4]؛
۶. شناخت نیازمندی‌های ساخت نمایه‌ساز ماشینی در پژوهشگاه علوم و فناوری اطلاعات ایران؛ و
۷. ارزیابی وضع کنونی پژوهشگاه علوم و فناوری اطلاعات ایران برای ساخت نمایه‌ساز ماشینی.

٤. مدل مفهومی نمایه‌سازی ماشینی برای منابع فارسی

٤-١. نیازسنجی

همان‌گونه که پیش‌تر اشاره شد، فرایند متداول انتخاب اصطلاح‌های نمایه‌ای زبان

1. Unified Modeling Language: UML
2. model driven architecture
3. use case view
4. logical view

طبیعی از متن که همه محتوای آن را بازنماید، دارای گام‌های کم و بیش روشنی است (Salton 1989, 303). بیشتر پروژه‌های موفق نمایه‌سازی ماشینی، همه یا برخی از این گام‌ها را دنبال می‌کنند. بنابراین فرایند کلی یا گام‌های اصلی نمایه‌سازی ماشینی برای همه زبان‌ها یکی است و تنها باید زیرسیستم‌ها را برای یک زبان، سفارشی‌سازی و کارآیی سیستم را پس از تغییر، بررسی و در صورت نیاز آن را تنظیم[1] کرد. بنابراین، آیا نیازی به ساخت یک سیستم نمایه‌ساز ماشینی برای زبان فارسی وجود دارد؟ این پرسش را می‌توان از سه دیدگاه پاسخ داد:

الف) امکان سفارشی‌سازی سیستم‌های موجود. برای اینکه بتوان سیستم‌های نمایه‌سازی کنونی را برای زبان فارسی سفارشی‌سازی کرد، این سیستم‌ها باید منبع‌ـباز[2] باشند و قابلیت تغییر در آنها نیز وجود داشته باشد. برای تغییر در سیستم‌های تجاری نیز باید با تولیدکننده توافق و هزینه‌های آن را نیز پرداخت کرد.

ب) انعطاف سیستم‌های موجود برای همخوانی با زبان فارسی. همانگونه که اشاره شد نمایه‌ساز ماشینی سیستمی است که باید به مرور زمان بالغ شود و این بلوغ وابسته به تنظیم و تغییر همه زیرسیستم‌هاست. از این‌رو قابلیت سفارشی‌سازی زبان برای یک سیستم به‌تنهایی کافی نیست و باید انعطاف لازم برای همخوان‌سازی با زبان فارسی را در طول زمان نیز داشته باشد.

پ) قابلیت توسعه‌های آتی. افزون بر این، ممکن است برای ارتقاء سیستم نیاز به افزودن یک یا چند زیرسیستم تازه یا پیاده‌سازی الگوریتم‌هایی مانند تکنیک‌های هوش مصنوعی باشد که تا به‌حال استفاده نشده‌اند.

بر این پایه، ساخت یک نمایه‌ساز ماشینی بر بنیاد سیستم‌های موجود، راه‌حلی کوتاه‌مدت است و دستیابی به بسیاری از قابلیت‌های یک نمایه‌ساز ماشینی بالغ را ناممکن خواهد کرد. از این رو، ساخت یک سیستم نمایه‌ساز ماشینی ویژه زبان فارسی با معماری دقیق و زیرسیستم‌های قابل تغییر و توسعه، رویکرد بهتری خواهد بود.

1. tuning
2. Open-source

٤-٢. طراحی زیرسیستم‌ها

زیرسیستم‌های نمایه‌ساز ماشینی زبان فارسی بر پایه اجزایی طراحی شدند که از سوی «سالتون» ارائه شده‌اند (Salton 1989, 303). این زیرسیستم‌ها عبارت‌اند از:

◇ زیرسیستم مبدل دیجیتال

این زیرسیستم ورودی نمایه‌ساز ماشینی است. در صورتی که اسناد از ابتدا دیجیتالی باشند به‌طور مستقیم وارد مراحل بعدی می‌شوند وگرنه ابتدا دیجیتالی و سپس برای تجزیه کلمات و مراحل بعدی وارد نمایه‌ساز می‌شوند.

◇ زیرسیستم تحلیل واژگانی

این زیرسیستم، متن را به واژه‌ها تفکیک می‌کند و ماهیت هر کلمه را تشخیص می‌دهد و تشخیص نوع واژه و شناسایی فعل‌ها، الفاظ، و اصطلاح‌ها را در بر دارد. واژه‌شکن در ابتدا زنجیره کلمات را جدا و نوع آنها را تشخیص می‌دهد. در حالت‌های ویژه مانند اختصارها یا نام‌های خاص (مانند نام سازمان‌ها)، واژه با جداول مربوطه تطبیق داده می‌شود. از آنجایی که یکی از منابع خوب برای گزینش واژه‌های کلیدی پیشنهادهای نویسنده است، زیرسیستم تحلیل واژگانی، واژه‌های پیشنهادی نویسنده را نیز به‌عنوان نامزد، به سیاهه واژه‌های کلیدی می‌افزاید.

◇ زیرسیستم گزینش عبارت‌ها و اصطلاح‌ها و نرمال‌سازی

این زیرسیستم رابط میان نمایه‌ساز ماشینی و اصطلاحنامه است و از یک‌سو، اصطلاح‌ها و واژه‌های استخراج‌شده از متن را با واژه‌های استاندارد مطابقت می‌دهد و واژه نرمال را جایگزین می‌کند. از سوی دیگر نیز اصطلاح‌هایی را که در اصطلاحنامه نیستند، شناسایی و برای افزودن پیشنهاد می‌کند.

◇ زیرسیستم فهرست حذفی

این زیرسیستم، واژه‌های پرکاربردی را که نمی‌توانند به‌عنوان واژه کلیدی برگزیده شوند (مانند حرف اضافه)، به‌عنوان فهرست حذفی نگهداری و آنها را از واژه‌های نامزد برای کلیدواژه‌ها حذف می‌کند. افزون بر این، بر پایه نظر «بالرینی» و دیگران واژه‌هایی که عدد آستانه مربوط به حرف را دارند نیز حذف می‌شوند (Ballerini et al. 1997). برای ازدست‌ندادن واژه‌های کوتاه که ارزش نمایه‌شدن را دارند از لیست واژه‌های ضدحذفی استفاده می‌شود. به این ترتیب حجم لیست نامزد نمایه کاهش بسیاری خواهد یافت.

◇ زیرسیستم ریشه‌گیری واژه‌ها

برای افزایش کیفیت نمایه‌سازی، در این زیرسیستم واژه‌ها ریشه‌گیری و به دو گونه فعل و لفظ یا واژه تقسیم می‌شوند. با ریشه‌گیری، همخوانی میان اصطلاح‌های نمایه‌ای پرسش و متن مدرک بهبود می‌یابد. به باور «سالتون» (Salton 1986) ریشه‌گیری جامعیت را افزایش و دامنه اصطلاح‌های کلیدی را در متن گسترش خواهد داد. ابزار توسعه‌ای به‌نام تجزیه و تحلیل معناشناختی نیز در این زیرسیستم فراهم شده است تا واژه‌های دارای چند معنی یا جمله‌های بدون بارمعنایی شناخته شوند. بررسی معناشناختی می‌تواند متن‌های دارای معنای مشترک را گروه‌بندی کند تا با استفاده از این ارتباط بتوان آنها را بازیابی کرد.

◇ زیرسیستم وزن‌دهی به اصطلاح‌ها

پس از گزینش اصطلاح‌ها و واژه‌های نامزد نمایه نمی‌توان همه آنها را به‌عنوان نمایه برگزید. از این‌رو سازوکاری باید که نمایه‌ها را اولویت‌بندی کند و شماری از آنها را به‌عنوان کلیدواژه‌های پایانی برگزیند. زیرسیستم وزن‌دهی به اصطلاح‌ها به هر اصطلاح یک وزن عددی تخصیص می‌دهد که وابسته به فراوانی اصطلاح در مدرک و همچنین فراوانی آن در مدارک دیگر است. وزن‌دهی به‌صورت رفت و برگشتی انجام و پس از استفاده از یک الگوریتم وزن‌دهی، کیفیت آن بر پایه مدارک بازیابی‌شده با تغییر الگوریتم یا مقدار فاکتورهای آن تنظیم می‌شود.

٤-٣. تأثیر زبان فارسی بر اجزای نمایه‌ساز فارسی

فرایندهای نمایه‌سازی ماشینی و بسیاری از زیرسیستم‌ها و الگوریتم‌های آن از یک الگوی کم و بیش ثابت پیروی می‌کنند و تأثیر ویژگی‌های زبانی، بیشتر بر جزئیات یا داده‌های ورودی به این الگوریتم‌ها، توالی فرایندها، یا طراحی قواعد است. چون در اینجا مدل مفهومی نمایه‌ساز ماشینی ارائه می‌شود، تنها دو تأثیر زبان فارسی یادآوری می‌شود:

◇ مسائل ریخت‌شناسی

زبان فارسی ساختار ریخت‌شناسی کم و بیش پیچیده‌ای دارد. شمار بسیارپسوندها و شمار کم‌پیشوندها، سیستم پیچیده ترکیب‌ها، و تفاوت در واحدهای معنایی از جمله ویژگی‌های این ساختار هستند. مهم‌ترین تأثیر این ساختار، بر زیرسیستم تحلیل واژگانی و زیرسیستم ریشه‌گیر واژه‌هاست.

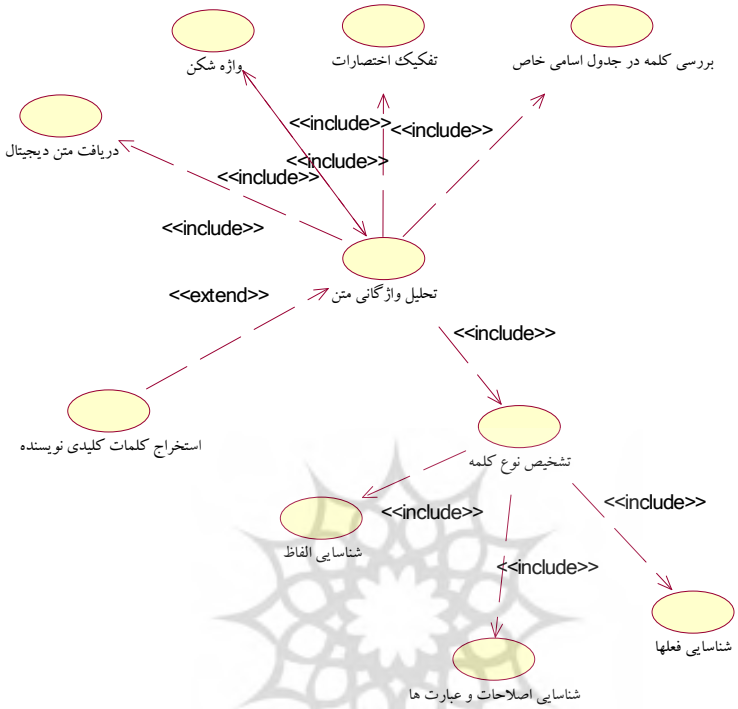
◇ مسائل نحوی

ویژگی‌هایی مانند بی‌ترتیب‌بودن، چندمعنایی‌بودن، و حذف به قرینه لفظی و معنوی از جمله ویژگی‌های نحوی زبان فارسی هستند که بر سیستم نمایه‌ساز ماشینی تأثیر می‌گذارند. تأثیر اصلی این تفاوت‌ها هم بر زیرسیستم‌های تحلیل‌واژگانی و ریشه‌گیر و هم در برخی موارد بر زیرسیستم گزینش عبارت‌ها و اصطلاح‌ها و نرمال‌سازی خواهد بود.

٤-٤. طراحی مدل مفهومی

همان‌گونه که اشاره شد در این پژوهش از روش‌شناسی «آر.یو.پی» برای طراحی نمایه‌ساز ماشینی، و از زبان مدل‌سازی «یو.ام.ال» برای ترسیم مدل‌ها استفاده شده است. نکته‌ای که در مورد دیاگرام‌ها و مدل‌های استفاده‌شده قابل اشاره است، ماهیت رفتاری آنهاست. از آنجایی که در اینجا ساخت مدل مفهومی و طراحی اسکلت کلی سیستم نمایه‌ساز ماشینی مدّنظر بود، از مجموعه دیاگرام‌هایی استفاده شد که رفتار سیستم را مدل می‌کردند و در موارد اندکی به جزئیات ساختار سیستم و درواقع مدل‌های استاتیک و نزدیک به ساختار واقعی و نهایی نرم‌افزار پرداخته شد. ولی همان‌گونه که گفته شد در معماری استفاده‌شده، این مدل‌ها مبنای اصلی توسعه سیستم قرار خواهند گرفت و تحلیل‌گر و سازنده سیستم آن را مانند نقشه‌ای مبنای کار خود قرار خواهند داد.

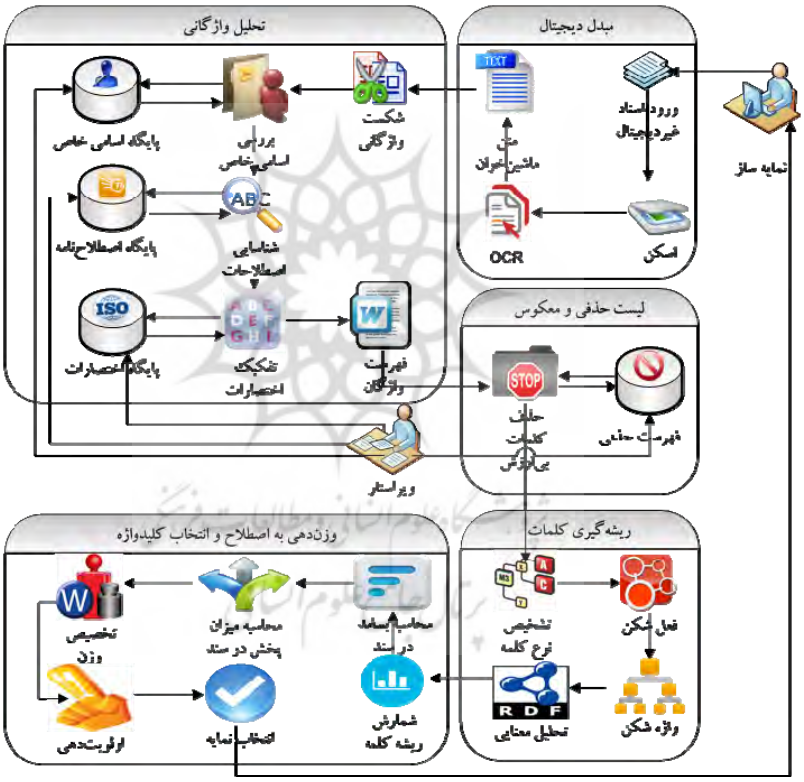
اولین و بنیادی‌ترین دیاگرام که طراحی سیستم با آن شروع می‌شود، نمودار مورد کاربرد است. دیدگاه مورد کاربرد مفهومی است که به درک تحلیل‌گر از چگونگی رفتار سیستم کمک می‌کند. در این نما ارتباط میان کاربران و سیستم روشن و تا حد نیاز زیرسیستم‌ها نیز تحلیل می‌شوند. در این پروژه برای هر یک از زیرسیستم‌هایی که پیش‌تر به آنها اشاره شد یک نمودار مورد کاربرد مستقل و در مجموع پنج نمودار ترسیم شد. شکل ۱، یکی از این نمودارها را که برای زیرسیستم تحلیل واژگانی ترسیم شده است، برای نمونه نشان می‌دهد.



شکل ۱. نمودار مورد کاربرد زیرسیستم تحلیل واژگانی

شکل ۲، شمای کلی نمایه‌ساز ماشینی و فرایند استخراج ماشینی واژه‌های کلیدی را نشان می‌دهد. بر این پایه فرایند نمایه‌سازی با تبدیل متون غیردیجیتال به دیجیتال آغاز و در گام بعدی وارد زیرسیستم تحلیل واژگانی می‌شود. در این گام متون پیوسته به واژگان شکسته‌شده و وجود نام‌های خاص، اصطلاحات، و اختصارات در آن بررسی می‌شوند. سپس بر پایه فهرست حذفی، واژه‌هایی که ارزش نمایه‌شدن را ندارند، حذف و واژگان دیگر برای ریشه‌گیری وارد زیرسیستم ریشه‌گیری واژه‌ها می‌شوند. در این گام ابتدا نوع واژه مشخص و ریشه‌گیری بر پایه نوع انجام می‌شود. اگر واژه از نوع فعل باشد، زمان، نوع، بن، و ویژگی‌های دیگر آن و اگر فعل نباشد، ریشه کلمه تشخیص داده می‌شود. در این گام همه واژه‌ها در حالت مفرد در نظرگرفته و اگر نگاه به واژه‌ها نگاه معنایی باشد، از

دیدگاه معنایی نیز دسته‌بندی می‌شوند. برای نمونه در این حالت ممکن است واژه «جغرافیا» و «زمین» در یک گروه معنایی قرار گیرند. آخرین گام از فرایند نمایه‌سازی، محاسبه شاخص‌هایی همچون پخش و بسامد و تخصیص وزن است. در پایان بر پایه وزن‌هایی که به واژه‌ها و اصطلاحات تخصیص داده شده‌اند، اولویت‌بندی انجام می‌شود. اینکه چه تعداد واژه به‌عنوان نمایه برگزیده شوند، می‌تواند جزئی از ترجیحات نمایه‌ساز باشد. چراکه با انتخاب تعداد نمای بیشتر، احتمال بازیابی[1] بیشتر می‌شود و دقت[2] پایین می‌آید.

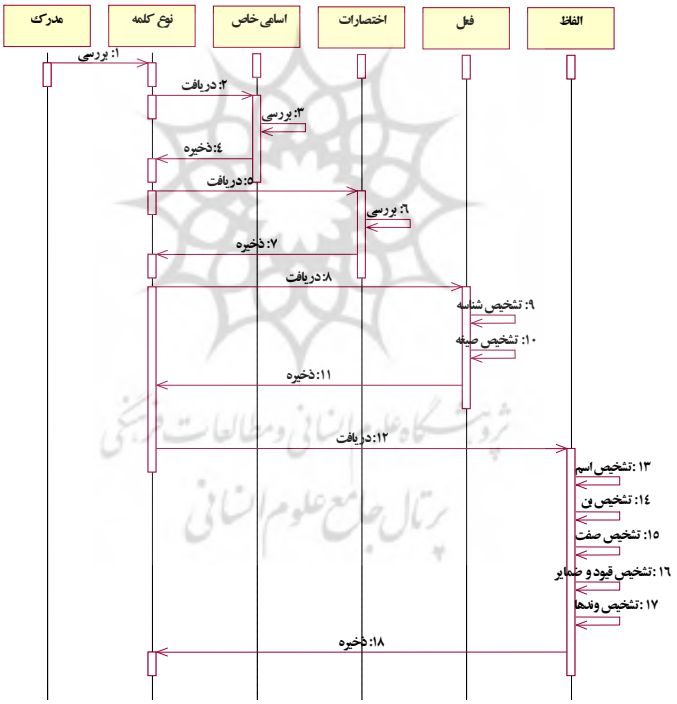


شکل ۲. شمای کلی نمایه‌ساز ماشینی (نمای فرایندی)

1. recall
2. precision

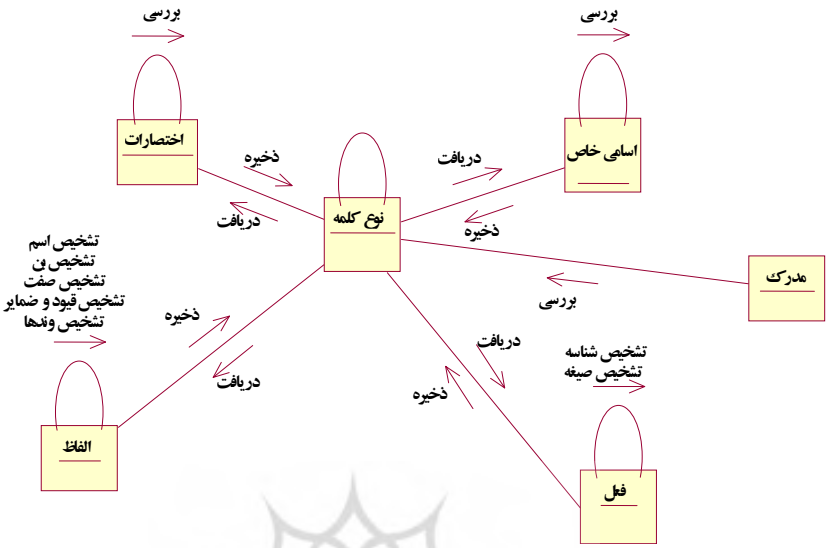
٤-٥. طراحی دیدگاه منطقی

در این دیدگاه بر پایه دیاگرام‌ها و تحلیل‌های انجام‌شده در دیدگاه شماتیک مورد کاربرد، اجزای اصلی سیستم شناسایی و در قالب «دیاگرام‌های کلاس»[1] نمایش داده شدند. در هر مرحله نیز دیاگرام توالی[2] ترسیم شد که چگونگی تعامل میان اجزاء را در طول زمان نشان می‌دهند. دیاگرام همکاری، اجزایی مشابه دیاگرام توالی دارد، ولی زمان انجام هر فعالیت را نشان نمی‌دهد. این دیاگرام کلیّت سیستم را بهتر از دیاگرام توالی نشان می‌دهد. شکل ۳، توالی عملیات را برای زیرسیستم تحلیل واژگانی نشان می‌دهد که پیش‌تر دیدگاه مورد کاربردِ آن طراحی شده بود. شکل ۴، دیاگرام همکاری را برای همین زیرسیستم نشان می‌دهد، با این تفاوت که زمان و توالی در آن دیده نشده‌اند.



شکل ۳. دیاگرام توالی زیرسیستم تحلیل واژگانی

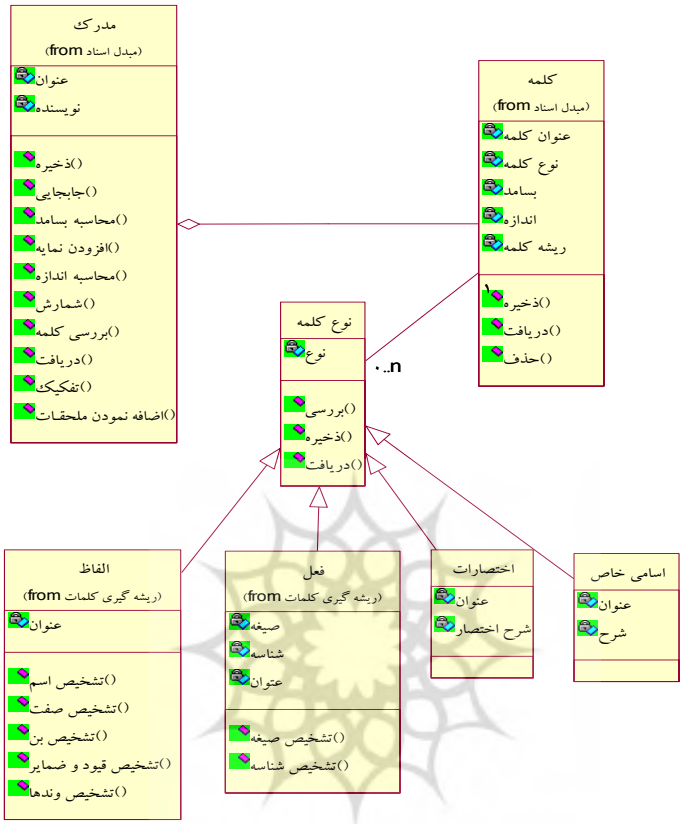
1. Class diagram
2. Sequence diagram



شکل ٤. دیاگرام همکاری زیرسیستم تحلیل واژگانی

شکل ۵، دیاگرام کلاس زیرسیستم تحلیل واژگانی را نمایش می‌دهد. البته این دیاگرام خلاصه‌شده و برخی مؤلفه‌های جزئی مربوط به توسعه نرم‌افزار از آن حذف شده است. همان‌گونه که دیده می‌شود این دیاگرام شباهت زیادی به مدل داده‌ای[1] دارد و مبنای اصلی توسعه پایگاه داده نرم‌افزار است.

1. Data model



شکل ۵. دیاگرام کلاس زیرسیستم تحلیل واژگانی

٤-٦. شناخت نیازمندی‌های ساخت نمایه‌ساز ماشینی در پژوهشگاه علوم و فناوری اطلاعات ایران

پژوهشگاه علوم و فناوری اطلاعات ایران که پیش‌تر با نام‌های دیگری شناخته می‌شد، از دیرباز در حوزه نمایه‌سازی کار تخصصی می‌کرده (ایرانشاهی ۱۳۸۲؛ علیدوستی و کاظم‌پور ۱۳۸۴) و وظیفه سازمان‌دهی بسیاری از مدارک علمی از جمله پایان‌نامه‌ها را در کشور برعهده داشته است[۱]. پایگاه وب این پژوهشگاه نیز سال‌هاست که

۱. نگاه کنید به «درباره پژوهشگاه علوم و فناوری اطلاعات ایران» http://irandoc.ac.ir/about-us/about-us.html

درگاهی بنیادین برای دسترسی به مدارک علمی به‌ویژه پایان‌نامه‌ها به‌شمار می‌رود. از سوی دیگر افزایش قابل توجه حجم اطلاعات تولیدشده در کشور، در کنار کمبود نیروی انسانی می‌تواند دشواری‌های روزافزونی را در سازمان‌دهی اطلاعات برای این پژوهشگاه پدید آورد که نمایه‌ساز ماشینی پاسخی برای آنها خواهد بود.

پیش‌تر کارهای غیرمتمرکزی در حوزه‌های مرتبط با نمایه‌سازی ماشینی در پژوهشگاه انجام شده‌اند که می‌توانند بنیادی برای ساخت سیستم متمرکز نمایه‌ساز ماشینی باشند. از جمله مهم‌ترین این کارها می‌توان به پروژه‌های واژه‌شکن فارسی، فعل‌شکن فارسی، اصطلاحنامه جامع، و «او سی آر»[1] اشاره کرد.

◇ پروژه واژه‌شکن فارسی

برنامه‌ای رایانه‌ای است که کلمات پیچیده زبان فارسی را به اجزای سازنده آن تجزیه، و مقوله دستوری و معنایی آن را مشخص می‌کند. این نرم‌افزار دارای حافظه‌ای است که کلمات بسیط و تک‌واژها و قواعد ساخت آنها در آن قرار دارد. نرم‌افزار با عرضه کلمه مفروض به آن، ابتدا صورت کامل واژه را در حافظه جست‌وجو و مقوله دستوری و معنایی آن را تعیین می‌کند. چنانچه صورت کامل واژه در حافظه وجود نداشت، در پی تجزیه آن برمی‌آید و اگر اجزاء واژه طبق قواعد داده‌شده در کنار هم آمده باشند مقوله دستوری و معنایی، آن را به کاربر عرضه می‌کند. چنانچه اجزاء مطابق قواعد نباشد، صرفاً به تجزیه واژه پیچیده اکتفا می‌کند. واژه‌شکن فارسی هم کلمات مشتق و هم کلمات مرکب و مرکب بن‌ساختی را تجزیه می‌کند و هم قواعد ترکیب حوزه اشتقاقی و حوزه کلمات مرکب و کلمات بن‌ساختی را می‌شناسد (سمائی ۱۳۸۵).

این نرم‌افزار می‌تواند یکی از اجزای اصلی نمایه‌ساز ماشینی، و یکی از کارکردهای آن نیز شناسایی ریشه واژه‌ها برای شمارش واژه‌های هم‌ریشه یا هم‌خانواده آن درون متن باشد.

◇ فعل‌شکن فارسی

پروژه فعل‌شکن فارسی طرح نرم‌افزاری است که برای استخراج قواعد تشخیص و تجزیه فعل‌های ساده طراحی شده است؛ که البته در آن فعل‌های مرکب و پیشوندی

1. Optical Character Recognition: OCR

بررسی نشده‌اند (سمائی ۱۳۸۸). چون این کار به مرحله اجرا درنیامده است، نیاز به تکمیل و تبدیل به یک بخش نرم‌افزاری از نمایه‌ساز ماشینی دارد. همان‌گونه که پیش‌تر نیز اشاره شد، فعل‌شکن زبان فارسی یکی از کلیدی‌ترین بخش‌های نمایه‌ساز ماشینی است. استفاده از این قابلیت به نمایه‌ساز این اجازه را می‌دهد که ریشه افعال گوناگون را تشخیص دهد و فراوانی آنها را بر اساس ریشه فعل محاسبه کند.

◇ اصطلاحنامه جامع

اصطلاحنامه جامع[1] که هم‌اکنون دارای ده اصطلاحنامه در زمینه‌های گوناگون مانند «زیست‌شناسی»، «زمین‌شناسی»، «شیمی»، «فنی ـ مهندسی»، «فیزیک»، «ریاضی» و «کشاورزی» است، دارای کارکردهای گوناگونی از جمله خودکارسازی نمایه‌سازی مدارک علمی فارسی است (بهشتی ۱۳۷۸). اصطلاحنامه‌ها از مهم‌ترین داده‌های ثابت در سیستم نمایه‌ساز ماشینی هستند و هر چه دقیق‌تر و جامع‌تر باشند، دقت نمایه‌سازی را بهتر خواهند کرد. اصطلاحنامه جامع، مشکل تعدد مدخل‌ها در حوزه‌های موضوعی مختلف را نیز برطرف می‌کنند.

◇ نرم‌افزار «او سی آر»

نرم‌افزار «او سی آر» دارای قابلیت شناسایی خودکار بخش چکیده پایان‌نامه‌ها و تبدیل آنها به متن با استفاده از قابلیتهای نرم‌افزارهای موجود و تشخیص شماره صفحات مربوط به بخش عناوین، فهرست مطالب، و شماره صفحات اصلی پایان‌نامه‌ها و ایجاد پیوند میان آنها و متن است. این نرم‌افزار هم می‌تواند بر روی پایگاه‌های داده، نصب و هم به‌صورت جداگانه استفاده شود (فرامرزی ۱۳۸۲). «او سی آر» نقش مبدل اطلاعات غیردیجیتال به دیجیتال را در نمایه‌سازی ماشینی ایفا می‌کند و بدون آن نرم‌افزار نمی‌تواند مدارک غیردیجیتالی را نمایه‌سازی کند.

٤-٧. ارزیابی وضعیت پژوهشگاه علوم و فناوری اطلاعات ایران برای ساخت نمایه‌ساز ماشینی

بررسی وضعیت پژوهشگاه برای ساخت نمایه‌ساز ماشینی نشان می‌دهد که در زمینه زیرسیستم‌های گوناگون دارای سطوح بلوغ گوناگونی است. وضعیت کنونی پژوهشگاه در زمینه هر یک از زیرسیستم‌های نمایه‌ساز ماشینی در جدول ۱ نشان داده شده‌اند.

1. Meta-Thesaurus

جدول ۱. ارزیابی وضعیت پژوهشگاه در زمینه زیرسیستم‌های نمایه‌ساز ماشینی

زیرسیستم	وضعیت پژوهشگاه
مبدل اسناد	مهم‌ترین بخش در این زیرسیستم تبدیل اسناد غیردیجیتال به دیجیتال است که بیشتر آن با «اوسی‌آر» انجام می‌شود. پژوهشگاه هم‌اکنون تجربه کافی برای ساخت این زیرسیستم را دارد، ولی نرم‌افزار موجود برای نیازهای سیستم نمایه‌ساز ماشینی کافی نیست. افزون بر این، پژوهشگاه تجربه قابل توجهی در حوزه دیجیتال‌سازی مدارک دارد.
تحلیل واژگانی	دو پروژه فعل‌شکن فارسی و واژه‌شکن فارسی در این زیرسیستم نقش بسیاری دارند، ولی این دو برای تحلیل نحوی متن کافی نیستند و باید جزء قوی‌تری از ترکیب این دو ساخته شود. جزء تازه باید قوانین نحوی جمله‌ها را در حد نیاز در خود داشته باشد. از سوی دیگر نیز در زیرسیستم تحلیل واژگانی، جدول نام‌های خاص هست که اکنون به‌صورت دستی است و روزآوری می‌شود و رایانه‌ای‌کردن آن هنوز به سرانجام نرسیده است. برای تشخیص اختصارهای تازه نیز باید یک پایگاه قواعد ساخته شود.
فهرست حذفی	تاکنون کاری در این زمینه در پژوهشگاه انجام نشده است، ولی دانش نهان آن در میان کارشناسان وجود دارد. فهرست حذفی همچنین می‌تواند بر اساس قواعد خاصی نیز در هر مستند ساخته شود. به‌عنوان نمونه کلمات با بسامد زیاد و طول کم (که در جدول اسامی خاص و اختصارات هم نیستند) به احتمال زیاد می‌بایست حذف شوند. فهرست‌های حذفی می‌توانند در طول زمان و توسط ویراستار کامل‌ترشوند.
اصطلاح‌نامه	پژوهشگاه در زمینه اصطلاحنامه، تجربه‌و سرمایه قابل‌توجهی دارد که شاید نتوان آن را با هیچ سازمان دیگری مقایسه کرد. هم‌اکنون شمار زیادی اصطلاحنامه تخصصی در پژوهشگاه تألیف یا ترجمه شده و همه آنها نیز دیجیتالی هستند. در این باره، تجربه و نرم‌افزارهای اصطلاحنامه‌ای برخط (آنلاین) پژوهشگاه بنیاد خوبی را تشکیل می‌دهند، ولی به اصطلاحنامه‌های بیشتر و قوی‌تر و همچنین یکپارچگی بیشتری نیاز است. از سوی دیگر برای سیستم‌های بالغ‌تر باید این مجموعه‌ها و پیوند میان آنها بسیار قوی‌تر، و سازوکارهای یادگیری نیز به آنها افزوده شود.

زیرسیستم	وضعیت پژوهشگاه
ریشه‌گیری واژه‌ها	درباره ریشه‌گیری واژه‌ها، نرم‌افزار واژه‌شکن بخش زیادی از این زیرسیستم را تأمین می‌کند، ولی نیاز به تکمیل دارد. همچنین نرم‌افزار فعل‌شکن هم می‌تواند پس از پیاده‌سازی کمبودهای نرم‌افزار واژه‌شکن را تکمیل کند.
وزن‌دهی به اصطلاح نمایه	تاکنون کاری در این زمینه در پژوهشگاه انجام نشده است و تجربه چندانی هم وجود ندارد. کار در این زمینه نیازمند تجربه‌های آماری و نرم‌افزاری در حوزه الگوریتم‌های جست‌وجو و بازیابی است. الگوریتم‌های بازیابی اطلاعات اعم از آماری یا بُرداری نیازمند آزمایش و خطا و تنظیم‌های ویژه‌ای هستند تا به بهترین عملکرد خود دست یابند. این تنظیم‌ها می‌توانند از یک بافت به بافت دیگر تغییر کنند. الگوریتم‌های یادگیری ماشینی مکمل خوبی برای الگوریتم‌های بازیابی هستند تا کیفیت نمایه‌سازی را بیشینه کنند.

۵. نتیجه‌گیری

مدل ارائه‌شده برای نمایه‌ساز ماشینی در این مقاله، ابعاد مفهومی و نرم‌افزاری آن را به‌عنوان یک سیستم و نیز مبنایی را برای مطالعه بیشتر و انجام پروژه‌های عملیاتی در این زمینه در پژوهشگاه علوم و فناوری اطلاعات ایران نشان می‌دهد.

ساخت یک سیستم بالغ برای نمایه‌ساز ماشینی در این پژوهشگاه که همه ویژگی‌های مدل مفهومی را پشتیبانی کند، نیازمند گام‌های بسیاری است که آن را به یک پروژه بزرگ و نیازمند پشتیبانی‌های زیادی از دیدگاه مالی و دانش فنی بدل می‌کند. تأمین منابع برای انجام چنین پروژه‌ای در کوتاه‌مدت برای پژوهشگاه ناممکن است، از همین‌رو می‌توان آن را در زمان بلندتر و با گام‌های کوتاه‌تری اجرا کرد. پیروی از روش‌شناسی «آر.یو.پی.» نیز چنین امکانی را فراهم می‌کند.

بر این پایه در گام نخست، نمایه‌ساز ماشینی می‌تواند دستیار نمایه‌ساز باشد. در این گام سیستم جست‌وجوی اصطلاح‌های متن و بسامد واژه‌ها با اصطلاحنامه‌های کنونی انجام و در اختیار نمایه‌ساز قرار می‌گیرند. در گام بعدی، نمایه‌ساز کار سیستم را رهگیری خواهد کرد و نمایه‌های استخراج‌شده از سوی سیستم را ویرایش و نهایی می‌کند. گام

پایانی، استقلال سیستم نمایه‌ساز ماشینی و بازبینی دوره‌ای یا تصادفی نمایه‌هاست. در این گام کاستی‌های نمایه‌ساز ماشینی برطرف و پایگاه داده اصطلاحنامه‌ها، فهرست حذفی، پایگاه قواعد نحوی، و فرمول‌های انتخاب نمایه تکمیل و تنظیم می‌شوند. دستیابی به این سطح از بلوغ نمایه‌ساز ماشینی زمان‌بر و در بلندمدت شدنی است.

٦. منابع

ایرانشاهی، محمد. ۱۳۸۲. بررسی میزان همخوانی کلیدواژه‌های عنوان و توصیفگرهای نمایه‌سازان در پایگاه چکیده پایان‌نامه‌های ایران. گزارش طرح پژوهشی. تهران: پژوهشگاه اطلاعات و مدارک علمی ایران.

بهشتی، ملوک‌السادات. ۱۳۷۸. اصطلاحنامه جامع. گزارش طرح پژوهشی. تهران: پژوهشگاه اطلاعات و مدارک علمی ایران.

بهشتی، ملوک‌السادات. ۱۳۸۲. کاربرد اصطلاح‌شناسی و واژه‌گزینی در نمایه‌سازی ماشینی و بازیابی اطلاعات. *علوم اطلاع‌رسانی* ۱۸ (۳-۴): ۳۱-۴۴.

جلالی‌منش، عمار. ۱۳۹۰. طراحی مدل مفهومی یکپارچه نمایه ساز ماشینی برای منابع فارسی در پژوهشگاه علوم و فناوری اطلاعات ایران. گزارش طرح پژوهشی. تهران: پژوهشگاه علوم و فناوری اطلاعات ایران.

حسینی بهشتی، ملوک‌السادات. ۱۳۸۹. ارزیابی تأثیر کاربرد اصطلاحنامه جامع در نمایه‌سازی و بازیابی اطلاعات. گزارش طرح پژوهشی. تهران: پژوهشگاه اطلاعات و مدارک علمی ایران.

داوودآبادی، مرضیه. ۱۳۸۴. پردازش معنایی جملات و اجرای دستورات صادرشده به زبان فارسی. پایان‌نامه کارشناسی ارشد هوش مصنوعی، دانشکده مهندسی برق و کامپیوتر، دانشگاه صنعتی اصفهان.

دیانی، محمدحسین. ۱۳۷۹. *مواد و خدمات مرجع: فشرده درس‌ها*. تهران: انتشارات کتابخانه رایانه‌ای.

سمائی، سیدمهدی. ۱۳۸۵. *واژه‌شکن فارسی*. گزارش طرح پژوهشی. تهران: پژوهشگاه اطلاعات و مدارک علمی ایران.

سمائی، سیدمهدی. ۱۳۸۸. *فعل‌شکن فارسی*. گزارش طرح پژوهشی. تهران: پژوهشگاه اطلاعات و مدارک علمی ایران.

سنجی، مجیده. ۱۳۸۷. *شناسایی واژه‌های غیرمفهومی در نمایه‌سازی خودکار مدارک فارسی*. پایان‌نامه کارشناسی، دانشکده علوم تربیتی و روان‌شناسی، دانشگاه فردوسی مشهد.

علیدوستی، سیروس، و زهرا کاظم‌پور. ۱۳۸۴. *فرایند نمایه‌سازی*. گزارش طرح پژوهشی. تهران: پژوهشگاه اطلاعات و مدارک علمی ایران.

فرامرزی، اسماعیل. ۱۳۸۲. فاز مطالعاتی و امکان‌سنجی شناسایی شماره صفحات پایان‌نامه‌ها در بخش فهرست مطالب و صفحات اصلی جهت زمینه‌سازی برقراری ارتباط خودکار بین آنها. گزارش طرح پژوهشی. تهران: پژوهشگاه اطلاعات و مدارک علمی ایران.

نوروزی، علیرضا. ۱۳۸۰. نمایه‌سازی کتاب: راهنمای برای ناشران، نمایه‌سازان، کتابداران، مؤلفان و مترجمان. تهران: چاپار.

نیاکان، شهرزاد. ۱۳۸۳. بررسی کاربرد نمایه‌سازی ماشینی در کتابخانه‌ها. *علوم اطلاع‌رسانی* ۱۶ (۳-۴): ۴۹-۵۵.

Ballerini, J.-P., M. Büchel, R. Domenig, D. Knaus, B. Mateev, E. Mittendorf, P. Schäuble, P. Sheridan, and M. Wechsler. 1997. SPIDER retrieval system at TREC-5. In E. M. Voorhees & D.K. Harman (Eds.), *Information Technology: The Fifth Text Retrieval Conference (TREC-5)* (pp. 217-228). Gaithersburg, MD: NIST SP 500-238.

Calvert, Drusilla, and Hilary Calvert. 2010. Macrex, MACREX VERSION 8. http://www.macrex.com/ (accessed 28 Jan 2014).

Cisco, Susan L.1998. One foot in front of the other. *Inform* 12 (5): 20-32.

Cleveland, Donald B., and Ana D. Cleveland. 2001. *Introduction to indexing and abstracting.* 3rd ed. Englewood: Libraries Unlimited.

Editorium, The. 2010. DEXter for Microsoft Word. http://www.editorium.com/DEXter.htm (accessed 16 June 2010).

El-Beltagy, S. R., and A. Rafea. 2009. KP-Miner: Akeyphrase extraction system for English and Arabic documents. *Information Systems* 34:132– 144.

Ellis, David, Nigel Ford, and Jonathan furner. 1998. In search of the unknown user: Indexing, hypertext and the world wide web. *Journal of Documentation* 54 (1): 28-47.

Fuji, H., and W. B. Croft. 1993. A Comparison of Indexing Techniques for Japanese Text Retrieval. In the Proceedings of the 16th annual international ACM SIGIR Conference on Research and Development in Information Retrieval, 237 – 246. Pittsburg, PA, USA.

Indexer Research. 2010. *Cindex: Indexing Software for Windows and Macintosh*. http://indexres.com/soft.php (accessed 28 Jan 2014).

Lee, J. H. and J. S. Ahn. 1996. *Using n-Grams for Korean Text Retrieval*. In the Proceedings of the 19th annual international ACM SIGIR Conference on Research and Development in Information Retrieval, 216-224. Zurich, Switzerland.

Luhn, H. P. 1957. A statistical approach to mechanized encoding and searching of literary information. *IBM Journal of Research and Development* 1 (4): 309-317.

Mansour, N., R. A. Haraty, W. Daher, and M. Houri. 2008. An auto-indexing method for Arabic text. *Information Processing and Managemen*t 44 (4): 1538–1545.

Matsumura, N., Y. Ohsawa, and M. Ishizuka. 2002. *PAI: Automatic Indexing for Extracting Asserted Keywords from a Document*. American Association for Artificial Intelligence. www.aaai.org (accessed 28 Jan 2014).

Ohsawa, Y., N. E. Benson, and M. Yachida. 1998. KeyGraph: Automatic Indexing by Co-occurrence Graph based on Building Construction Metaphor. In the *Proceedings of the Advances in Digital Libraries Conference*, 12. http://portal.acm.org/citation.cfm?id=785950 (accessed 28 Jan 2014).

Online Dictionary for Library and Information Science (ODLIS). 2009. Automatic Indexing.

http://lu.com/odlis (accessed 28 Jan 2014).

Salton, G. 1986. Another look at automatic text-retrieval systems. *Communications of the ACM* 29 (7): 648-656.

Salton, G. 1989. Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer. Addison-Wesley.

Sky Software. 2010. Sky Index Professional. http://www.sky-software.com/contact_info/ContactInfo.htm (accessed 28 Jan 2014).

Woodruff G. and C. Plaunt. 1994. GIPSY: Automated Geographic Indexing of Text Documents. *Journal of the American Society for Information Science* 45 (9): 645–655.

Machine Indexer for Persian Resources: An Integrated Model for the Iranian Research Institute for Information Science and Technology

Ammar Jalalimanesh*
MS in Industrial Engineering, Lecturer; Iranian Research Institute for Information Science and Technology
Tehran, Iran

Sirous Alidousti[1]
PhD in Management; Assistant Professor; Iranian Research Institute for Information Science and Technology; Tehran, Iran

Mahmood Khosrowjerdi[2]
MD, Knowledge and Information Science;
Young Researchers club, Central Tehran Branch;
Islamic Azad University; Tehran, Iran

Iranian Journal of Information Processing & Management

Iranian Research Institute Iranian for Science and Technology
ISSN 2251-8223
eISSN 2251-8231
Indexed in LISA, SCOPUS & ISC
Vol.29 | No.2 | pp: 425-451
Winter 2014

Abstract: Machine indexer is referred to as a kind of indexing system in which keywords are extracted from the title or body of a text and organized in the index entries using computer algorithm. Althogh there is an increasing need for applying computer in indexing machine indexer has not been developed for Persian language yet. Therefore in this paper the results of investigating and identifying of theoretical and practical aspects of such system in the Iranian Research Institute for Information Science and Technology are presented. To do so, firstly the relationship between indexing and information retrieval, theoretical debates and the machine indexing revolution, the practical and procedural viewpoints of machine indexing, successful projects in different countries and for different languages, and the requirements of machine indexer for Persian language are discussed. Then the elements of a machine indexer for Persian language are defined and its conceptual model and also the relationships between the elements, the details of subsystems, and the indexer logical system are designed. Finally, the requirments of developing the system for the Iranian Research Institute for

* Corresponding Author:
jalalimanesh@irandoc.ac.ir
1. alidousti@irandoc.ac.ir
2. mkhosro@gmail.com

Information Science and Technology are discussed and the state of the art of the implemented projects and existing experiences in this regard are defined.

Keywords: Indexing; Machine indexing; UML; conceptual model; Iranian Research Institute for Information Science and Technology