

پژوهشگاه علوم و فناوری اطلاعات ایران

طراحی مدل مفهومی یکپارچه نمایه سازی ماشینی

منابع فارسی

عمار جلالی منش

1389

فهرست مطالب

1	خلاصه
2	مقدمه
5	1.نمایه‌سازی و بازیابی اطلاعات
12	2. نمایه‌سازی ماشینی: سیر تکامل و تحول نظری
13	2-1. مقدمه
14	2-2. مدل‌های فضای برداری
15	2-2-1. سیر تکامل مدل‌های فضای برداری
23	2-3. مدل‌های احتمالی
24	2-3-1. فضای پیشامد(رخداد)
25	2-4. مفروضه‌های مدل‌های بازیابی اطلاعات
25	2-5. مدل مفهومی بازیابی احتمالی اطلاعات
27	2-6. ربط و احتمال ربط
27	2-7. اصل رتبه بندی احتمال
29	2-8. مدل‌های احتمالی ربط
32	2-9. مدل‌های استنباطی عدم قطعیت
32	2-10. مسائل کلی مطروح در مدل‌های احتمالی بازیابی اطلاعات
36	3.نمایه‌سازی ماشینی: ویژگی‌ها و مفاهیم
37	3-1. نمایه‌سازی ماشینی
38	3-2. فنون مشترک در نمایه‌سازی ماشینی
40	3-2-1. تحلیل واژگانی
41	3-2-2. استفاده از فهرست حذفی
43	3-2-3. ریشه‌گیری کلمات
45	3-2-4. گزینش عبارت‌ها و اصطلاحات و نرمال‌سازی اصطلاحات
46	3-2-5. بازشناسی اسامی خاص
47	3-2-6. وزن‌دهی به اصطلاح نمایه
48	3-3. انواع نمایه‌سازی ماشینی
48	3-3-1. نمایه سازی کوئیک
49	3-3-2. نمایه کواک
49	23-4. نمایه سازی با کمک رایانه
52	4. نمایه‌سازی ماشینی: پروژه‌های موفق
53	4-1. تجارب نمایه‌ساز/نمایه‌سازی ماشینی
68	4-2. مروری بر پاره‌ای از نرم‌افزارهای موجود نمایه‌سازی ماشینی

72	5. نیازمندیهای نمایه‌ساز ماشینی برای زبان فارسی
73	5-1. تفاوت‌های نمایه‌سازی به زبان فارسی با سایر زبان‌ها
76	5-2. نیازمندیهای نمایه‌ساز ماشینی به زبان فارسی
78	5-3. اجزاء نمایه‌ساز ماشینی فارسی
78	5-3-1. زیرسیستم مبدل دیجیتال
79	5-3-2. زیرسیستم تحلیل واژگانی
79	5-3-3. زیرسیستم گزینش عبارتها و اصطلاحات و نرمال‌سازی
79	5-3-4. زیرسیستم فهرست حذفی
79	5-3-5. زیرسیستم ریشه‌گیری کلمات
80	5-3-6. زیر سیستم وزن‌دهی به اصطلاحات
80	5-4. تأثیرات زبان فارسی در اجزاء نمایه‌ساز فارسی
81	5-4-1. مسائل ریخت‌شناسی
81	5-4-2. مسائل نحوی
82	6. نمایه‌سازی ماشینی: طراحی مدل مفهومی
83	6-1. متدولوژی و معماری طراحی سیستم
85	6-2. دیدگاه مورد کاربرد
95	6-3. دیدگاه منطقی
96	6-3-1. زیرسیستم مبدل اسناد
99	6-3-2. زیرسیستم تحلیل واژگانی
104	6-3-3. زیرسیستم گزینش اصطلاحات و نرمال‌سازی
107	6-3-4. زیرسیستم فهرست حذفی
110	6-3-5. زیرسیستم ریشه‌گیری کلمات
114	6-3-6. زیرسیستم وزندهی به اصطلاح نمایه
118	7. نمایه‌ساز ماشینی: نیازمندیهای ایجاد سیستم در پژوهشگاه
120	7-1. واژه‌شکن
120	7-2. فعل‌شکن فارسی
121	7-3. اصطلاح‌نامه جامع
122	7-4. نرم‌افزار OCR
123	7-5. ارزیابی وضعیت فعلی پژوهشگاه برای ایجاد نمایه‌ساز ماشینی
126	8. نتیجه‌گیری
128	منابع

خلاصه

نمایه‌سازی یکی از ابزارهای تسهیل‌کننده بازیابی اطلاعات و سرویس‌های اطلاع‌رسانی است. این موضوع از این جهت مورد اهمیت است که هم جامعیت و هم مانعیت را در دسترسی به مدارک بالا می‌برد. به دلیل زمانبر بودن فرایندهای نمایه‌سازی و وابستگی شدید به انسان این ابزار هنوز نتوانسته است قابلیت‌های خود را در حوزه بازیابی اطلاعات گسترش داده و فراگیر شود. از اینرو سالهاست مطالعه بر روی کاهش وابستگی و یا بی‌نیاز کردن این فرایند از انسان شروع شده و فعالیت‌های زیادی برای عملیاتی کردن آن صورت گرفته است. این حوزه با نام نمایه‌سازی ماشینی و یا خودکار شناخته می‌شود. متأسفانه در کشور ما هنوز فعالیت‌های قابل اتکائی در حوزه مذکور صورت نگرفته و می‌توان گفت از حوزه‌های مغفول در رشته‌های زبان‌شناسی، کتابداری و اطلاع‌رسانی است. با توجه به سابقه پژوهشگاه علوم و فناوری اطلاعات ایران در زمینه‌های مطالعاتی و اجرایی زبان‌شناسی رایانه‌ای شاید بتوان گفت یکی از آماده‌ترین بسترها برای پایه‌ریزی و توسعه علمی و عملیاتی نمایه‌ساز ماشینی برای زبان فارسی باشد.

در همین راستا مطالعه حاضر به دنبال بررسی و شناخت جنبه‌های علمی و اجرایی این سیستم و بررسی نیازمندی‌های زبان فارسی و نیازمندی‌های اجرایی نمودن این پروژه در پژوهشگاه علوم و فناوری اطلاعات است. گزارش حاضر نتیجه مطالعات در این زمینه و مصاحبه با خبرگان حوزه‌های زبان‌شناسی و کتابداری و مشاوره متخصصان حوزه مهندسی نرم‌افزار می‌باشد.

در این گزارش به رابطه بین نمایه‌سازی و بازیابی اطلاعاتی (جستجو)، مباحث نظری و سیر تحول نمایه‌سازی ماشینی از دیدگاه تئوریک، نمایه‌سازی ماشینی به صورت عملیاتی و از نگاه فرایندی، پروژه‌های موفق انجام شده در دنیا و برای زبان‌های مختلف و نیازمندی‌های نمایه‌ساز ماشینی در زبان فارسی پرداخته شده است. سپس اجزاء یک نمایه‌ساز زبان فارسی تشریح گردیده و مدل مفهومی نمایه‌ساز ماشینی به صورت دقیق ترسیم شده و روابط بین اجزاء، جزئیات زیرسیستم‌ها و مدل منطقی اولیه سیستم طراحی شده است. در انتها نیازمندی‌های توسعه نمایه‌ساز ماشینی برای پژوهشگاه بحث شده و وضعیت پروژه‌هایی که در حال حاضر انجام شده و یا تجربیاتی که وجود دارد و می‌توانند در اجرای این پروژه مفید باشد مورد توجه قرار گرفته است.

مقدمه

یکی از مهمترین مسائلی که در حوزه مدیریت اطلاعات مورد توجه است، سازماندهی اطلاعات است. در این راستا نمایه‌سازی از عمومی‌ترین تکنیک‌های مورد استفاده است. با استفاده از نمایه‌سازی مستندات و منابع اطلاعاتی قابلیت بازیابی مناسب می‌یابند. در این روش نمایه‌ساز توصیف‌گرهایی را که هر یک نشانگر ارزش معنایی بخشی از مدرک هستند انتخاب می‌کند. امروزه به خاطر افزایش قابل توجه حجم دانش در دنیا و در نتیجه آن افزایش حجم مدارک علمی، نمایه‌سازی دستی، که کاری بسیار زمانبر است، جوابگوی نیازهای اطلاعاتی نیست. از اینرو نرم‌افزارهایی برای انجام فرایند نمایه‌سازی به صورت اتوماتیک طراحی شده‌اند که تا حد قابل قبولی سازماندهی اطلاعات را از نمایه‌سازی دستی بی‌نیاز می‌کنند. با این وجود، تاکنون در کشور ما به این مسئله به خوبی پرداخته نشده و تجربیات قابل اتکایی نیز در این زمینه وجود ندارد. ایجاد یک مکانیزم نمایه‌سازی ماشینی نیازمند وجود زیرسیستم‌های مختلفی از جمله اصطلاحنامه‌های دیجیتالی، الگوریتم‌های تطابق و پایگاه‌های دستور زبانی و منطقی است. البته نباید از نظر دور داشت که حتی سیستم‌های تجاری نمایه‌ساز ماشینی در دنیا از بلوغ کافی برخوردار نیستند و هنوز موتورهای جستجوی بزرگ از الگوریتم‌های آنها طبیعت نمی‌کنند.

با وجود سابقه طولانی نمایه‌سازی متون در زبان فارسی در کشور و وجود متخصصان زیادی در این حوزه تنها چند مطالعه نظری در این حوزه صورت گرفته و هنوز پروژه و یا نرم‌افزار قابل اتکایی در این حوزه وجود ندارد. همانطور که اشاره شد یک نمایه‌ساز ماشینی قوی از بلوک‌های ساختاری همچون اصطلاحنامه‌های دیجیتالی جامع، موتورهای استنتاج نحوی، پایگاه قواعد دستوری، الگوریتم‌های وزن‌دهی مختلف و مواردی از این دست تشکیل شده است. در واقع می‌توان گفت مباحث زبان-شناسی رایانه‌ای مبانی اصلی این سیستم‌ها را تشکیل می‌دهد و با توجه به اینکه این مباحث در کشور نسبتاً مغفول مانده بلوک‌های ساختاری مورد نیاز برای ساخت نمایه‌ساز ماشینی که از جهاتی نیز شبیه سیستم‌هایی همچون مترجم ماشینی و یا سیستم‌های تشخیص گفتار است هنوز آماده ارائه نیستند.

پژوهش حاضر در راستای حرکت در جهت تحقق این مهم سعی در انجام مطالعه بر روی نیازمندی‌های ایجاد یک سیستم نمایه‌ساز ماشینی برای زبان فارسی با تکیه بر نظریات، تجربیات و پروژه‌های موفق این حوزه دارد. از دیگر اهداف این تحقیق رسیدن به یک مدل مفهومی جامع و مدلسازی با استفاده از زبان استاندارد UML است. در نهایت بررسی نیازمندی‌های ایجاد این سیستم در پژوهشگاه علوم و فناوری اطلاعات ایران از خروجی‌های دیگر آن می‌باشد.

در راه انجام طرح مشکلات زیادی پیش رو بود که مهمترین آنها نبود مطالعات کافی در حوزه زبان‌شناسی رایانه‌ای برای زبان فارسی بود. همچنین عدم اشراف متخصصان زبان‌شناسی به مباحث مهندسی نرم‌افزار و تفاوت زاویه دید آنها با آنچه ما در این پژوهش دنبال می‌کردیم کار را تا حدودی مشکل می‌کرد.

در این پژوهش در ابتدا مطالعاتی در حوزه جایگاه نمایه‌سازی در بازیابی اطلاعات و مباحث نظری حوزه نمایه‌سازی ماشینی صورت گرفت و سپس مدل‌های عملیاتی در حوزه مورد بررسی قرار گرفت در قدم بعدی برخی پروژه‌ها و نرم‌افزارهای موفق در این حوزه مورد توجه و مطالعه قرار گرفت. سپس اقتضائات زبان فارسی و تاثیرات آن بر سیستم نمایه‌ساز ماشینی استخراج شده و بر این اساس بلوک‌های تشکیل‌دهنده نمایه‌ساز ماشینی برای زبان فارسی استخراج گردید. در ادامه مدل مفهومی با استفاده از زبان استاندارد UML ترسیم شد و در انتهای کار وضعیت پژوهشگاه و آمادگی آن برای اجرای چنین سیستمی مورد ارزیابی قرار گرفت.

بر همین اساس فصل اول این گزارش به رابطه بین نمایه‌سازی و بازیابی اطلاعاتی (جستجو) می‌پردازد. فصل دوم به مباحث نظری و سیر تحول موضوع نمایه‌سازی ماشینی از دیدگاه تئوریک می‌پردازد. در فصل سوم به نمایه‌سازی ماشینی به صورت عملیاتی و از نگاه فرایندی پرداخته شده است. در این فصل به تکنیک‌هایی که به صورت مشترک در حوزه مورد استفاده قرار می‌گیرد ارائه گردیده و انواع مختلف نمایه‌سازی ماشینی تشریح گردیده است. در فصل چهارم پروژه‌های موفق انجام شده در دنیا و برای زبان‌های مختلف مورد مطالعه قرار گرفته و خصوصیات و توانایی‌های آنها ذکر شده است. فصل پنجم به نیازمندی‌های نمایه‌ساز ماشینی در زبان فارسی می‌پردازد. در ابتدا نمایه‌سازی به زبان فارسی در مقایسه با سایر زبان‌ها مورد بررسی قرار گرفته و تفاوت‌های زبانی ذکر شده است. سپس اجزاء یک نمایه‌ساز زبان فارسی تشریح گردیده است. در فصل ششم مدل مفهومی نمایه‌ساز

ماشینی به صورت دقیق ترسیم شده و روابط بین اجزا، جزئیات زیرسیستم‌ها و مدل منطقی اولیه سیستم طراحی شده است. در فصل هفتم در مورد نیازمندیهای توسعه نمایه‌ساز ماشینی برای پژوهشگاه بحث شده و وضعیت پروژه‌هایی که در حال حاضر صورت گرفته و یا تجربیاتی که وجود دارد و می‌توانند در اجرای این پروژه مفید باشد مورد توجه قرار گرفته است. در انتها نیز یک نتیجه‌گیری کلی از پژوهش ارائه شده است.

۱. نمیه سازی و بازیابی اطلاعات

یک مخزن اطلاعات¹ تا زمانی که مدارک موجود در آن سازماندهی نشوند، تنها اتاقی پر از مدرک است. در این میان، سازماندهی اطلاعات فرایندی است که در آن اطلاعات در شکلهای گوناگون برای استفاده کاربران دستکاری شده و با ابزارهای جست و جو قابل بازیابی می شوند (Cleveland and Cleveland 2001, 29, 34). بدین ترتیب سازماندهی اطلاعات را می توان مجموعه فعالیتهایی تعریف کرد که در طی آن مدارک و منابع اطلاعاتی تحلیل، طبقه بندی و قابل بازیابی می شوند. بنابراین هدف آن دسترسی بهتر و سریع تر به اطلاعات است. چنین فرایندی برای شناسایی محتوا و خصیصه های مدارک با هدف جست و جو و بازیابی سریع، ساهاست که موضوع کار متخصصان بوده است. یکی از فعالیتهای مهم در این فرایند، نمایه سازی یا به عبارتی انتساب مجموعه ای از کدها یا واژه ها به مدرک برای نمایش محتوای موضوعی آن است (Ellis, Ford and Furner 1998). نمایه سازی فعالیتی حیاتی در اطلاع رسانی محسوب می شود. اغلب کاربران کتابخانه ها نیازمند ابزارهای جست و جویی هستند که در یافتن مدارک موضوعی خاص، آثار نویسنده ای معین، یا عنوانی مشخص مفید باشند. ایجاد نمایه ها تلاشی در پاسخ به نیاز آنها است که طی فرایند نمایه سازی تولید می شوند (Booth 2001, 2).

بدین ترتیب، نمایه سازی یکی از روشهای مؤثر در سازماندهی و جست و جوی اطلاعات است. نمایه سازی در کاهش زمان بازیابی مدارک نقش دارد و از نمایه ها برای گروه بندی رکوردها، شناسایی مکان فیزیکی آنها، و بازیابی سریع استفاده می شود. بدین ترتیب، کاربران، جست و جو میان تعداد انبوهی از مدارک غیر ضروری وقت خود را هدر نمی دهند (Cisco 1998). بر این اساس، فرایند نمایه سازی دو هدف اصلی را دنبال می کند:

1. به حداقل رساندن زمان و تلاشها در جست و جوی اطلاعات.

2. به حداکثر رساندن موفقیت جست و جو.

با انتخاب بهترین واژه ها که زبان جست و جوی کاربر را با نظامی از ارجاعات صحیح و کامل متناسب می کند، هر دو هدف برآورده می شوند (Cleveland and Cleveland 2001, 31). «گین گراند» نیز هدف نمایه سازی را هم جهت شدن با اهداف جست و جوی کاربران در یافتن مدارک، بر اساس شناخت محتوای آنها می داند. گسترش این فرایند برای تعیین مکان صحیح مدارک و یافتن سریع آنها مفید است (Gingrande 2004; Sauperl 2004).

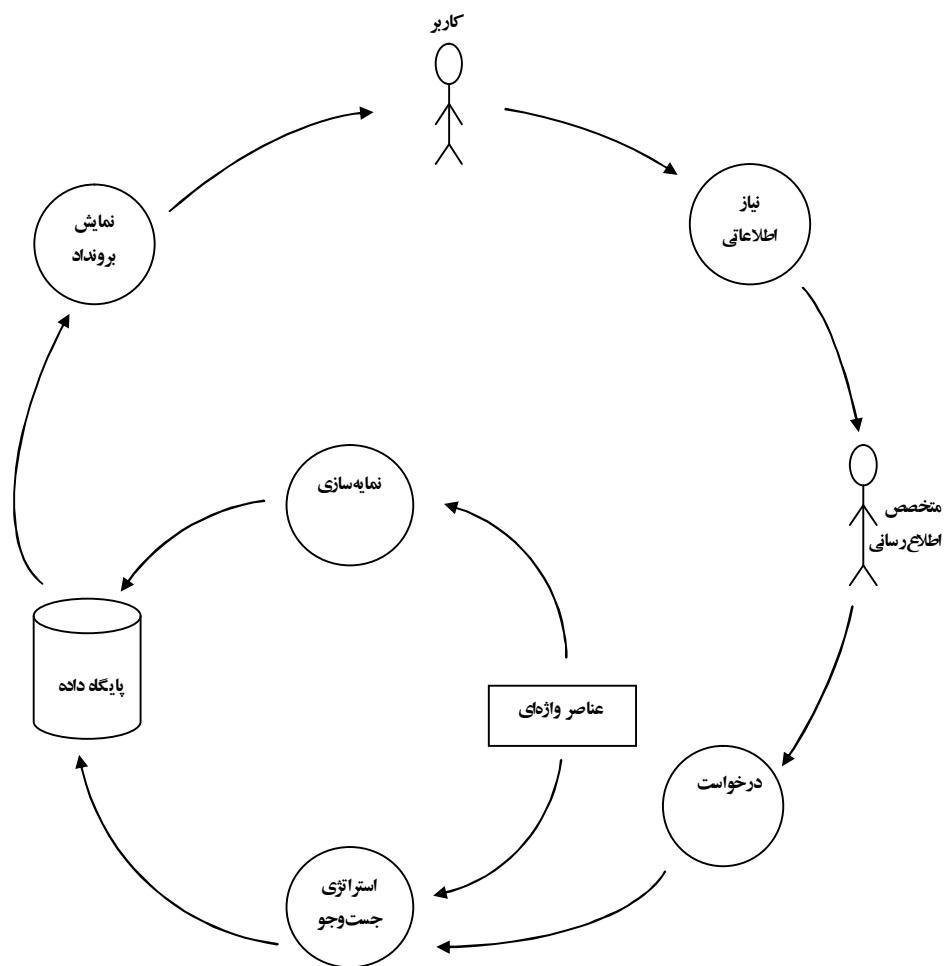
بدین ترتیب هدف نمایه سازی راهنمایی کاربر به محتوا و مکان فیزیکی مدارک برای بازیابی است (Cisco 1998).

¹ information store

از طرفی می‌توان هدف این فرایند را دسترس‌پذیری اطلاعات و سازماندهی فایل‌های قابل جست‌وجو دانست (Croghan 2002) و می‌توان هدفهای آن را به صورت زیر دسته‌بندی کرد:

1. صرفه‌جویی در وقت کاربران هنگام جست‌وجوی اطلاعات.
2. تعیین محتوا و مکان مدارک مورد نیاز.
3. افزایش دسترس‌پذیری اطلاعات.
4. انتساب نمادهای موضوعی به مدرک.

«لنکستر» در تبیین فرایند بازیابی اطلاعات که در یک پایگاه اطلاعات رخ می‌دهد (شکل یک)، رابطه میان نمایه‌سازی و جست‌وجوی اطلاعات را بهتر نمایش داده است.



شکل 1-1. فرایند بازیابی اطلاعات در پایگاه اطلاعات (Lancaster 1991, 75)

بدین ترتیب می‌توان گفت که نمایه‌سازی است که در بازیابی مدارک حاوی اطلاعات به کاربر کمک می‌کند. با رشد کمی مدارک، پیچیدگی مدارک انتشار یافته و با درک اهمیت اطلاعات در تحقیق و تصمیم‌گیری، اهمیت نمایه پیوسته رو به رشد است. «پراشر» کاربردهای نمایه را موارد زیر برمی‌شمارد (Prasher 1989, 2-3):

1. نمایه‌ها کاربران را به سمت مدارک هدایت می‌کنند.
2. نمایه‌های اسامی نویسندگان¹، کاربران را به نامهای افراد راهنمایی کرده و محققان را به سمت مدارک وابسته به موضوع هدایت می‌کنند (زیرا نویسندگان به سمت تخصصی شدن تمایل دارند).
3. نمایه‌ها کاربر را به سمت منابعی که از وجود آنها مطلع است یا به سمت مدارکی که از وجود آنها ناآگاه است هدایت می‌کنند (بنابراین نمایه در جهت پاسخ به پرسشهای فراخواندن² یا کشف کردن³ مدارک استفاده می‌شود).
4. بدون نمایه‌ها، محقق وقت خود را صرف جست‌وجوی صفحه به صفحه مدارک می‌کند بنابراین نمایه‌ها در وقت صرفه‌جویی کرده و موجب جست‌وجوی سودمند می‌شوند.
5. نمایه‌ها در شکلی بسیار فشرده، اطلاعات مفید درباره یک نویسنده یا یک زمینه موضوعی فراهم می‌کنند. در نمایه‌های اسامی نویسندگان، زمینه‌های جدید مورد علاقه نویسنده به سهولت بررسی می‌شود همچنین تغییر در علایق یک نویسنده در طول زمان با مجموعه‌ای از این نمایه‌ها مشخص می‌شود.
6. ارجاعات در نمایه‌های موضوعی کاربر را به واقعیتهای پذیرفته‌شده‌ای از یک زمینه موضوعی هدایت می‌کنند و وجود واژه‌های مترادف در نمایه‌ها برای ورود به زمینه‌های موضوعی جدید مؤثر هستند.
7. فهرست اسامی و اصطلاحات، اغلب به وسیله نمایه‌ها به شیوه‌ای مؤثر ایجاد می‌شوند.
8. سازماندهی زمینه‌ای از دانش با نمایه رده‌ای یا به وسیله ارجاعات در نمایه موضوعی امکان‌پذیر می‌گردد.

¹ author indexes

² recall

³ discovery

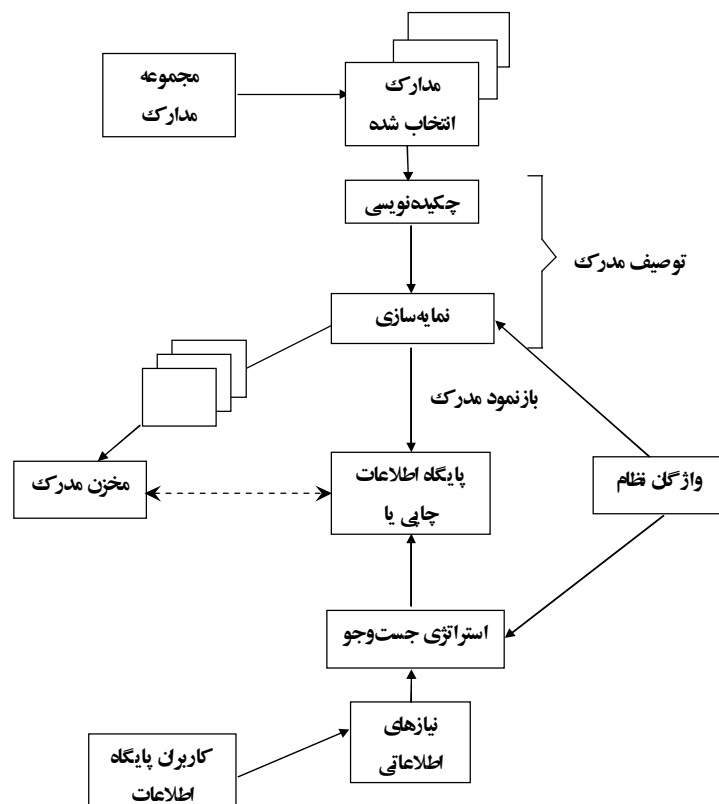
9. نمایه‌ها به حل مشکلات تنوع زبانها کمک می‌کنند. تعداد کمی از کاربران با بیش از یک زبان آشنایی دارند، نمایه‌ها فرد را به منابع موجود در یک زبان هدایت می‌کنند و به محقق در تعیین روابط میان نمایه و مدرک کمک می‌کنند.

10. مشکل انتشار تعداد بسیار زیاد مدارک بطور سالانه، با نمایه‌ها قابل حل است زیرا نمایه‌ها انتخاب سریع منابع مرتبط را به وجود می‌آورند و محقق به سهولت می‌تواند از میان منابع بی‌شمار در مدت زمان کوتاهی به انتخاب بپردازد.

«نیزو» کاربرد نمایه را فراهم‌آوری ابزار نظام‌یافته و کارآمدی برای تعیین مکان مدارک یا بخشهایی از آنها که با نیازها و درخواستهای اطلاعاتی متناسب هستند می‌داند. بنابراین نمایه باید (NISO 1997):

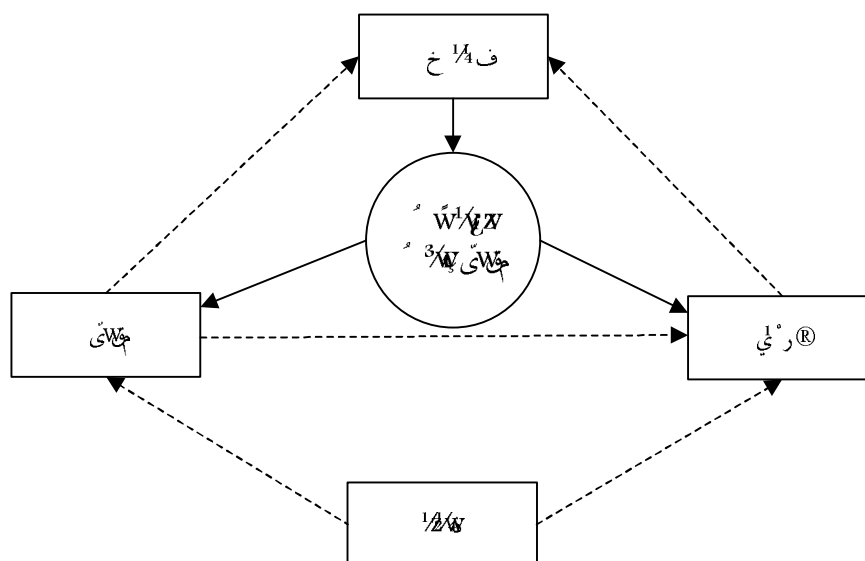
1. مدارکی را شناسایی کند که موضوعها یا ویژگیهای خاصی در بر دارند؛
2. همه موضوعها یا مشخصه‌های مهم یک مدرک را مطابق با سطح جامعیتی نشان دهد که برای نمایه در نظر گرفته شده است؛
3. میان موضوعهایی که دارای اهمیت زیاد هستند و آنهایی که از اهمیت کمتری برخوردارند تمایز قائل شود؛
4. امکان دسترسی به موضوعها یا مشخصه‌های مدارک را به وسیله واژه‌های احتمالی کاربران فراهم کند؛
5. تا جایی که زبان نمایه‌سازی اجازه می‌دهد از واژه‌های تخصصی استفاده کند؛
6. امکان دسترسی از طریق اصطلاحات مترادف و معادل را فراهم کند؛
7. کاربران را به اصطلاحاتی که مفاهیم وابسته را نشان می‌دهند (اصطلاحات اخص، اعم و وابسته) راهنمایی کند؛
8. امکان ترکیب اصطلاحات را برای سهولت شناسایی جنبه‌های خاص موضوعها، مشخصه‌ها، یا حذف جنبه‌هایی که مورد درخواست نیست، فراهم کند؛
9. ابزاری برای جست‌وجوی موضوعها یا مشخصه‌های خاص به وسیله ترکیب نظام‌یافته‌ای از مدخلها (در نمایه‌های قابل مشاهده) یا با روشی مستند برای ورود، ترکیب، و اصلاح واژه‌ها برای ایجاد گزاره جست‌وجو و بازبینی موارد بازیابی شده (در نمایه‌های غیر قابل مشاهده) فراهم کند.

«لنکستر» نقش نمایه‌سازی در جست‌وجو و بازیابی اطلاعات را در شکل دو نمایش می‌دهد. همانطور که در شکل دیده می‌شود؛ ابتدا طرح پایگاه اطلاعات، از میان انبوهی از مدارک انتشار یافته بر اساس معیارهای خاص پایگاه، جایگزین‌ها را گزینش می‌کند. آشکارترین معیار، موضوع مدرک است و دیگر معیارها می‌توانند نوع مدرک، زبان، یا منبع آن باشند. موارد منتخب برای ورود به پایگاه به شیوه‌های گوناگون مانند نمایه‌سازی و چکیده‌نویسی توصیف می‌شوند. اصطلاحاتی که در نمایه‌سازی استفاده می‌شوند از واژگان کنترل شده (واژگان نظام) استخراج می‌شوند. اما در مواردی می‌توان در نمایه‌سازی از اصطلاحات مدرک استفاده کرد. فعالیتهای توصیفی موجب ایجاد بازنمودهایی برای مدارک می‌شوند. در سوی دیگر کاربران وجود دارند که با اهداف گوناگون از پایگاه اطلاعات استفاده می‌کنند. آنها نیاز اطلاعاتی خود را مطرح می‌کنند؛ این نیاز باید تغییر کند و به شکلی از استراتژی جست‌وجو تبدیل شود. در اینجا نیز از زبان نمایه استفاده می‌شود، بنابراین نقش نمایه در بازیابی مدارک مرتبط با نیاز اطلاعاتی آشکار می‌شود و تنها با استفاده از نمایه می‌توان میان مدارک موجود در پایگاه و درخواستهای کاربران انطباق برقرار کرد و مدارک مرتبط را بازیابی نمود (Lancaster 1991, 1-4).



شکل 1-2. نقش نمایه‌سازی در بازیابی اطلاعات (Lancaster 1991, 1-4)

نمایه‌سازی و بازیابی اطلاعات در عمل نمی‌توانند از یکدیگر مجزا باشند. هنگامی که پرسش جست‌وجو مطرح می‌شود؛ سؤال یا مسئله اطلاعاتی را نمایه‌سازی می‌کنیم. همچنین مدارک هنگام ورود به پایگاه نمایه‌سازی می‌شوند. بدین ترتیب می‌توان میان پرسش جست‌وجو و مدارک با استفاده از اصطلاحات نمایه‌ای ارتباط برقرار کرد. پس نمایه‌ها ابزارهایی برای یافتن اطلاعات مورد نیاز کاربران هستند که به اطلاعات مفید اشاره دارند. نمایه پلی بر روی آبهای ناآرام اطلاعات است که کاربر را به منبع اطلاعات مرتبط هدایت می‌کند و نقش آن در بازیابی اطلاعات کاملاً آشکار است. می‌توان رابطه میان نمایه‌سازی و جست‌وجوی اطلاعات را در شکل سه نشان داد (Cleveland and Cleveland 2001, 32-34):



شکل 1-3. رابطه نمایه‌سازی و جست‌وجوی اطلاعات (Cleveland and Cleveland 2001, 32-34)

۲. نمای سازی ماشینی:
سیرتقابل و تحول نظری

2-1. مقدمه

اندیشه استفاده از نمایه سازی ماشینی را می توان هم زاد با علاقه به مدل های برداری و احتمالی بازیابی اطلاعات در دهه 1960 دانست. تحقیقاتی که بر روی پروژه 9IR/X توسط «لوسینگ»¹ و «فانگ مه یو»² صورت گرفت منجر به بین المللی شدن این حرکت ها (در راستای نمایه سازی ماشینی) در سال های 1968 تا 1969 شد. به موازات آن، پژوهش هایی نیز در ارتباط با نمایه سازی بر اساس روش های زبان شناختی صورت پذیرفت. در آلمان از آغاز سال های دهه 70 مطالعات ریخت شناسی بر روی اجزای سیستم های ترجمه با هدف بازیابی اطلاعات صورت پذیرفت. طرحی به نام «پاسات»³ که در اصل در زمینه مطالعه زبان های رایانه ای شکل گرفته بود، باعث به وجود آمدن چالش هایی در سال 1969 در این حوزه گردید. در دهه 80 علاقه به سیستم های هوشمند که بسیاری از دانشمندان در تحلیل محتوا به آن اعتماد داشتند افزایش یافت و منجر به انجام برخی پروژه ها با این رویکرد گردید. برای نمونه، سیستم دانشگاه «ییل»⁴ و «برلین»⁵ را می توان نام برد (نیاکان 1383). با توجه به موارد ذکر شده، در ادامه به مطالعه تاریخی مدل های برداری و احتمالی که بنیان مدل های نمایه سازی ماشینی امروزی قلمداد می شوند پرداخته می شود. این رویکردها را می توان دقیقاً همان رویکردهای مطرح در نمایه سازی ماشینی تلقی کرد.

¹ Lusting

² Fangmeyer

³ PASSAT

⁴ Yale

⁵ Berlin

2-2. مدل‌های فضای برداری

مدل‌های فضای برداری بازیابی اطلاعات زیرمجموعه‌ای از فنون بازیابی اطلاعات است که در سال‌های اخیر مورد توجه واقع شده‌اند. مدل‌های فضای برداری بر این فرض استوارند که می‌توان معنای یک مدرک را از اصطلاحات متشکله مدرک بدست آورد.

در این مدل‌ها، مدارک به صورت بردارهایی از اصطلاحات

$$d = t_1, t_2, \dots, t_n$$

نشان داده می‌شوند که $t_i (1 \leq i \leq n)$ مقداری غیر منفی است و بر دفعات وقوع اصطلاح i در مدرک d دلالت دارد. بنابراین هر اصطلاح واحد در مجموعه مدرک، متناظر با بعدی در فضا است و

$$q = t_1^{\wedge}, t_2^{\wedge}, \dots, t_m^{\wedge}$$
 هر پرسش نیز به صورت یک بردار

بیان می‌شود که اصطلاح $t_i^{\wedge} (1 \leq i \leq m)$ مقداری غیر منفی است و به تعداد وقوع اصطلاح $t_i^{\wedge}$ در پرسش، اشاره دارد. هم بردارهای مدرک و هم بردارهای پرسش، مکان اقلام اطلاعاتی را در فضای اصطلاح - مدرک مشخص می‌کند. با محاسبه فاصله بین پرسش و اقلام اطلاعاتی دیگر موجود در فضا، آن‌هایی که دارای محتوای معنایی مشابه با پرسش هستند، بازیابی خواهند شد.

مدل‌های فضای برداری به منظور رفع پاره‌ای از مشکلات مربوط به فنون تطبیق واژگانی¹ به وجود آمدند (Letsche 1996). از آنجاییکه معمولاً کلمات معانی مختلفی دارند (چند معنایی)² مشکل است که بتوان بدون درک متنی که واژه یا اصطلاح مورد نظر در آن به کار برده شده است و صرفاً با روش تطبیق واژگانی، بین دو مدرکی که به طور مشترک از یک واژه یا اصطلاح استفاده کرده‌اند، تمایزی قائل شد. چون ممکن است معانی مورد نظر هر یک از مدارک برای آن اصطلاح خاص، یکسان نباشد. از طرف دیگر روش‌های متفاوتی برای توصیف یک مفهوم خاص وجود دارد (هم معنایی)³ و این احتمال وجود دارد که مدارک مرتبط به یکدیگر، برای توصیف مفاهیم مشابه از اصطلاحات یکسانی استفاده نکنند. بنابراین پرسشی که یکی از اصطلاحات دو مدرک را به کار می‌برد، مدرک مربوط

¹ Word matching

² Polysemy

³ Synonymy

دیگر را بازیابی نخواهد کرد. در بدترین حالت، پرسشی که از واژگانی متفاوت از اصطلاحات بکار رفته در مجموعه مدارک مربوط استفاده می‌کند، هیچ مدرکی را با استفاده از تطبیق واژگانی بازیابی نخواهد کرد علی‌رغم اینکه مجموعه ما مدارک مربوط را داراست.

مدل‌های فضای برداری با قرار دادن اصطلاحات، مدارک و پرسش‌ها در یک فضای اصطلاح-مدرک و محاسبه همانندی بین این درخواست‌ها و اصطلاحات یا مدارک، این امکان را به وجود می‌آورد که نتایج پرسش بر اساس میزان همانندی، رتبه‌بندی شود.

برخلاف فنون تطبیق واژگانی که هیچ نوع رتبه‌بندی انجام نمی‌دهند و یا از یک نظام رتبه‌بندی خام و ابتدایی استفاده می‌کنند (به عنوان مثال، یک مدرک را فقط به خاطر اینکه دارای تعداد وقوع بیشتری از اصطلاحات کاوش شده است، قبل از مدرک دیگری قرار می‌دهد)، مدل‌های فضای برداری، رتبه‌بندی را بر اساس فاصله اقلیدسی¹ یا مقیاس زاویه‌ای بین پرسش و اصطلاحات یا مدارک موجود در فضا انجام می‌دهد و قادر است کاربر را به مدارکی هدایت کند که تشابه مفهومی بیشتری با پرسش وی داشته و از مدارک دیگر کاربردی‌تر باشد (Letsche 1996).

علاوه بر آن، مدل‌های فضای برداری این مزیت را دارد که روش دقیق‌تری برای اجرای بازخورد ربط² فراهم کند.

2-2-1. سیر تکامل مدل‌های فضای برداری

اندیشه استفاده از مدل‌های فضای برداری برای بازنمون اطلاعات از زمان ظهور رایانه وجود داشته است. طی این سال‌ها، راه حل‌های زیادی جهت بهبود عملکرد بازیابی این مدل‌ها پیشنهاد شده است که در ذیل به پاره‌ای از این کوشش‌ها اشاره می‌شود:

اچ. پی. لون³ و خاستگاه مدل‌های فضای برداری

در سال 1953، «اچ. پی. لون» مباحث ابتدایی استفاده از مدل‌های فضای برداری را در بازیابی اطلاعات منتشر کرد که چکیده‌ای از موضوعات و مفاهیم کلیدی ست و هنوز مورد توجه قرار می‌گیرند (Luhn 1953). لون به این موضوع پرداخت که واژگان کنترل شده و طرح‌های رده‌بندی

¹ Euclidean

² relevance feedback

³ H. P. Luhn

مورد استفاده در نمایه سازی سنتی در خلال زمان تغییر می کنند و با طبقه بندی صرف مفاهیم موجود در یک مدرک که در در یک دوره زمانی مهم جلوه می نماید، از جنبه هایی از مدرک که ممکن است در آینده از اهمیت بیشتری برخوردار شود غفلت می شود.

بنابراین وی یک فضای مفهومی برای حل مشکل رده بندی و عدم یکپارچگی در کاربرد واژگان پیشنهاد کرد. در مدل پیشنهادی لون هر مدرک باید در یک فضای مفهومی¹ و به صورت مجموعه ای نشان داده شود که عناصر آن با مفاهیم گوناگون (اندیشه های مستقل از زبان) مجموعه مدارک متناظر است. برای کاوش مجموعه، مفاهیم موجود در پرسش، مانند مفاهیم موجود در مدرک رمز گذاری می شوند و هر مجموعه با مجموعه های دیگر موجود در فضا مقایسه می شود.

گرچه مدل پیشنهادی لون فاقد بسیاری از جزئیات اجرایی بود اما مدل وی، بنیانی برای اقدامات و اصلاحات بعدی قرار گرفت. حداقل بخشی از نظام های بسیار پیچیده کنونی بر ایده های لون استوارند. از دیگر مزایای مدل «لون» می توان به این امر اشاره نمود که وی توجه صرف به بسامد و شمارش کلمات پربسامد را به عنوان شاخص اصلی نمایه سازی با چالش روبه رو ساخت و مفهوم را جایگزین کلمه ساخت.

شایان ذکر است بسیاری از مسائلی که حدود 50 سال پیش به وجود آمد (مثلا چگونگی فائق آمدن بر مشکلات هم معنایی و چند معنایی) هنوز به طور کامل حل نشده اند.

«بورکو» و «برنیک» در راه طبقه بندی مدارک بر اساس فضای کاهشی².

اندکی پس از آنکه ایده های «لون» منتشر شد، اچ. بورکو و ام. برنیک (Borko and bernick 1963) روشی ارائه کردند که از طریق آن، مدارک به صورت خود کار در مقوله های از پیش تعریف شده طبقه بندی می شدند.

گرچه طبقه بندی مدارک دارای اهدافی سوای اهداف بازیابی اطلاعات است اما رویکرد بورکو و برنیک به طبقه بندی مدارک می تواند به عنوان حالت خاصی از بازیابی اطلاعات مورد توجه قرار گیرد. نظیر همه روش های فضای برداری، بورکو و برنیک عقیده داشتند که اصطلاحات موجود در یک

¹ Notion-space

² Reduced-Space document classification

مدرک بیانگر مطلوبی از محتوای معنایی مدرک است. آن‌ها امیدوار بودند مدارکی که از نظر رده-بندی هم مقوله‌اند، دارای واژگان مشابهی باشند. در آزمون‌های آن‌ها دو دسته مدرک، شامل یک دسته آزمایشی و یک دسته گواه مورد استفاده قرار گرفت.

دسته آزمایشی برای اشتقاق مقوله‌های رده‌بندی و بدست آوردن تابعی جهت پیش‌بینی مقوله صحیح مدرک موردنظر بکار گرفته شد و دسته گواه برای آزمون اینکه فرمول پیش‌بینی شده به درستی مدارک را مقوله‌بندی می‌کرد یا خیر.

به منظور تعریف مقوله‌های رده‌بندی، بورکو و برنیک یک ماتریس اصطلاح-مدرک برای مجموعه مدرک مورد آزمون ساختند که از بسامد واژگان در هر مدرک به عنوان عناصر ماتریس استفاده می‌کرد، سپس ضرایب همبستگی هر اصطلاح را نسبت به همه اصطلاحات دیگر موجود در ماتریس به دست آوردند.

نتایج نشان داد که تقریباً نیمی از مدارک از طریق این نظام، به درستی مقوله‌بندی می‌شوند. نمایه-سازی معانی پنهان نیز درست‌نظیر نظام رده‌بندی بورکو و برنیک از یک فضای کاهشی برای بیان اصطلاحات و مدارک استفاده می‌کند. علی‌رغم روش بورکو و برنیک، نمایه‌سازی معانی پنهان از تجزیه مقدار ویژه¹ (SVD) که یک روش تحلیل عاملی دو راهه است برای کاهش ابعاد فضایی استفاده می‌کند.

بنابراین، به طور کلی می‌توان گفت که رده‌بندی مقوله‌ها توسط بورکو و برنیک به پیشبرد فرایند بازیابی و جستجوی اطلاعات کمک‌شایانی نمود. به گونه‌ای که می‌توان حتی آن را شکل ابتدایی مدل نمایه‌سازی معانی پنهان به شمار آورد. به نظر می‌رسد عیب این روش، طراحی رده‌ها و قرار دادن مدارک در این مقوله‌های رده‌بندی باشد.

سالتون و نظام بازیابی اطلاعات اسمارت (SMART):

نظام بازیابی اطلاعات اسمارت در سال 1964 بوسیله سالتون (Salton 1956) و همکارانش در دانشگاه هاروارد شروع شد و در سال 1968 در دانشگاه کرنل ادامه یافت. این نظام دارای بسیاری از تکنیک‌های بازیابی اطلاعات بود که امروزه در نظام‌های فضای برداری مدرن شاهد آن هستیم. سالتون

¹ Singular Value Decomposition (SVD)

و پژوهشگران دیگر از اسمارت برای مطالعه وزن‌دهی اصطلاح¹، بازخورد ربط، دسته‌بندی و مسئله مترادف‌ها استفاده کردند و استفاده از عبارت‌ها را برای بهبود عملکرد نظام بازیابی پیشنهاد کردند (Salton 1971).

اسمارت از تطابق نحوی عبارت‌ها برای شناسایی مفاهیم مشابه در مدارک استفاده می‌کرد و باعث می‌شد مدارکی بازیابی شوند که دارای مفاهیمی مشابه باشند. این نظام از تطابق آماری عبارت‌ها نیز جهت کشف هم‌آیندی² مفاهیم موجود در جملات مدد می‌گرفت (Salton and Lesk 1965).

سالتون با مقایسه اسمارت با مدلارز³ (یک نظام معمولی بازیابی اطلاعات) دریافت که نمایه استاندارد اسمارت بدتر از جستجوی استاندارد مدلارز عمل می‌کند، اما وی با حذف واژگان مشترک، واژگانی که فقط یک بار در مجموعه مدرک ظاهر می‌شوند، و واژگانی که همانندی درون مدرکی⁴ را در مجموعه مدرک افزایش می‌دهند به این نتیجه رسید که اسمارت تقریباً به خوبی مدلارز عمل می‌کند. سرانجام سالتون پی برد که با تدوین اصطلاح‌نامه‌ای برای یاری اسمارت و بکارگیری یک یا دو برهمکنش بازخورد ربط، مزیت نظام اسمارت نسبت به مدلارز، 18 تا 30 درصد افزایش می‌یابد. بنابراین، سالتون به این نتیجه رسید که واژه‌های پربسامد و کم‌بسامد را حذف نماید و به گونه‌ای می‌توان وی را از پیشگامان استفاده از فهرست حذفی به شمار آورد.

اسپارک جونز⁵ و وزن‌دهی عام⁶:

اسپارک جونز در سال 1972، استفاده از نظام‌های وزن‌دهی عام را برای بهبود عملکرد نظام‌های بازیابی اطلاعات، مورد آزمایش قرار داد. وی نشان داد اصطلاحاتی که به طور متناوب در یک مجموعه ظاهر می‌شوند با آنکه نتیجه پرسش‌هایی که شامل آن اصطلاحات هستند را منحرف می‌کنند اما برای بازیابی موثر اطلاعات ضروری هستند (Sparck Jones 1972).

به هر حال برای بهتر متمایز ساختن مدارک به هنگام تعیین مدرک مرتبط با یک پرسش، باید به اصطلاحاتی که متناوباً در مجموعه حضور نمی‌یابند نسبت به اصطلاحات که به طور مداوم در مجموعه حضور می‌یابند، ارزش بیشتری داده شود.

¹ Term-Weighting

² Co-occurrence

³ MEDLARS

⁴ Inter document

⁵ Sparck Jones

⁶ Global-Weighting

به عبارت دیگر، جونز نیز به حذف موارد پراستناد پرداخت و به واژه‌ها و اصطلاحاتی اهمیت داد که گاهی در مجموعه‌های مدارک حضور می‌یابند. از دیگر مزایای روش وی، وزن‌دهی اصطلاحات بود که کمک شایانی به بهبود نظام‌های نمایه سازی و بازیابی اطلاعات نموده است.

جونز از تابع $f(n) = m$ که $(2^{m-1} < n \leq 2^m)$ برای وزن‌دهی عام سه مجموعه استاندارد آزمایشی خود استفاده کرد. در مجموعه‌ای که دارای N مدرک است، اصطلاحی که n بار در مجموعه حضور می‌یابد وزن $f(N)-f(n)+1$ می‌گیرد.

در هر یک از سه مجموعه مدرک، بهبود عملکردی فراهم شده از طریق وزن‌دهی عام از هر پیشرفت دیگر نظام‌های بازیابی اطلاعات که در متون پژوهشی پیشین ارائه شده‌اند، پیشی گرفته است. وزن‌دهی عام به همراه وزن‌دهی خاص¹، حتی عملکرد بازیابی نظام‌های اطلاعاتی ساده را نیز افزایش داده است.

کُل² و مدل‌های فضای کاهشی در بازیابی اطلاعات:

کُل، در سال 1979 اندیشه فضای کاهشی اصطلاح - مدرک بورکو و برنیک را بسط داد و WEIRD را پدید آورد (Koll 1979).

سیستم WEIRD یک نظام بازیابی مدرک خودکار بود که در دانشگاه «سیراکوس»³ طراحی و پیاده‌سازی شد. این نظام تلاش می‌کرد تا هنر بازیابی رایانه‌ای را از حالت تطابق واژگانی به قضاوت درباره تشابه مفهومی واژه‌ها ارتقا دهد. این نظام از یک مدل فضای برداری برای بازنمایی روابط میان اصطلاحات و مدارک سود می‌جست. اقلام موجود در فضا (پیشامد) مطابق با معنای آن‌ها مکان‌یابی می‌شدند که همان مجاورت اقلام با همه اقلام دیگر در داده پایگاه است که از طریق شمارش همابندی واژگان صورت می‌پذیرفت (Koll 1979).

کُل از مجموعه آزمایشی اش هفت مدرک را به عنوان محورهای عمود برهم فضای اصطلاح - مدرک انتخاب کرد سپس هر یک از مدارک باقی مانده و واژگان مشترک حذف شده را مورد آزمون قرار داد و ریشه اصطلاحات را به منظور حذف پیشوندها و پسوندها بدست آورد. وی با بکارگیری اصطلاحات در مدارکی که قبلاً در فضای اصطلاح - مدرک قرار گرفته‌اند، مدارک را

¹ Local-Weighting

² Koll

³ Syracuse

موقتاً در موضع میانگین فاصله‌ای اصطلاحات آن‌ها قرار داد. سپس اصطلاحات باقی مانده در فضا دوباره مکان‌یابی شدند تا معانی آن‌ها در مدرک جدید منعکس شود. هنگامی که تمام مدارک پردازش شدند این نتیجه حاصل شد که به طور کلی، مدارک مربوط نسبت به مدارک نامرتب در جایگاه بالاتری قرار دادند. کل همچنین تعداد ابعاد لازم در جستجوی موثر را مورد بررسی قرار داد و شیوه‌هایی جهت انتخاب مدارک به عنوان محورهای فضای اصطلاح - مدرک پیشنهاد کرد ولی با این وجود وی در حل مسائل مربوط به مدل‌های فضای برداری به نتایج معینی نرسید (Koll 1979).

بنابراین مزیت مدل کل را می‌توان در استفاده از فضا-مکان در بازیابی و نمایه‌سازی اطلاعات جست و عیب آن را در گم شدن در مسائل و مشکلات مربوط به مدل‌های فضای برداری تصور کرد.

«سالتون» و «بوکلی» به سوی بازخورد ربط:

در سال 1990، سالتون و بوکلی فرایند بازخورد ربطی که در اواسط دهه 1960 معرفی شده و سالتون در نظام بازیابی اطلاعات اسمارت از آن استفاده کرده بود را مورد آزمون قرار دادند. آن‌ها با اعمال چندین روش بازخورد ربط روی پنج مجموعه آزمایشی استاندارد دریافتند هنگامی که بازخورد ربط برای اصلاح پرسش کاربر به کار می‌رود، عملکرد بازیابی به طور متوسط 90 درصد بهبود می‌یابد و در صورتی که بازخورد ربط چندین بار تکرار شود بهبود عملکرد تا 100 درصد نیز امکان پذیر است (Salton and Buckley 1990). سایر نتایج مهم آزمایش سالتون و بوکلی به قرار زیر است:

- 1- وزن‌دهی اصطلاحات موجود در مجموعه مدرک، کارآیی بازخورد ربط را افزایش می‌دهد.
- 2- استفاده تمام متن از مدارک بازخورد ربط (بجای بکارگیری صرف اصطلاحاتی که بیشترین رخداد را در مجموعه بازخورد دارند) در یک پرسش، موفقیت آن را بهبود می‌بخشد.
- 3- پرسش‌هایی که مبهم بیان شده‌اند (به عنوان مثال، فقط شامل چند واژه‌اند) نسبت به پرسش‌هایی که از اصطلاحات بیشتری استفاده می‌کنند، از بازخورد ربط، بیشتر بهره می‌برند.
- 4- مجموعه‌های مدرکی که در آغاز کار از نظر قابلیت جستجو مشکل هستند نسبت به آن‌هایی که به راحتی جستجو می‌شوند از بازخورد ربط سود بیشتری می‌برند.

5- مجموعه‌های تکنیکی و فنی نسبت به مجموعه‌های عام، با درخواست‌های بازخورد ربط، بهتر کنار می‌آیند زیرا در فضای اصطلاح - مدرک کوچک‌تری محصور می‌شوند و این امکان حاصل می‌شود که بازخورد ربط بهتری جهت تمایز مدارک صورت گیرد.

به عبارت دیگر، سالتون و بوکلی به نقش و اهمیت فرایند بازخورد ربط در بازیابی و نمایه سازی تأکید ویژه‌ای داشتند. آن‌ها همچنین، بسط یا محدودیت درخواست‌های جستجو را نیز بنیان نهاده‌اند. علاوه‌براین، آن‌ها ابهام اولیه در درخواست‌های جستجو را در ابتدای کار مشکل، اما با بهره‌گیری از بازخورد ربط، تسهیل می‌کردند.

دومایس (Dumais 1991) گزارش داد که بازخورد ربط، عملکرد نمایه‌سازی معانی پنهان را بین 33 تا 67 درصد بهبود می‌بخشد. بازخورد ربط به گونه‌ای که در مدل نمایه سازی معانی پنهان به کار می‌رود در ادامه مورد بحث و بررسی قرار می‌گیرد.

نمایه سازی معانی پنهان (نمپ)

نمایه سازی معانی پنهان (نمپ)¹ یک شیوه آماری بازیابی اطلاعات است و برای فائق آمدن بر دو اشکال رایج در بازیابی به وجود آمده است. اشکال اول، هم معنایی² است و به این موضوع اشاره دارد که برای یک شیء چندین نام وجود دارد. مثلاً واژه picture هم به معنای image و هم به معنای photograph به کار می‌رود. اشکال دوم، که چندمعنایی³ نام دارد دال بر این مطلب است که اکثر واژگان، بیش از یک معنا دارند (Hofmann 1999; Hong 2000). مثلاً جگوار⁴ هم به یک حیوان (شبه پلنگ)، هم به صفحه فرمان آتاری‌های بازی قدیمی، هم به یک اتوموبیل گران قیمت و هم به یک ماشین مجازی جاوا⁵ اطلاق می‌شود که در بخش علوم کامپیوتر دانشگاه «برکلی» ابداع شد (Hong 2000).

هم معنایی، با بسط واژه یا اصطلاح مورد تهاجم قرار گرفته است. این تکنیکها به بهبود جامعیت⁶ کمک می‌کند اما در مقایسه با چندمعنایی به کاهش دقت¹ منجر می‌شود. چون واژگانی که از بسط

¹ Latent Semantic Indexing (LSI)

² Synonymy

³ Polynomy

⁴ Jaguar

⁵ Virtual Java Machine

⁶ Recall

کلمه یا یک اصطلاح نامه تهیه می شود ممکن است بیش از یک معنا داشته باشد، بنابر این کاربرد اصطلاحات مورد نظر فقط جامعیت را بهبود می بخشد اما اسناد و مدارکی را نیز بازیابی می کند که نامربوط به پرسش اصلی است و باعث کاهش دقت می شود.

چند معنایی نشان داده است که از لحاظ نگارشی بسیار مشکل دارد چون با ابهام ذاتی زبان طبیعی روبه رو است. یک نگارش کلی این است که برای رفع ابهام اصطلاحات پرسش، از واژگان کنترل شده استفاده شود اما چون واژگان کنترل شده مستعد نابهنجاری اند و از نظر یادآوری مشکل هستند واضح است که این رویکرد دارای مشکل کاربردپذیری است. علاوه بر این ممکن است اصطلاحات و واژگان موجود در واژگان کنترل شده، در یک مدرک دلخواه نباشد و بدین معنی ست که مدرک مورد نظر هرگز با اصطلاح صحیحی از واژگان، نمایه نشده است. عده ای بر این عقیده اند که می توان از یک سلسله مراتب استفاده کرد اما این کار نیز بیش از اندازه دقیق و انعطاف ناپذیر است و به یقین، حالت هم معنایی به خود می گیرد. مثلاً ممکن است شخصی یک مدرک را در یک رشته یا شاخه موضوعی قرار دهد و دیگری همان مدرک را در شاخه ای دیگر قرار دهد.

در اقدامات اولیه روی «نمپ»، پژوهشگران به دنبال رویکردی نوین بودند که دارای ویژگی های

زیر باشد:

1. قدرت سازگاری با مجموعه ای خاص
2. توان بازنمایی روشن اصطلاحات و مدارک به منظور تسهیل امر بازیابی
3. قابلیت کنترل از نظر محاسباتی

بنابر این نمایه سازی معنایی نهان به منظور پاسخگویی به این مسائل به وجود آمد.

بینش اصلی در نمایه سازی معنایی این است که جدای از بازیابی مدارک از طریق اصطلاحات خاص، می توان مدارک را بر اساس مفاهیم موجود در مدرک بازیابی کرد. مثلاً مدرک می تواند از طریق یک پرسش و به علت شباهت مفهومی با مدرک دیگر بازیابی شود. یعنی مدرکی که از طریق یک پرسش بازیابی می شود مدرک دیگری را به ما می نمایاند که با مدرک موجود شباهت مفهومی دارد. بعد دیگر این بینش این است که هر مدرک و هر پرسش در یک فضای مفهومی با ابعاد کوچک تر ترسیم می شود و در این فضای مفهوم است که بسیاری از عملیات صورت می گیرد (Hong

¹ Precision

(2000). به طور خلاصه، «نمپ» از یک مدل جبری بازیابی مدرک استفاده می‌کند که روشی ماتریسی است و تجزیه مقدار ویژه (SVD) نام دارد. این شیوه در ادامه این مبحث به تفصیل مورد بررسی قرار می‌گیرد.

3-2. مدل‌های احتمالی

رویکرد احتمالی¹ به بازیابی اطلاعات نخستین بار در سال 1960م. توسط «مارون» و «کوهنز»² مطرح شد (مارون و کوهنز 1960). از آن زمان تا کنون، این شیوه بازیابی به روش‌های مختلف آزموده و به کار گرفته شده است که از آن جمله می‌توان به آثار «کوپر»³ (Cooper 1991, 1995)، «وان ریجسبرگن»⁴ (1979, 1977)، «کرافت»⁵ (1986)، «کرافت» و «هارتر»⁶ (1979)، «فاهر» و «بوکلی»⁷ (1991)، «هارتر»⁸ (1975a, b)، «کوپر»، «چان»⁹، و «گی»¹⁰ (1994)، «کواک»¹² (1995)، «روبرتسون»¹³ (1977, 1990)، «روبرتسون» و «بووی»¹⁴ (2000)، «والکر»¹⁵ (1994, 2000)، «روبرتسون»، «مارون»، و «کوپر» (1982)،

¹ Probabilistic Approach

² Maron and Kuhns

³ Cooper

⁴ van Rijsbergen

⁵ Croft

⁶ Harper

⁷ Fuhr and Buckley

⁸ Harter

⁹ Chen

¹⁰ Gey

¹¹ Hiemstra

¹² Kwok

¹³ Robertson

¹⁴ Bovey

¹⁵ Walker

«اسپارک جونز»¹، «والکر»، و «روبرتسون» (1998, 2000a, b)، «آماتی» و «کرسطانی»² (1999)، «بوکشتاین»³ (1983a, b, 1985)، و «فاهر» (1989a, b, 1990) اشاره کرد.

به منظور آشنایی با مفاهیم اساسی مدل‌های احتمالی بازیابی احتمالات، تعریف دقیق پاره‌ای مفاهیم ضروری می‌نماید. این مفاهیم از اجزای جدایی ناپذیر مدل مفهومی (مدل فاهر) هستند که در ادامه به آن اشاره می‌شود.

2-3-1. فضای پیشامد (رخداد)⁴

به‌طور کلی، مدل‌های احتمالی⁵، مجموعه $Q \times D$ را به عنوان فضای پیشامد در نظر می‌گیرند، که Q که مجموعه همه درخواست‌های⁶ محتمل، و D همه مدارک⁷ مجموعه را می‌نمایاند. تفاوت‌های میان مدل‌های مختلف، در استفاده آنان از بازنمون‌ها⁸ و توصیف‌های مختلف برای درخواست‌ها و مدارک ریشه دارد. در اکثر مدل‌ها، درخواست و مدرک با استفاده از توصیفگر⁹ نشان داده می‌شود. این توصیفگرها به صورت بردارهایی با ارزش دودویی¹⁰ بیان می‌شوند که هر آرایه¹¹ به یک اصطلاح متناظر است. بیشتر مدل‌های پیچیده، یا از بردارهای دارای ارزش واقعی استفاده می‌کنند و یا به روابط میان اصطلاحات و مدارک می‌پردازند (Crestani et al. 1998).

هر درخواست بیانگر یک نیاز اطلاعاتی می‌باشد. در مدل‌های احتمالی هر درخواست، پیشامدی بی‌بدیل است، مثلاً؛ اگر دو کاربر درخواست واحدی را داشته باشند و یا اگر درخواستی واحد بوسیله یک کاربر و در دو موقعیت متفاوت ابراز شود، این دو درخواست متفاوت از یکدیگر انگاشته می‌شوند. درخواست کاربر به نظام اطلاعاتی فرستاده می‌شود که هدف این نظام، یافتن اطلاعات مرتبط با نیاز اطلاعاتی موجود در درخواست است.

¹ Sparck Jones

² Amati and Crestani

³ Bookstein

⁴ Event Space

⁵ Probabilistic Models

⁶ Query

⁷ Document

⁸ Representation

⁹ Descriptor

¹⁰ Real-Valued Vectors

¹¹ Array

مدرک، هر چیزی است که حامل اطلاعات باشد: یک قطعه متن، یک تصویر، صدا یا ویدئو. اکثریت قریب به اتفاق نظام‌های کنونی بازیابی اطلاعات، تنها با متن سروکار دارند، و می‌توان علل این امر را در مشکلاتی جستجو کرد که در بازنمون‌های کنونی موارد غیرمتنی وجود دارد.

2-4. مفروضه‌های مدل‌های بازیابی اطلاعات

مفروضه‌هایی که در تمام مدل‌های بازیابی اطلاعات مشترک است به قرار زیرند (Crestani et al. 1998):

1. دریافت کاربران از نیازهای اطلاعاتی خود، که در هر جلسه جستجو متغیر است، پیرو اصلاح و بهبود مستمر است و از طریق درخواست‌های مختلفی بیان می‌شود؛
2. بازیابی اطلاعات خواسته شده بر مبنای بازنمون‌های درخواست‌ها و مدارک صورت می‌گیرد نه بر اساس خود درخواست‌ها و مدارک؛
3. بازنمایی «شی»¹ها در نظام‌های بازیابی اطلاعات دارای قطعیت² نیست، مثلاً؛ استخراج اصطلاحات نمایه‌ای از یک مدرک یا درخواست جهت بازنمایی محتوای اطلاعاتی آن، فرایندی است کاملاً همراه با عدم قطعیت³ و نامطمئن که در نتیجه، فرایند بازیابی نیز دارای عدم قطعیت می‌شود. همین مفروضه سوم است که منجر به مطالعه مدل‌های احتمالی بازیابی می‌شود. بنابراین بهتر آن است که چنین نتیجه بگیریم که «نظریه احتمال صرفاً سروکار داشتن با عدم قطعیت است» و مشاهده عدم قطعیت در مدل‌های احتمالی چندان تعجب‌آور نیست.

2-5. مدل مفهومی⁴ بازیابی احتمالی اطلاعات

به اهمیت مدل‌سازی مفاهیم در حوزه‌هایی مانند نظام‌های مدیریت پایگاه داده⁵ و نظام‌های اطلاعاتی¹ توجه بسیار زیادی شده است. «فاهر» در سال 1989م. مدلی را پیشنهاد کرد که هم ساده و

¹ Objects

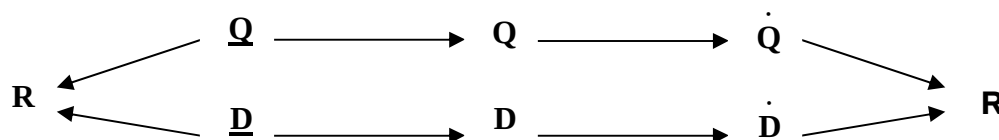
² Certainty

³ Uncertain

⁴ Conceptual Model

⁵ Database Management Systems

هم کلی است و می‌تواند مبنای مفهومی تمام مدل‌های احتمالی قرار گیرد که ما بدان‌ها اشاره خواهیم کرد گرچه برخی از این مدل‌ها به قبل از وی برمی‌گردد.



شکل 2-1. مدل مفهومی «فاهر» (Fuhr 1989)

همان گونه که در «مدل فاهر» می‌بینیم (شکل یک)، اجزای اساسی یک نظام بازیابی اطلاعات عبارتند از مجموعه معینی از مدارک «D» (مانند کتاب‌ها، مقالات، و یا تصاویر)، و مجموعه معینی از درخواست‌ها «Q» (مانند نیازهای اطلاعاتی). توجه داشته باشید که ما به مجموعه‌ای از درخواست‌ها و نه به درخواست واحدی اشاره داریم، زیرا؛ یک کاربر ممکن است نیازهای اطلاعاتی متنوعی داشته باشد. اگر ما R را مجموعه معینی از قضاوت‌های احتمالی ربط² بدانیم، به عنوان مثال؛ در موارد دودویی³ که $R=R, \dot{R}$ (یعنی یک مدرک ممکن است مربوط به یک درخواست باشد یا نباشد)، آنگاه وظیفه نظام بازیابی اطلاعات، بازنمایاندن هر زوج «مدرک - درخواست» بوسیله یک عضو R است.

متأسفانه، نظام‌های بازیابی اطلاعات به طور مستقیم با درخواست‌ها و مدارک سروکار ندارند، بلکه با بازنمون‌های آن‌ها در ارتباطاند (مثلاً بازنمایی یک مدرک از طریق یک متن، یا یک درخواست از طریق یک بیان بولی⁴). بنابراین، همان‌طور که قبلاً نیز گفتیم نوع روش بازنمایی بکار گرفته در یک مدل بازیابی اطلاعات، آن را از دیگری متمایز می‌کند. پس، در مدل بالا بازنمون‌های Q و D، بر مجموعه‌های $\dot{Q}$ و $\dot{D}$ نگاشته می‌شوند (Fuhr 1989).

¹ Information Systems

² Probabilistic Relevance Judgments

³ Binary

⁴ Boolean Statement

2-6. ربط¹ و احتمال ربط²

«ربط»، یکی از مفاهیم چالش برانگیز و اساسی نظام بازیابی اطلاعات است. در هیچ یک از منابع مطالعه، تعریفی دقیق و جامع از ربط ارائه نشده است و پژوهشگران از ارائه تعریفی رسمی از مفهوم ربط اجتناب ورزیده‌اند. گرچه تلاش‌های بدین منظور صورت گرفته (Saracevic 1970, Cooper 1971, Mizzaro 1996)، اما بر تعریف واحدی اتفاق نظر نشده است. سوای همه مطالب پیش گفته، می‌توان تعریف زیر را در همه منابع پیشین مشترک دانست:

«ربط به ارتباطی اطلاق می‌شود که بین یک مدرک و یک کاربر نظام اطلاعاتی برقرار می‌شود و یا نمی‌شود. به این معنا که اگر کاربر، مدرک بازیابی شده را مطالبه کند می‌گوییم این ارتباط برقرار شده است و اگر عکس این قضیه اتفاق افتد می‌گوییم فرایند ربط صورت پذیرفته است و یا مدارک بازیابی شده خورند نیاز کاربر نبوده‌اند.»

با توجه به مدل پیشین (مدل فاهر)، ربط (R)، رابطه بین یک مدرک (D) و نیاز اطلاعاتی کاربر (Q) است. اگر کاربر مدرک مربوط به نیازش را بخواهد آنگاه می‌گوییم مدرک مربوط است.

دو اصطلاح دیگر در این زمینه وجود دارند که بعضی مواقع شبهه برانگیزند: مدرک مربوط³ و مدرک مناسب⁴. یک مدرک ممکن است مربوط باشد اما مناسب نباشد، یعنی؛ توسط کاربر نهایی انتخاب نشود. بنابراین ملاک مربوط یا مناسب بودن، انتخاب نهایی مدرک توسط کاربر است.

2-7. اصل رتبه بندی احتمال⁵

یکی از ویژگی‌های مشترک کلیه مدل‌های احتمالی که در بازیابی اطلاعات نمود و توسعه یافته‌اند، پشتیبانی آن‌ها از حقانیت نظری این مدل‌هاست که در قالب «اصل رتبه بندی احتمال» بیان می‌شود (Robertson 1977).

¹ Relevance

² Probability of Relevance

³ Relevant

⁴ Pertinent

⁵ Probability Ranking Principle

اصل رتبه‌بندی احتمال بر این عقیده است که عملکرد بهینه¹ بازیابی هنگامی صورت می‌گیرد که مدارک، بر مبنای احتمال مرتبط بودن آن‌ها با یک درخواست، رتبه‌بندی شوند. این احتمالات باید در نهایت دقت و بر اساس کلیه داده‌هایی صورت گیرد که بدین منظور فراهم شده است.

این اصل، بازیابی بهینه را از بازیابی کامل مجزا می‌داند. بازیابی بهینه² دقیقاً برای بازیابی احتمالی اطلاعات تعریف می‌شود، زیرا؛ این نوع بازیابی با توجه به بازنمون‌ها (توصیفات) مدارک و نیازهای اطلاعاتی محقق می‌شود. از سوی دیگر، بازیابی کامل³ با خود عناصر نظام‌های بازیابی اطلاعات مانند مدارک و نیازهای اطلاعاتی در ارتباط است.

تعریف پذیرفته شده اصل رتبه‌بندی احتمال بدین گونه است:

اگر C را هزینه بازیابی یک مدرک مربوط، و $\dot{C}$ را هزینه بازیابی یک مدرک نامربوط بنامیم، «قاعده تصمیم»⁴، که مبنای اصل رتبه‌بندی احتمال است، بیان می‌کند که تنها زمانی یک مدرک (d_m) باید بر اساس درخواست کاربر (q_k) بازیابی شود که برتر از هر مدرک دیگر موجود در مجموعه (d_i) است. بنابراین،

$$C.P(R|q_k, d_m) + \dot{C}.(1-P(R|q_k, d_m)) \leq P(R|q_k, d_i) + \dot{C}.(1-P(R|q_k, d_i))$$

شایان ذکر است، کاربرد اصل رتبه‌بندی احتمال در مدل‌های احتمالی بازیابی اطلاعات متضمن مفروضه‌های زیر است:

1. به طور کلی، روابط بین مدارک نادیده انگاشته می‌شوند و مدارک به تنهایی مورد توجه قرار می‌گیرند. بنابراین، ربط یک مدرک با یک درخواست، مستقل از ربط مدارک دیگر موجود در مجموعه است.

¹ Optimal Performance

² Optimal Retrieval

³ Perfect Retrieval

⁴ Decision Rule

2. فرض می‌شود که مثلاً، احتمال $P(R|q_k, d_i)$ در تابع تصمیم به بهترین روش ممکن تخمین زده می‌شود و برای برآورد کردن قضاوت ربط واقعی کاربر، و در نتیجه مرتب‌سازی مدارک کافی است.

البته باید خاطر نشان شود که استراتژی‌های بازیابی دیگری وجود دارند که عملکرد بالایی دارند و با اصل رتبه‌بندی احتمال سازگار نیستند. به عنوان مثال، چنین استراتژی‌هایی را می‌توان در مدل‌های بولی و خوشه‌ای¹ مشاهده کرد که این مدل‌ها ماهیتاً احتمالی نیستند (Robertson 1977). با توجه به متون پژوهشی (که در مقدمه به آن‌ها اشاره شد) در حوزه بازیابی اطلاعات ما با دو نوع کلی از مدل‌های احتمالی بازیابی اطلاعات بر می‌خوریم که عبارتند از: مدل‌های ربط² و مدل‌های استنباطی³. در ادامه به این دو نوع مدل احتمالی می‌پردازیم.

2-8. مدل‌های احتمالی ربط

مدل‌های ربط، بر شواهدی استوار است که دربارهٔ مرتبط بودن یک مدرک با یک درخواست معین است. بنابراین، وظیفهٔ اصلی نظام‌های بازیابی اطلاعات در مدل‌های ربط، ارزیابی احتمال مربوط بودن یک مدرک است. این کار از طریق برآورد احتمال $P(R|q_k, d_i)$ برای هر مدرک d_i در مجموعه صورت می‌گیرد که امری بسیار مشکل است و تنها با ساده‌سازی مفروضه‌ها امکان‌پذیر است. برای فایق آمدن بر این مشکل، دو رویکرد وجود دارد، رویکرد مدل‌گرا و رویکرد توصیف‌گرا (Crestani et al. 1998).

رویکرد مدل‌گرا⁴: رویکردهای مدل‌گرا بر استقلال مفروضه‌های احتمالی⁵ استوارند و به عناصر بکاررفته در بازنمایی مدارک یا درخواست‌ها اشاره دارند. بنابراین، ابتدا احتمالات این «عناصر ویژه

¹ Cluster Models

² Relevance Models

³ Inference Models

⁴ Model-Oriented Approach

⁵ Probabilistic Independence Assumptions

بازنمایی¹ برآورد می‌شوند و با استفاده از مستقل بودن مفروضه‌ها، احتمال بازنمایی‌های مدارک از طریق آن‌ها تخمین زده می‌شود. نمایه‌سازی دودویی مستقل²، مدل‌های بازیابی، و «مدل پواسن مرتبه³ ان» نمونه‌هایی از این رویکردها به بازیابی هستند.

رویکرد توصیف‌گرا⁴: رویکردهای توصیف‌گرا، از ماهیت اکتشافی⁵ بیشتری برخوردارند. با در نظر گرفتن بازنمایی درخواست‌ها و مدارک، مجموعه‌ای از ویژگی‌ها برای زوج‌های «درخواست - مدرک» تعریف می‌شود که به هر زوج امکان می‌دهد بر این ویژگی‌ها نگاشته شوند. سپس با بکارگیری تعدادی داده آموزشی، که شامل زوج‌های «درخواست - مدرک» و قضاوت‌های ربط مربوط به آن‌هاست، احتمال ربط با در نظر گرفتن این ویژگی‌ها تخمین زده می‌شود. بهترین نمونه‌ای که از این رویکرد استفاده کرده، «مدل نمایه‌سازی دارمشتاد»⁶ است.

نمونه‌ای از مدل‌های احتمالی ربط، مدلی است که در سال 1979 م. توسط «وان ریجسبرگن» ارائه شد. تمرکز مدل وی که به «مدل‌سازی احتمالی به مثابه یک استراتژی تصمیم»⁷ معروف است، بر «تابع مخاطره یا شکست»⁸ است (van Rijsbergen 1979). در این مدل، و مدل «فاهر» که قبلاً به آن اشاره شد، شیوه‌های بازنمایی و توصیف درخواست‌ها و مدارک، یکسان تلقی می‌شوند. درخواست‌ها و مدارک با مجموعه‌ای از اصطلاحات نمایه‌ای توصیف می‌گردند. $t = \{t_1, t_2, t_3, \dots, t_n\}$ مجموعه اصطلاحات بکاررفته در مجموعه مدارک را نشان می‌دهد. درخواست q_k با اصطلاحات متعلق به مجموعه t نشان داده می‌شود و به همین ترتیب، مدرک d_j نیز مجموعه‌ای از اصطلاحات فرض می‌شود که در آن ظاهر می‌شود. با یک نمایش دودویی، d_j به صورت بردار دودویی $X = (X_1, X_2, X_3, \dots, X_n)$ بیان می‌شود که اگر $t_i \in d_j$ باشد $X_i = 1$ است در غیر این صورت $X_i = 0$ می‌شود.

تمام مدل‌های احتمالی ربط بر این فرض استوارند که توزیع اصطلاحات در مجموعه مدارک، اطلاعاتی راجع به ربط یک مدرک با درخواستی معین فراهم می‌کند، زیرا؛ فرض می‌شود توزیع

¹ Individual Representation Elements

² Binary Independence Indexing

³ N-Poisson Model

⁴ Description-Oriented Approach

⁵ Heuristic Nature

⁶ Darmstadt Indexing Model

⁷ Probabilistic Modeling as a decision strategy

⁸ Risk or Loss Function

اصطلاحات در مدارک مربوط و نامربوط متفاوت از یکدیگرند. این فرض، به «فرضیه خوشه‌ای»¹ مشهور است.

اگر توزیع واژگان در مجموعه مدارک مربوط و نامربوط یکسان باشد امکان ایجاد یک ضابطه تشخیص بین آن‌ها وجود ندارد و در این صورت بازنمایی‌های متفاوتی از محتوای اطلاعاتی مدارک ضروری به نظر می‌رسد. توزیع (پراکندگی) اصطلاحات، اطلاعاتی درباره احتمال ربط یک مدرک با یک درخواست در اختیار می‌گذارد. اگر قایل به قضاوت دودویی ربط باشیم توزیع اصطلاحات، اطلاعاتی درباره احتمال $P(R|qk, di)$ فراهم می‌کند.

از آنجایی که نمی‌توان کمیت $P(R|qk, X)$ را، که X نمایش دودویی مدرک است مستقیماً اندازه‌گیری کرد، به جای آن از «قضیه بیز» استفاده می‌شود (Pearl 1988):

$$P(R|qk, X) = [P(R|qk) \cdot P(X|R, qk)] / P(X|qk)$$

برای آنکه بدانیم این فرمول چگونه بدست آمده است از قانون احتمال شرطی آغاز می‌کنیم. پس داریم:

$$P(a, b) = P(a|b)P(b)$$

حال با جای‌گذاری m به جای b ، و $m=b, c$ داریم:

$$P(a, m) = P(a|m)P(m) = P(a|b, c)P(b, c) = P(a|b, c)P(b|c)P(c)$$

پس داریم:

$$P(a, b, c) = P(a|b, c)P(b|c)P(c)$$

حال با جای‌گذاری X, R, qk به جای a, b, c داریم:

$$P(X, R, Qk) = P(X|R, qk)P(R|qk)P(qk) \rightarrow P(R|X, qk) = [P(X|R, qk)P(R|qk)] / P(X|qk)$$

مدل‌های احتمالی ربط که در بازیابی اطلاعات به کار رفته‌اند بسیار زیاد می‌باشند. اما آنچه که از متون پژوهشی این حوزه بر می‌آید بیانگر آن است که مدل‌های احتمالی بسیار متنوع هستند و هر پژوهشگر با توجه به رویکردی که به مسأله بازیابی دارد یکی از این مدل‌ها را به کار می‌برد.

برخی از معمول‌ترین مدل‌های احتمالی ربط عبارتند از: مدل‌سازی احتمالی به مثابه یک استراتژی تصمیم، مدل بازیابی دودویی مستقل²، مدل نمایه‌سازی دودویی مستقل¹، مدل نمایه‌سازی «دارمشتاد»¹

¹ Cluster Hypothesis

² Binary Independence Retrieval Model

بازیابی با مدل نمایه‌سازی احتمالی²، مدل استنتاجی احتمال³، و مدل نمایه‌سازی «پواسن مرتبه‌ان»⁴، که پرداختن به هر یک از این مدل‌ها با بسیاری از مباحث پیچیده ریاضی و احتمالات در ارتباط است و بسیار خسته‌کننده می‌نماید و در چارچوب بحث حاضر نیست.

2-9. مدل‌های استنباطی عدم قطعیت⁵

این مدل‌ها بر این اندیشه استوارند که بازیابی اطلاعات، یک فرایند استنباطی عدم قطعیت است. مدل‌های استنباطی عدم قطعیت در مقایسه با مدل‌های ربط، بر شکل‌های پیچیده‌تری از ربط استوارند، که اصولاً با تخمین‌های آماری احتمال ربط مدارک صورت می‌گیرد. در مدل‌های استنباطی عدم قطعیت، اطلاعاتی که در فرمول‌بندی درخواست مورد استفاده قرار نگرفته، ممکن است در ارزیابی ربط یک مدرک بکار رود. پاره‌ای از این اطلاعات عبارتند از اطلاعات مربوط به کاربران، بازخورد ربط کاربران، و دانش موضوعی. ویژگی دیگر این مدل‌ها آن است که مدل‌های استنباطی برخلاف مدل‌های ربط، چندان به مجموعه وابسته نیستند. پارامترها در مدل‌های ربط فقط برای مجموعه جاری⁶ معتبر است درحالی‌که، مدل‌های استنباطی می‌تواند از دانش مربوط به کاربران، در سایر مجموعه‌ها نیز استفاده کند. یعنی بحث تعمیم در این مدل‌ها امکان‌پذیر است. بطور کلی ما با دو دسته از مدل‌های استنباطی عدم قطعیت روبرو هستیم: دسته‌ای که بر «منطق غیر کلاسیک»⁷ استوار است و احتمالات بر اساس آن برآورد می‌شود، و دسته دوم که بر «استنباط‌های بیزین»⁸ استوار است.

2-10. مسائل کلی مطروح در مدل‌های احتمالی بازیابی اطلاعات

¹ Binary Independence Indexing Model

² Retrieval with Probabilistic Indexing Model

³ Probabilistic Inference Model

⁴ N-Poisson Indexing Model

⁵ Uncertain Inference Models

⁶ Current Collection

⁷ Non Classical Logic

⁸ Bayesian Inferences

مدل‌های احتمالی بازیابی اطلاعات، با سه مسئله اساسی روبرو هستند که عبارتند از تخمین پارامتر، بسط درخواست، و بازنمایی مدارک و درخواست‌ها.

1. تخمین پارامتر¹: یعنی اینکه هر مدل احتمالی بازیابی اطلاعات قبل از آنکه به کار گرفته شود باید قادر به تخمین پارامترهای معینی باشد.

2. بسط درخواست²: منظور این است که مشخص شود چه اصطلاحاتی باید در فرمول‌بندی درخواست در نظر گرفته شوند.

3. بازنمایی مدارک و درخواست‌ها: در این باره حرف و حدیث بسیار است اما به طور کلی بر سه مسئله تأکید بیشتری شده است:

§ عبارات³، نوع متفاوتی از اصطلاحات⁴ در مقایسه با واژگان⁵ است. بدین دلیل، ممکن است کاربرد «شماهای وزن‌دهی استاندارد»⁶ بکار رفته برای واژگان مناسب نباشد؛

§ هنگامی که عبارات و اجزای آنها به عنوان اصطلاحات درخواست به کار می‌روند این اصطلاحات بسیار به همدیگر وابسته‌اند و بنابراین، باید در آزمون مدارک مجموعه موردنظر، یک مدل وابستگی⁷ مورد استفاده گیرد.

§ مجموعه مدارکی که در آزمون‌ها مورد بررسی قرار می‌گیرد باید بسیار کوچک باشد تا بتوان هر مزیت حاصل از کاربرد عبارات را نمایاند. گرچه عبارات و واژگان، در مجموعه مدارک بزرگی نیز با موفقیت آزموده شده‌اند (Fuhr 1991).

به طور خلاصه در این فصل مروری بر مطالعه تاریخی مدل‌های برداری و احتمالی و مزایا و معایب آن‌ها صورت پذیرفت. خلاصه‌ای از مدل‌های یادشده در جدول زیر نمایانده شده‌اند.

مدل‌های بازیابی اطلاعات مطروحه تا کنون و کانون توجه آن‌ها

ردیف	پژوهشگر	راهکار پیشنهادی	عیب
1	اچ.پی. لون	• دگر دیسی در واژگان کنترل شده و طرح‌های رده‌بندی مورد استفاده در نمایه سازی	• عدم موفقیت در حل مشکل هم‌معنایی و چندمعنایی واژگان و اصطلاحات

¹ Parameter Estimation

² Query Expansion

³ Phrases

⁴ Terms

⁵ Words

⁶ Standard Weighting Schemes

⁷ Dependence Model

		• ارائه یک فضای مفهومی برای حل مشکل رده‌بندی و عدم یکپارچگی در کاربرد واژگان نمایه	نمایه‌ای
2	بورکو و برنیک	• طبقه‌بندی مدارک در مقوله‌های از پیش تعریف شده • اندیشه فضای کاهشی • مدارک هم‌مقوله دارای واژگان مشابه هستند.	• طراحی رده‌ها و مقوله‌های بنیانی برای جای‌دهی مدارک
3	سالتون	• آغاز نظام بازیابی اسمارت • وزن‌دهی اصطلاح • بازخورد ربط • دسته‌بندی و مقوله‌بندی • استفاده از عبارت‌ها در نظام بازیابی	• مشکلات تدوین اصطلاحنامه و فهرست حذفی
4	جونز	• توجه به واژه‌ها و اصطلاحات پربسامد ارزش‌دهی کمتر به آن‌ها • وزن‌دهی اصطلاحات	-
5	گل	• بسط اندیشه فضای کاهشی بورکو و برنیک و پدیدآوردن WEIRD • شاخص قضاوت مفهومی به جای تطابق واژگانی	• روبه‌رو شدن با مسائل کلی مربوط به مدل‌های فضای برداری
6	سالتون و بوکلی	• آزمون متوالی بازخورد ربط • وزن‌دهی باعث افزایش کارایی بازخورد ربط • رابطه مستقیم تعداد کلمات درخواست کاربر با استفاده از بازخورد ربط	-
7	نمایه‌سازی معانی پنهان	• استفاده از مدل جبری بازیابی مدرک • تکنیک تجزیه مقدار ویژه • شباهت مدرک با مدرک بازیابی شده دلیل بر شباهت مفهومی است.	• مشکل بودن تجزیه مقدار ویژه
8	مارون و کوهنر	• رویکرد احتمالی به بازیابی اطلاعات • بازیابی اطلاعات بر مبنای بازنمون درخواست نه خود درخواست • اصل عدم قطعیت	-

در فصل بعد به فرایند نمایه سازی ماشینی به طور اخص پرداخته می شود. همچنین، در فصل یاد شده، با تشریح تکنیک های مورد استفاده در نمایه سازی ماشینی جزئیات بیشتر این فرایند مدنظر قرار می گیرند.

۳. بنیہ سازی ماشین:

ویژگی ماو معاہم

3-1. نمایه‌سازی ماشینی

نمایه‌سازی ماشینی به گونه‌ای 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).

سابقه استفاده از رایانه در نمایه‌سازی به اوایل دهه 1950 میلادی (دایره‌المعارف آنالین کتابداری و اطلاع‌رسانی، زیر مدخل نمایه‌سازی) و یا به قول «کوئل» به اواسط قرن بیستم میلادی بر می‌گردد (Coyle 2008). «هیمویچ» و همکاران در طی انجام پروژه نمایه‌سازی ماشینی کتابخانه پزشکی «ولش» اذعان می‌دارند که ماشین قادر است نمایه‌های سریع‌تری تولید نماید اما نمایه‌های کامل نیاز به درک انسانی است (Himwich et al. 1955) که شاید هنوز هم مصداق داشته باشد. اما تحول عمده در این خصوص در دهه 1980 میلادی و همزمان با پیشرفت‌های فناوری اطلاعات صورت گرفت. این پیشرفت‌ها برای شرکت‌های بزرگ نمایه‌سازی این امکان را به‌وجود آورد که برنامه‌های پایگاه‌های اختصاصی را برای استفاده در نمایه‌سازی نشریات ارائه و گسترش دهند. نرم‌افزارهای ویژه نمایه‌سازی با آنچه در برنامه‌های واژه‌پرداز برای نمایه‌سازی تعبیه شد، کاملاً متفاوت است. این نرم‌افزارها می‌توانستند کارهایی نظیر الفبایی کردن، تنظیم تورفتگی و بیرون آمدگی مدخل‌ها و سایر امور مربوط به صفحه‌آرایی را انجام دهند (دایره‌المعارف آنالین کتابداری و اطلاع‌رسانی، زیر مدخل نمایه‌سازی).

نمایه‌سازی ماشینی معمولاً بر فنون تحلیل آماری و همابندی واژه‌ها در متون استوار است. روش‌های جدا کردن ریشه کلمات هم به این شیوه مهم مدد می‌رساند. گاهی برای شناسایی واژه‌های عبارتی که بهتر است با آن‌ها به عنوان یک واحد برخورد شود، از تقطیع معنایی متن استفاده می‌شود.

جنبه‌های معنایی هم می‌تواند با استفاده از مقابله واژه‌ها و عبارات با یک اصطلاح‌نامه جامع، به‌منظور تهیه واژگانی یکدست و برقراری پیوند بین واژگان مرتبط، به نمایه‌سازی ماشینی اضافه شود. اکنون مدت‌هاست که از همایندی واژه‌ها، لغات، عبارات، و استنادات کتابشناختی برای کمک به یافتن واژه‌ها و مدارک مربوط به هم استفاده می‌شود. از همین خاصیت در تهیه رابط‌های جست‌وجو در کتابخانه‌های دیجیتالی هم استفاده می‌شود. از همایندی استنادات کتابشناختی در تهیه نمایه‌های استنادی نیز استفاده می‌شود. این نمایه‌ها به پژوهشگران در ردیابی نویسندگان و مدارکی که به آثار مورد علاقه قبلی استناد کرده‌اند، کمک می‌کند. این همایندی می‌تواند برای خوشه‌بندی دسته‌هایی از مدارک که استناد یکسان دارند، به کار رود. همچنین از همایندی استنادی و شناخت گروه‌هایی از اسناد جدید که به یکدیگر استناد کرده‌اند، برای شناخت الگوهای پژوهشی در حال ظهور استفاده می‌شود (دایره‌المعارف اینترنتی کتابداری و اطلاع‌رسانی، زیر مدخل نمایه‌سازی).

البته از نظر سرعت و یکدستی، نمایه‌سازی ماشینی بسیار کارآمدتر از نمایه‌سازی انسان است، به گونه‌ای که هرگاه متون مورد نمایه‌سازی ماشین‌خوان باشند و پارامترهای انتخاب واژگان به نحو دقیق در برنامه گنجانده شده باشد، رایانه می‌تواند مقادیر عظیمی از متون را با سرعتی غیرقابل تصور اسکن کند و بدون داشتن خطا، یک واژه را هر چند بار که در متن آمده باشد، انتخاب و جای‌نمای مناسب را برای آن بگنجانند. اما تشخیص موارد اندکی از اشتباهات املائی، یا تفاوت‌های جزئی در املا، و یا نوشته شدن همان واژه با مقداری فاصله بین حروف برای ماشین دشوار است، کاری که برای انسان آسان است (دایره‌المعارف اینترنتی کتابداری و اطلاع‌رسانی، ذیل واژه نمایه‌سازی؛ Himwich et al. 1955).

3-2. فنون مشترک در نمایه‌سازی ماشینی

بیشتر روش‌های نمایه‌سازی ماشینی کنونی از اصطلاحات نمایه‌ای زبان طبیعی استفاده می‌نمایند. اصطلاح نمایه‌ای مستخرج در این روش‌ها، شامل کلمات ساده و عبارت‌های چندکلمه‌ای است و فرض آن است که آن‌ها محتوای مدرک را باز می‌نمایند. می‌توان اصطلاحات یاد شده را مستقیماً از عنوان، چکیده، و کل متن مدرک نیز استخراج کرد. «لون» (1957) نخستین کسی بود که پیشنهاد کرد

می‌توان کلمات معینی به طور خودکار از متن استخراج کرد که محتوای متن را بنمایاند (Luhn 1957).

به هر حال، همه کلمات موجود در مدرک اصطلاحات نمایه‌ای مناسبی نیستند و کلماتی که اصطلاحات نمایه‌ای مناسب قلمداد می‌گردند نیز در تعریف محتوای یک متن نقش هم‌اندازه‌ای ندارند. برخی از تکنیکها به شناسایی و وزن‌دهی اصطلاحات درست موجود در مدرک کمک می‌کنند. یک فرایند متداول انتخاب اصطلاحات نمایه‌ای زبان طبیعی از متون که همه محتوای آن را باز می‌نماید دارای مراحل زیر است (Salton 1989, 303):

1. شناسایی کلمات مفرد متن، که تحلیل واژگانی¹ نامیده می‌شود.
2. جابجایی یا حذف کلمات اضافی و اصطلاحاتی که به کرات در متن تکرار شده‌اند.
3. کاهش اختیاری کلمات باقی مانده به شکل ریشه‌ای آن‌ها، که فرایند ریشه‌گیری² نام دارد.
4. شکل‌دهی اختیاری عبارتها به عنوان اصطلاحات نمایه‌ای.
5. جانشینی اختیاری کلمات، ریشه‌های کلمات، یا عبارت‌های مربوط به آن‌ها.
6. محاسبه وزن هر کلمه یا ریشه آن، یا اصطلاحات موجود در رده مورد نظر در اصطلاح‌نامه، یا اصطلاح عبارتی.

ترتیب متفاوتی از مراحل یاد شده نیز امکان‌پذیر است. برای نمونه، می‌توان شناسایی عبارت‌ها را قبل از جابجایی کلمات اضافی یا فعل‌های کمکی انجام داد.

انتخاب اصطلاحات نمایه‌ای زبان طبیعی معمولاً به شیوه‌های فرعی یا غیرمستقیم ارزیابی می‌شود (Sparck Jones and Galliers 1996, 19). ارزیابی فرعی به قضاوت درباره کیفیت اصطلاحات نمایه بر اساس چگونگی عملکرد اصطلاحات نمایه‌ای در وظایف دیگر می‌پردازد. این ارزیابی فرعی معمولاً بر اساس چگونگی تأثیر بر کارآیی یا اثربخشی ارزیابی امکان‌پذیر می‌شود. اثربخشی ارزیابی بر حسب جامعیت³ و دقت (مانعیت) سنجیده می‌شود. و از دو فرمول زیر به دست می‌آیند:

$$\text{recall} = \frac{\text{number of relevant documents retrieved}}{\text{total number of relevant documents in the collection}}$$

و

¹ lexical analysis

² stemming

³ recall

$$\text{precision} = \frac{\text{number of relevant documents retrieved}}{\text{total number of documents retrieved from the collection}}$$

به عبارت دیگر:

- جامعیت برابر است با تعداد مدارک مرتبط بازیابی شده تقسیم بر تعداد کل مدارک مرتبط موجود در مجموعه.
- دقت یا مانعیت برابر است با تعداد مدارک مرتبط بازیابی شده تقسیم بر تعداد کل مدارک بازیابی شده از مجموعه.

3-2-1. تحلیل واژگانی

تحلیل واژگانی زمانی آغاز می‌شود که متن به طور الکترونیکی ذخیره شده است و می‌توان آن را به صورت یک سلسله حروف به حساب آورد. تحلیل واژگانی، فرایند تبدیل یک جریان ورودی از حروف به یک جریان کلمات یا اشتقاقات آن‌ها است (Fox 1992). یک کلمه یا اشتقاق آن به عنوان زنجیره‌ای از حروف تعریف می‌شود که با یک فاصله سفید یا / و علامت سجاوندی از هم مجزا می‌شوند.

تحلیل واژگانی، اصطلاحات نمایه‌ای نامزد را تولید می‌کند که می‌توان در مراحل بعدی آن‌ها را پردازش کرد و نهایتاً آن‌ها را به عنوان یک اصطلاح کلیدی انتخاب کرد. به نظر می‌رسد که بازشناسی کلمات مفرد فرایندی ساده باشد در حالی که همواره این گونه نیست. برای نمونه:

1. اختصارات (برای نمونه، «نک») برخی مواقع گیج کننده می‌شوند. برای نمونه اگر این «نک» در آخر جمله و جزیی از یک کلمه باشد، تمیز آن از اختصارات امری ساده نیست.
2. یک تصمیم متفاوت و وابسته به زبان درباره شکستن کلمات خط فاصله دار به کلمات تشکیل دهنده، یا نگهداری آن‌ها به عنوان یک اصطلاح نمایه‌ای مفرد لازم است. در زبان انگلیسی و نیز آلمانی، برخی از کلمات مرکب با خط میان آن‌ها نشان داده می‌شوند. جداسازی این کلمات مرکب در حالی که جامعیت سیستم را افزایش می‌دهند مانعیت را کاهش می‌دهند. همچنین، خط فاصله می‌تواند جزئی از اسامی خاص باشد (برای نمونه MS-DOS).

3. اعداد در متن‌ها معمولاً اصطلاحات نمایه‌ای مناسبی قلمداد نمی‌گردند و نادیده گرفته می‌شوند.

3-2-2. استفاده از فهرست حذفی

نمایه سازی رایانه‌ای عمدتاً به دوشیوه زیر انجام می‌شود: حفظ اصطلاح و حذف اصطلاح (دیانی 1379).

الف. حفظ اصطلاح:

این شیوه توسط دکتر «سوزان آرتاندی» در سال 1963 به کار گرفته شد. در نمایه سازی به شیوه حفظ اصطلاح یک تعریف عملیاتی از اصطلاحات نمایه‌ای لازم و ضروری است تا رایانه بتواند راهنمایی برای انتخاب موضوع‌های مدرک تهیه کند، علاوه به راین باید برنامه یا مجموعه برنامه‌های سازگار با این شیوه ساخته شود. در نظام حفظ اصطلاح سیاهه یا فهرستی به رایانه داده می‌شود که حاوی کلمات کلیدی است و به نظام گفته می‌شود، هریک از کلمات این لیست را که در متن بود، حفظ کند و آنرا به عنوان اصطلاح نمایه‌ای انتخاب کند. در این روش استفاده از اصطلاحات از پیش انتخاب شده راهنمای خوب و مفیدی است و امکان انتخاب واژه‌های استفاده شده به وسیله مولف مدرک را فراهم می‌کند. نقطه ضعف این روش این است که تضمینی وجود ندارد تا تمامی مفاهیم موضوعات جدید و مهم نمایه‌سازی شوند.

ب. حذف اصطلاح:

در شیوه حذف اصطلاح یک فهرست حذفی به رایانه داده می‌شود که کلمات فهرست را هرکجا که بود حذف کند و بقیه را به عنوان کلید واژه انتخاب نماید. نمایه سازی کوئیک و کواک از نوع حذف اصطلاح هستند. در این شیوه آنچه را که نباید کلیدواژه تلقی شود، حذف می‌شود (دیانی 1379).

بنابراین، یک فهرست حذفی یا واژه‌نامه منفی¹ فهرستی از واژه‌های ماشین خوان را دارا است (کلمات حذفی) که نمی‌توان آن‌ها را به عنوان اصطلاحات نمایه‌ای انتخاب کرد (Salton 1975, 30; Salton 1989, 279; Fox 1992). به طور معمول، یک فهرست حذفی برای حذف کلماتی استفاده

¹ negative dictionary

می‌شود که محتوای مدرک را منعکس نمی‌کنند. هنگامی که در مراحل آغازین نمایه سازی ماشینی از فهرست حذفی استفاده گردد، حذف واژگانی زائد باعث بهبود پردازش آتی کلمات نامزد نمایه می‌شود و همچنین فضای ذخیره اصطلاحات و کلمات را کاهش می‌دهد. حجم فهرست‌های حذفی یا بازدارنده معمولاً متفاوت است (برای مثال، بسیاری از فهرست‌های حذفی در زبان انگلیسی دارای حدود 50 تا 400 کلمه هستند).

تکنیک‌های مختلفی برای ساختن فهرست حذفی وجود دارد که برخی از آنها به شرح زیر است.

1. برخی از کلاس‌های کلمات¹ توصیفگرهای بهتری برای محتوای متن محسوب می‌شوند در حالی که دسته‌ای دیگر دارای کلمات کمکی و اضافی هستند که به اهداف دستوری خدمت می‌کنند و به اشیاء یا مفاهیم اشاره‌ای ندارند (مانند «از»، «و»، «با» و غیره). بسیار ضروری است که دسته دوم از نامزدی اصطلاحات نمایه‌ای حذف گردند.

2. مرسوم‌ترین شیوه ساخت یک فهرست حذفی آن است که کلماتی که بیشترین بسامد را در متون دارند در فهرست حذفی گنجانند (Luhn 1957; Salton 1989, 279). این شیوه بر این اساس استوار است که بسامد حضور یک کلمه کمکی² بسیار بیشتر از بسامد وقوع یک کلمه محتوایی³ است. چه آنکه فهرست حذفی با توجه به پربسامدترین کلمات یک مجموعه فرضی ایجاد شده باشد که گستره وسیعی از موضوعات را در بر می‌گیرد و به آن فهرست حذفی عام⁴ گفته می‌شود، یا اینکه با مشاهده بسامد واژگان موجود در مجموعه مدرک که باید نمایه سازی شود این فهرست را ایجاد کرد که به آن فهرست حذفی موضوعی⁵ گفته می‌شود، در هر دو حالت نیز یک ارزش آستانه برای تعداد کلماتی که باید در فهرست حذفی مد نظر قرار گیرند لحاظ می‌گردد. در مواردی نادر، کلماتی که دارای وزن بسامد کوچک در مدرک مقلوب هستند به عنوان کلمات حذفی در نظر گرفته می‌شوند. به هر حال، بسامد وقوع یک کلمه در یک مجموعه مدرک یک معیار صد درصدی برای نشان دادن اهمیت محتوا نیست و استثنائاتی وجود دارند. برای نمونه، گاهی زیرمجموعه‌ای از داده‌پایگاه مدرک دارای متون مختلفی است که به یک موضوع

¹ word classes

² function word

³ content word

⁴ generic stoplist

⁵ domain-specific stoplist

پرداخته‌اند. همچنین، امکان دارد یک داده‌پایگاهِ متن خاص دارای کلماتی باشند که به درد اصطلاحات نمایه‌ای نخورند اما نه در زبان استاندارد پربسامد هستند و نه در داده‌پایگاهِ مدرک اثری از آن‌ها دیده می‌شود.

3. چون کلمات وابسته معمولاً کوتاه (دارای تنها چند حرف) هستند، گاهی همه کلمات کوتاهی که عدد ارزش آستانه مربوط به حروف را دارا می‌باشند از متن حذف می‌گردند (Ballerini et al. 1997). در این حالت، استفاده از یک فهرستِ «ضدِ کلمات حذفی»¹، از حذف کلمات کوتاه مهم جلوگیری می‌نماید.

3-2-3. ریشه‌گیری کلمات

تکنیک دیگری که کیفیت نمایه‌سازی ماشینی را بهبود می‌بخشد فرایند ریشه‌گیری است. ریشه‌گیری همان فرایند تجزیه و کاهش اجزای نحوی کلمات به ریشه آن‌ها (برای نمونه، ترسیم شکل‌های مفرد و جمع کلمات به یک ریشه) است. برنامه‌ای که این کار را انجام می‌دهد ریشه‌گیر² نام دارد. فرض بر این است که کلماتی که ریشه مشابهی دارند از نظر معنایی با یکدیگر مرتبط هستند و معنای مشابهی برای استفاده‌کننده از متن دارند. ریشه‌گیری در زمینه بازیابی اطلاعات در پی این هدف است که تطابق بین اصطلاحات نمایه‌ای پرسش و متن مدرک را بهبود بخشد. شانس تطابق هنگامی افزایش می‌یابد که اصطلاحات نمایه‌ای به ریشه‌هایشان تجزیه شوند. بنابراین، ریشه‌گیری یک ابزار افزایش جامعیت است تا دامنه اصطلاح کلیدی را در جستجوی متن گسترده‌تر سازد (Salton 1986).

همچنین، ریشه‌گیری، از طریق نگاشتِ اجزای نحوی به شکل استاندارد، تعداد اصطلاحات نمایه‌ای را کاهش می‌دهد. در نتیجه، از طریق ریشه‌گیری، اندازه بازنمایی متن کاهش می‌یابد، که یک مزیت در بحث ذخیره‌سازی محسوب می‌گردد. چهار رویکرد عمده برای ریشه‌گیری وجود دارد:

¹ anti-stopword-list

² stemmer

1. جستجوی جدولی¹: روش جستجوی جدولی ساده‌ترین روش در ریشه‌گیری محسوب می‌شود و نیازمند آن است که اصطلاحات و ریشه‌های آن‌ها در یک جدول یا یک واژه‌نامه ماشین‌خوان ذخیره گردند (Fakes 1992). ریشه‌گیری از طریق جستجوهای انجام می‌شود که در جدول صورت می‌پذیرد. مزیت این روش آن است که نتایج ریشه‌گیری معمولاً درست هستند. به هر حال، جدول بزرگ‌تر می‌شود و جدول‌های بزرگ نیازمند فضاهای ذخیره‌سازی بیشتر و الگوریتم‌های جستجوی کارآمدتری هستند (برای نمونه، درخت جستجوی دودویی، جدول «هش»²).

2. الگوریتم‌های حذف «وند»³: الگوریتم‌های حذف «وند» به کرات استفاده می‌گردند و پیشوندها و/یا پسوندها را از اصطلاحات، کاسته و به ریشه منتج می‌گردند (Fakes 1992). ماشین ریشه‌گیر «لوینز»⁴ (1968) پسوندها را با استفاده از بزرگ‌ترین الگوریتم تطابق انجام می‌دهد. این الگوریتم طولیل‌ترین زنجیره ممکن از حروف را از یک کلمه و بر حسب مجموعه‌ای از قواعد می‌زداید. این فرایند آنقدر تکرار می‌شود که دیگر امکان حذف حروف دیگری از کلمه امکان‌پذیر نباشد. حتی بعد از آنکه همه حروف وابسته از کلمه جدا شدند، ریشه‌ها ممکن است نیاز به دگردیسی داشته باشند. بنابراین، از دانش زبان‌شناختی برای کدگذاری مجدد ریشه‌ها استفاده می‌گردد.

3. ریشه‌گیر جانشین حروف متغیر⁵: این ریشه‌گیرنده از طریق استفاده از واژکهای بدنه عظیمی از واژگان موجود فرایند یادگیری را دنبال می‌نماید (Hafer and Weiss 1974). به عبارت دیگر، این ریشه‌گیر از بسامدهای متوالی حروف در یک مجموعه متون به عنوان بنیانی برای ریشه‌گیری استفاده می‌کند. با محاسبه مجموعه «تغییرات جانشین»⁶ برای یک کلمه آزمایشی و اشاره کردن به رکوردها، می‌توان واژکهای یک کلمه را از طریق این سیستم استخراج کرد.

4. در نهایت، روش «ان گرام» اصطلاحات را بر مبنای تعداد «ان گرام»هایی که به طور مشترک دارند تحلیل می‌کند. یک «ان گرام» یک توالی از n حروف پی‌درپی است. «آدامسون» و «بورهام» (1974) تعدادی تطابق جفت – گرام‌ها را محاسبه کردند. یک جفت – گرام شامل یک جفت از

¹ table-lookup

² hash table

³ affix

⁴ Lovins

⁵ Letter Successor Variety Stemmers

⁶ successor varieties

کلمات متوالی است. برخی دیگر از پژوهشگران از سه گرام استفاده کرده‌اند (Xu and Croft 1998). اصطلاحاتی که قویاً با تعداد «ان گرام»های مشترک مربوط هستند در گروه‌هایی از کلمات مرتبط دسته‌بندی می‌شوند. اکتشاف نیز کمک می‌کند که شکل ریشه‌ای کلمه را بتوان پیدا کرد و یا اینکه از روش‌های خوشه‌بندی فضایی¹ استفاده می‌گردد. سیستم‌های ریشه‌گیری زیادی برای زبان انگلیسی توسعه یافته‌اند. دو سیستم بسیار رایج در زبان انگلیسی عبارت‌اند از سیستم ریشه‌گیر «لوینز»² و سیستم ریشه‌گیر «پورتر»³ (Fakes 1992). شایان ذکر است، ریشه‌گیری زمانی سودمند است که نظام ریشه‌شناسی⁴ زبان غنی باشد (Krovetz 1993).

3-2-4. گزینش عبارت‌ها و اصطلاحات و نرمال‌سازی اصطلاحات

واضح است که عبارت‌ها معانی نحوی بیشتری نسبت به کلمات منفرد در خود دارند. به ویژه، اعتقاد بر این است که عبارت‌های اسمی و اضافی اجزای محتوادر هستند و در نتیجه توصیفگرهای خوبی از محتوای متن محسوب می‌گردند (Earl 1970; Salton et al. 1990; Smeaton 1992). برای نمونه، اصطلاح «سرمایه‌گذاری مشترک»⁵ یکی از اصطلاحات مهم در متون تجاری است در حالی که هیچ یک از این دو اصطلاح به تنهایی مهم تلقی نمی‌شوند. عبارت‌ها باعث بهبود اخصیت⁶ زبان نمایه‌سازی می‌گردند. همچنین، استفاده از عبارت‌ها به عنوان اصطلاحات نمایه‌ای باعث افزایش مانعیت سیستم نمایه‌سازی می‌گردد (Fagon 1989). به علاوه، معانی عبارت‌ها ابهام کمتری نسبت به معانی کلمات منفرد متشکله آن‌ها دارد.

هنگامی که عبارت‌ها به عنوان اصطلاحات نمایه‌ای زبان طبیعی به کار می‌روند دو جنبه بایستی خودکارسازی گردد (Croft, Turtle, and Lewis 1991): شناسایی آن‌ها در متن و نرمال‌سازی آن‌ها به شکلی استاندارد. دو تکنیک عمده برای شناسایی عبارت‌ها وجود دارد: بازشناسی آماری عبارت و

¹ spatial cluster methods

² Lovins

³ Porter

⁴ Morphology

⁵ joint venture

⁶ specificity

بازشناسی نحوی عبارت. عبارت‌هایی که اشاره به یک مفهوم واحد دارند نیز به شیوه‌هایی متفاوتی بیان می‌گردند. بنابراین، نرمال‌سازی عبارت‌ها به یک شکل استاندارد بسیار ضروری است. یک مورد بسیار مهم از بازشناسی عبارت متوجه تشخیص اسامی خاص است.

بازشناسی آماری عبارت، فرض می‌کند که هنگامی که مجموعه‌ای از کلمات همراه با هم در متون یک مجموعه مدرک ظاهر می‌شوند، مجموعه کلمات هم‌رویداد به یک عبارت تعلق دارند. ایده استفاده از روابط آماری بین کلمات به اوایل دهه 1960 بر می‌گردد. اغلب، جفت‌های هم‌جوار که در فهرست حذفی وجود ندارند به عنوان عبارت‌های کاندیدای نمایه سازی مدنظر قرار می‌گیرند (Salton, et al. 1975). یک عبارت آماری به صورت بسامد وقوع عبارت‌ها، هم‌رویدادی اجزای عبارت‌ها، و / یا هم‌جواری¹ اجزای تشکیل دهنده عبارت‌ها در متن تعریف می‌گردد.

یک عبارت آماری ممکن است از طریق بسامد وقوع، هم‌رویدادی اجزای متشکله، و/یا هم‌جواری اجزای آن در متن انتخاب گردد. اما همیشه یک رابطه نحوی بین اجزای تشکیل دهنده عبارت وجود دارد. یک عبارت نحوی، بخش دستوری جمله است و بر اساس معیار زبان‌شناختی شناسایی می‌گردد.

استفاده از عبارت‌های نحوی بر این مفروضه استوار است که کلماتی از یک متن که یک رابطه نحوی دارند اغلب دارای یک رابطه معنایی وابسته نیز هستند (Smeaton and Sheridan 1991). بازشناسی نحوی عبارت بیش از چند دهه است که رواج یافته است. یکی از ساده‌ترین روش‌های نحوی به این صورت است که از یک واژه‌نامه ماشین‌خوان یا اصطلاح‌نامه استفاده می‌گردد که دارای اصطلاحات عبارتی از پیش کدگذاری شده و در شکل‌های نحوی گوناگون است. چنین واژه‌نامه‌هایی باید شیوه‌های بسیاری برای ترکیب کلمات منفرد برای نمایاندن یک مفهوم داشته باشند.

3-2-5. بازشناسی اسامی خاص

یکی از موارد مهم در بازشناسی عبارت‌های موجود در متن انتخاب اسامی یا نام‌های خاص است (Rau 1992; Jacobs 1993; Mani and MacMillan 1996; Strzalkowski et al. 1997). نمایه‌سازی با استفاده از اسامی خاص در بسیاری از نرم‌افزارهای بازیابی سودمند است. اسامی خاص

¹ proximity

اشاره به نام‌های اشخاص، شرکت‌ها، سازمان‌ها، نشان‌های تجاری، مکان‌ها، و واحدهای پولی دارد. دو روش عمده برای شناسایی این نام‌ها وجود دارد:

1. استفاده از واژه‌نامه‌ای ماشین‌خوان از نام‌ها که نیازمند پایگاه اطلاعاتی موجود از نام‌ها است.
2. از آنجایی که بسیاری از نام‌های خاص (برای مثال، نام شرکت‌ها) پدیدار می‌شوند، از بین می‌روند، یا تغییر می‌کنند، شناسایی دقیق آن‌ها نیازمند اسامی جدید است. همچنین، برخی از القاب یا پیشوندها مانند آقای، خانم، دکتر، و غیره نیز بایستی حذف گردند.

3-2-6. وزن‌دهی به اصطلاح نمایه

به طور معمول، یک اصطلاح یک واژه، ریشه کلمه، یا یک عبارت است. اگرچه اصطلاحات متعلق به یک کلاس کلی کلمات، نشان دهنده محتوا هستند، اما این اصطلاحات الزاماً با توجه به محتوای متن، اهمیت یکسانی ندارند. شاخص اهمیت یا وزن اصطلاح، مرتبط به هر اصطلاح نمایه‌ای است. وزن‌دهی به اصطلاح، برای انتخاب کلمات نمایه‌ای مناسب جهت بازنمایی متن یا بهبود تطابق پرسش و جامعیت در محیط بازیابی بسیار ضروری است (Buckley 1993). وزن‌دهی، دقت بازیابی را افزایش می‌دهد (Ro 1988). بسیاری از توابع وزن‌دهی توسط پژوهشگران حوزه ذخیره و بازیابی ارائه شده‌اند (Sparck Jones 1973; Salton 1975, 4; van Rijsbergen 1979, 24; Noreault et al. 1981; Salton and McGill 1983, 59; Ro 1988; Salton and Buckley 1988; Fuhr and Buckley 1991; Tenopir et al. 1991, 144-146).

به طور کلی، پارامترهای زیر در محاسبه وزن اصطلاح نمایه نقشی اساسی دارند:

1. خود اصطلاح نمایه: برای نمونه، کلاس نحوی اصطلاح.
2. متنی که باید نمایه گردد: پارامترهایی که متن را توصیف می‌کند: برای نمونه، طول متن و تعداد اصطلاحات مختلف موجود در متن.
3. رابطه بین اصطلاح نمایه و متنی که باید نمایه‌سازی گردد: برای نمونه، بسامد وقوع اصطلاح در متن، محل اصطلاح در متن، رابطه با دیگر اصطلاحات موجود در متن، و زمینه اصطلاح در متن.
4. رابطه بین اصطلاح نمایه و مجموعه مدرک (یا مرجع دیگر): برای نمونه، بسامد وقوع در این مجموعه.

بسیاری از توابع وزن‌دهی بر الگوهای توزیع اصطلاحات در متنی که باید نمایه‌سازی گردد و/یا در مجموعه مرجع تأکید می‌ورزند، و از آمارهایی برای محاسبه وزن‌ها استفاده می‌نمایند. فقط تعداد اندکی از وزن‌های تخصیص داده شده به اصطلاحات نمایه‌ای بر اساس دانش خبره در حوزه اهمیت اصطلاح صورت می‌پذیرد (Sparck Jones 1973). وزن اصطلاح نمایه‌ای معمولاً دارای یک ارزش عددی است وزن‌های اصطلاح دارای ارزش صفر یا بیشتر هستند و در مورد وزن‌های نرمال شده بین صفر و یک متغیر است. که وزن یک به اصطلاحات بسیار مهم و وزن صفر به اصطلاحات بسیار ضعیف تخصیص می‌یابد (Salton and Buckley 1988). ارزش صفر به این معنا نیز هست که اصطلاح مورد نظر دارای هیچ ارزش محتوایی نیست.

3-3. انواع نمایه‌سازی ماشینی

دو نوع بارز نمایه که در پژوهش‌ها تبلور زیادی یافته‌اند عبارت‌اند از:

3-3-1. نمایه سازی کوئیک¹

در نمایه سازی «کوئیک» که همان نمایه درون بافتی است تمام کلمه‌های موجود در عناوین مدارک را که توسط برنامه‌ای به کامپیوتر داده شده، با سیاهه‌ای از واژه‌های غیرمجاز (واژه‌های غیرموضوعی که در فهرست حذفی قرار می‌گیرد تا در ردیف الفبایی خود به عنوان سرشناسه قرار نگیرد) در برنامه کامپیوتر مطابقت داده می‌شود. کلمه‌هایی که در سیاهه واژه‌های غیرمجاز وجود ندارد، بطور خودکار کلیدواژه محسوب می‌شود. مدخلی که باهریک از این کلیدواژه هادرست می‌شود به صورتی است که این واژه را در درون بافت خود، یعنی همراه با بقیه کلمه‌های آن عنوان، در شکل و نظم طبیعی‌اش نشان می‌دهد. این بافت شامل همه واژه‌ها از جمله واژه‌های غیرمجاز نیز می‌شود و به این ترتیب گرچه واژه‌های غیرمجاز در نمایه نامه مدخل قرار نمی‌گیرد، اما همراه با سایر کلمات عنوان، در مدخل می‌آید تا محیط کلیدواژه و معنا و مفهوم آن را دقیق‌تر آشکار سازد و ارتباط میان کلیدواژه‌های عنوان را بهتر نمایش دهد. به این ترتیب بی‌درنگ دریافت می‌شود که عنوان مربوط با

¹ KWIC: KeyWord In Context

نیاز مراجعه کننده مناسب دارد. برون داد رایانه ای می تواند به صورت چاپی در دسترس مراجعان قرار گیرد (دیانی 1379).

هر خط نمایه درون بافتی معمولاً شامل سه بخش است: کلیدواژه یا مدخل، بافت (متن) و نشانه بازیابی (جای نما). نمایه واژه ها به صورت ستونی در مرکز یا سمت چپ صفحه ویاسمت راست آن (در فارسی) به ترتیب الفبایی می آید و با همین کلمه هاست که جوینده، جستجوی خود را در نمایه آغاز می کند. برای مشخص کردن پایان عنوان، در هر نمایه از علامت قراردادی خاصی نظیر علامت جمع یا مساوی و یا خط اریب (اسلش /) و غیره استفاده می شود تا جستجوگر بتواند به سهولت ابتدای عنوان را پیدا کرده و عنوان را در شکل طبیعی خود به سرعت مرور کند. از مزایای این نمایه سرعت و سهولت آن می باشد، در ضمن به متخصص نمایه سازی نیازی ندارد زیرا که با وجود فهرست حذفی این کار را کامپیوتر انجام می دهد. از طرفی کلمه کلیدی در متن عبارت مورد جستجو قرار دارد و این مسئله به جستجوگر کمک شایانی می کند (دیانی 1379).

3-3-2. نمایه کواک¹

نمایه «کواک» همان نمایه برون بافتی است. این نمایه برای رهایی از مسائل و مشکلات ناشی از ضرورت کوتاه کردن عنوان و نیز به منظور ساده ساختن خواندن مدخل ها طراحی شده است. در این نمایه هر کلیدواژه به ترتیب از عنوان خود خارج شده و مقدم بر سایر اجزای عنوان قرار می گیرد. سپس عنوان مدرک به ترتیب طبیعی خود و به طور کامل در زیر این واژه یا به دنبال آن می آید. به این ترتیب برای هر واژه مهم یک مدخل ساخته می شود. در این نمایه نیز کلمات مجاز در متن عنوان وجود دارد و کلید واژه قرار نمی گیرد (دیانی 1379).

23-4. نمایه سازی با کمک رایانه

در این نظام نخست نمایه ساز با دقت متن را می خواند، شناسه های موضوعی را انتخاب و علامت گذاری می کند و نمایه دستنویس تهیه می کند. آن گاه اپراتور مدخل ها را بر اساس نمایه دستنویس از

¹ KWOC: KeyWord Out Of Context

طریق صفحه کلید وارد رایانه می‌کند. بعد از ورود، رایانه مدخل‌ها را بر اساس نظم خاصی مرتب می‌کند، توصیفگرها را زیر شناسه‌ها می‌برد، جای‌نماهای هر مدخل را با هم ادغام می‌کند و بعد یک نمایه الفبایی ارائه می‌دهد. در این برنامه هر اصطلاح نمایه‌ای همراه جای‌نماهای آن وارد می‌شود و اگر توصیفگر داشته باشد، پس از شناسه این توصیفگر با یک ممیز آورده می‌شود. رایانه نیز پس از بررسی مدخل‌ها بر اساس برنامه از پیش طراحی شده آن مداخل را تنظیم می‌کند. این نوع نمایه سازی با دوروش متفاوت انجام می‌شود (نوروزی 1380) که به شرح زیر است.

1- نمایه سازی در محیط گرافیکی (نمایه سازی درون کاشتی). امروزه بیشتر نرم افزارها در محیط‌های دارای رابط گرافیکی¹ کار می‌کنند. کار با این روش به صورت زیر است:

نمایه ساز یا ویراستار، پس از اتمام صفحه آرایی و مشخص شدن شماره صفحه‌های کتاب، صفحه‌ها را از صفحه نمایش کامپیوتر مرور می‌کند و کلیدواژه‌ها یا کلمات و عبارت‌های مهم را انتخاب می‌کند. این عبارت انتخاب شده از نظر سیستم یک کلیدواژه محسوب می‌گردد. نمایه ساز همزمان با انتخاب کردن این کلید واژه‌ها باید یک پنجره واژه پرداز مستقل دیگر را باز می‌کند و مداخل‌های ارجاعی را که درمورد برخی مدخل‌ها پیش می‌آیند به طور دستی تحریر می‌نماید. نتیجه این کار یک فایل فرعی نیز به دست می‌آید. نمایه ساز فایل اصلی مدخل‌ها و فایل فرعی را ادغام و توسط یک برنامه ترتیب بندی مناسب الفبایی می‌کند. نمایه ساز این مداخل الفبایی شده را مطالعه می‌کند و هرگاه به فرایند تبدیل با کما نیاز باشد فرایند به طور دستی در فایل انجام می‌دهد سپس برنامه الفبایی کننده شماره صفحه‌ها را به طور صعودی مرتب می‌کند.

2- نمایه سازی در محیط دستوری² (نمایه سازی فرمانی). این نرم افزار در محیط‌های رابط دستوری کار می‌کند. کار با این سیستم به این شکل است که نمایه ساز پس از اتمام صفحه آرایی و مشخص شدن شماره‌های صفحه‌های کتاب، صفحه‌ها را در صفحه نمایش کامپیوتر مطالعه می‌کند و کلیدواژه‌ها را با دو نماد کاملاً استثنایی چپ و راست که در مشابه آن در متن وجود نداشته باشد، علامت گذاری می‌کند (مثلاً کلیدواژه درون سه گروه قرار می‌گیرد) که این هم نوعی انتخاب کلیدواژه است. سپس نمایه ساز در همان فایل ارجاعات مورد نظر تحریر می‌کند. پس از اتمام

¹ Graphic User Interface (GUI)

² Command Line Interface

علامت‌گذاری با گروه‌ها و ارجاعات و تبدیل‌ها یک ابزار نرم‌افزار می‌تواند کل نمایه را تهیه و الفبایی کند. اگر نرم‌افزار دارای امکانات نمایه‌سازی نباشد ویراستار می‌تواند یک برنامه بسیار ساده (مثلاً به زبان بیسیک بنویسد) که فهرست کلیدواژه‌های علامت‌گذاری شده و شماره صفحه‌های آن‌ها را استخراج کند و یک برنامه یا ابزار ساده دیگر نمایه را اولویت‌بندی و مرتب کند. روش‌های نمایه‌سازی کتاب با کمک رایانه دارای مزایای سرعت ورود اصطلاحات نمایه‌ای، ظرفیت ذخیره سازی بالا، سهولت اصلاح، حروفچینی و تایپ، ویرایش و چاپ نمایه هستند و از لحاظ اقتصادی مقرون به صرفه می‌باشند (نوروزی 1380).

۴. نمایه سازی ماشینی:

پروژه های موفق

4-1. تجارب نمایه‌ساز/نمایه‌سازی ماشینی

در این بخش به برخی از سیستم‌های نمایه‌سازی ماشینی پرداخته می‌شود. این مدل‌ها دارای ابعاد موضوعی، جامعیت و مانعیت گوناگونی هستند. برخی از این مدل‌ها، ترکیبی از مدل‌های دیگر است و هر محقق با توجه به نیاز زبان‌شناختی و فرهنگی کشور خویش مدل خاصی را یا ابداع و یا بومی نموده است.

«کی‌پی ماینر»¹: سیستمی برای استخراج اصطلاحات کلیدی از متون انگلیسی و عربی

«ال‌بلتاگی»² و «رافیا»³ (2009) سیستمی با نام «کی‌پی ماینر» معرفی کردند که برای استخراج اصطلاحات کلیدی از متون عربی و انگلیسی مناسب است. شایان ذکر است که طول اصطلاحاتی که از طریق این سیستم استخراج می‌گردد متنوع است و ثابت نیست. برخلاف سیستم‌های استخراج کلمه فعلی، «کی‌پی ماینر» نیازی به یادگیری دربارهٔ یک مجموعه مدرک ندارد تا بتواند وظیفهٔ خویش را انجام دهد. این سیستم دارای این مزیت است که می‌توان تنظیمات ویژه‌ای در آن لحاظ کرد؛ زیرا قواعد و تصمیمات ذهنی که سیستم اتخاذ می‌نماید به ماهیت کلی مدارک و اصطلاحات کلیدی وابسته است.

«کی‌پی ماینر» طی چند مرحله به استخراج اصطلاحات کلیدی می‌پردازد که عبارت‌اند از:

1. انتخاب اصطلاح کلیدی نامزد: از آنجایی که الگوریتم به کار رفته در این سیستم هیچ دانشی نسبت به چیستی یک اصطلاح ندارد، از مجموعه‌ای از شرط‌ها به منظور فراخواندن

¹ KP-Miner

² El-Beltagy

³ Rafea

اصطلاح کلیدی نامزد بهره می‌گیرد. در مجموع، 187 از کلمات حذفی رایج (در، به، برای، ...) در مرحله انتخاب اصطلاحات کلیدی نامزد استفاده می‌شود.

2. محاسبه وزن برای اصطلاح کلیدی نامزد: همان‌گونه که در برخی پژوهش‌ها نیز اذعان شده است، موفقیت برنامه استخراج اصطلاح کلیدی به این معنی است که برنامه قادر به استخراج اصطلاحاتی است که به بهترین شکل مدرک را می‌نمایاند، برای نمونه، هدف آن فراخوانی ویژگی‌های کلیدی یک مدرک است. در کل، معادله زیر برای محاسبه وزن اصطلاحات کلیدی نامزد (بسیط یا جمع) استفاده می‌شود:

$$w_{ij} = tf_{ij} * idf * B_i * P_f$$

که w_{ij} وزن اصطلاح t_j در مدرک D_i است. tf_{ij} فراوانی اصطلاح t_j در مدرک مورد نظر، و idf برابر است با $\log_2 \bar{N}/n$ که N تعداد مدارک موجود در مجموعه و n تعداد مدارکی است که اصطلاح t_j دست‌کم یک‌بار در آن‌ها ظاهر شده است. اگر اصطلاح مرکب باشد، n برابر با 1 قرار داده می‌شود. B_i عامل کمکی است که به مدرک D_i وابسته است. P_f عامل مربوط به مکان اصطلاح است. اگر قواعد مربوط به مکان کلمه مورد نظر نباشد این مقدار 1 در نظر گرفته می‌شود.

3. پالایش فهرست نهایی اصطلاحات نامزد: «کی پی ماینر» به کاربر امکان می‌دهد تا عدد n مربوط به اصطلاحات کلیدی که وی می‌خواهد را مشخص نماید. این سیستم از فهرستی مرتب شده برای برگرداندن n اصطلاح کلیدی که در ابتدای فهرست جای دارد و مورد تقاضای کاربر است استفاده می‌نماید. عدد پیش فرض برای n عدد 5 است. هنگامی که اصطلاحات کلیدی نامزد تولید می‌شوند، طولانی‌ترین توالی ممکن از کلمات که به وسیله یک تعیین کننده کلمه مشخص می‌گردند جستجو و ذخیره می‌گردند.

4. تعدیل کردن سیستم برای کار با اصطلاحات عربی: تغییرات اندکی بر روی سیستم انگلیسی فعلی برای هم‌ساز شدن با اصطلاحات عربی صورت پذیرفته است. پاره‌ای از این تغییرات جزئی عبارت‌اند از استفاده از یک فهرست حذفی عربی و یک ریشه گیر¹ زبان عربی. در مورد فهرست حذفی، در کل، تعداد 585 کلمه و فعل رایج استفاده شده است. از آنجا که زبان عربی یک زبان بسیار صرفی و دارای تحولات لغوی غنی است، عمل ریشه‌گیری بسیار مشکل‌تر از زبان انگلیسی است.

¹ stemmer

نمایه‌سازی خودکار متون عربی

«منصور» و همکاران (2008) به ارائه روشی برای نمایه‌سازی خودکار متون عربی پرداختند. روش آن‌ها بسیار بر تحلیل ریخت‌شناختی و تخصیص وزن‌هایی به کلمات متکی است. در تحلیل ریخت‌شناختی، از تعدادی از قواعد دستوری برای استخراج ریشه کلماتی که نامزد نمایه می‌شوند استفاده می‌گردد. تکنیک‌های اختصاص وزن، وزن‌های این کلمات را با در نظر گرفتن مدرک حاوی آن‌ها محاسبه می‌کند. وزن‌دهی بر این اساس است که چگونه یک کلمه در یک مدرک توزیع شده است و نه فقط بر مبنای فراوانی حضور کلمه در مدرک. سپس، کلمات نامزد نمایه به ترتیب نزولی وزنشان مرتب می‌شوند به گونه‌ای که بازیابندگان اطلاعات بتوانند مهم‌ترین کلمات نمایه‌ای را برگزینند. این روش پس از تست شدن دارای جامعیت 46 درصد و مانعیت 64 درصد بوده است که درصد بالایی است.

روش آن‌ها شامل سه مرحله است:

1. **مرحله پیش‌پردازش:** قبل از استخراج ریشه کلمات، اولین مرحله که مرحله آماده‌سازی نیز نامیده می‌شود شامل گام‌هایی برای پیش‌پردازش کلمه است، که در ادامه تبیین می‌شوند.

الف) مرحله بساوند

در این مرحله، کلمات قبل از ریشه‌گیری، طبقه‌بندی می‌گردند. هر کلمه موجود در مدرک، از این نظر که اسم هست یا فعل، اسم مفرد است یا جمع، فعل زمان حال، گذشته یا آینده است، و اینکه ضمیر به آن‌ها وصل است یا خیر ارزیابی می‌شوند. در این مرحله، هر کلمه با مجموعه‌ای مناسب از بساونده‌ای از پیش تعریف شده مقایسه می‌شوند. برای نمونه، افعال با توجه به فعل «فَعَلَ» (انجام داد) بررسی می‌گردند زیرا این فعل به عنوان استاندارد در زبان عربی محسوب می‌شود.

ب) حذف اصطلاحات و عبارت‌های موجود در فهرست حذفی

در این مرحله اصطلاحات حذفی، حذف می‌گردند.

ج) شناسایی فعل‌ها و اسم‌ها

سیستم‌های بازیابی اطلاعات معمولاً از پردازش زبان طبیعی استفاده می‌کنند تا یک عنصر را به طور خودکار نمایه‌سازی نمایند. یک عنصر می‌تواند شامل یک مدرک، یک تصویر، یا اطلاعات شنیداری و غیره باشد. این نوع پردازش نیازمند اطلاعات واژگانی، نحوی، و دستوری از عنصر مورد

نظر می‌باشد و به طور زیادی به قواعد دستوری زبان مورد نظر بستگی دارد. روش این پژوهشگران دارای دو معیار برای تشخیص اسم یا فعل بودن یک کلمه است. اولین معیار این است کلمه‌ای که قبل از کلمه مورد نظر ما استفاده شده است مورد پرسش قرار گیرد. اگر کلمه پیشین یک اصطلاح فهرست حذفی باشد، برخی از اصطلاحات فهرست حذفی قبل از اسم می‌آیند و برخی دیگر قبل از فعل. برای نمونه، در عربی، بعد از «لَمْ»، فعل می‌آید در حالی که بعد از حروف اضافه و برخی ضمیرها اسامی می‌آیند.

معیار دوم، وزن کلمه مورد نظر است. اگر کلمه بر وزن «فاعل» یا «مفعول» باشد در بسیاری از موارد اسم است. در مواقعی که هر دو معیار صادق نباشند، ضمائر متصل ارزیابی می‌گردند. برخی از ضمائر تنها به فعل‌ها می‌چسبند در حالی که برخی دیگر فقط به اسم‌ها متصل می‌گردند.

2. مرحله استخراج ریشه کلمه

پس از شناسایی کل مدرک، شناسایی واژگان موجود در آن، حذف اصطلاحات و عبارت‌های حذفی، و شناسایی اسم‌ها و فعل‌های آن، از یک الگوریتم ریشه‌گیری برای هر کلمه استفاده می‌گردد. اشتقاق کلمه به ریشه‌های آن به عنوان مرحله‌ای مهم در نمایه‌سازی خودکار محسوب شده است. ریشه‌گیری طی چند مرحله انجام می‌شود که در ادامه می‌آیند.

الف) استخراج ریشه کلمات از فعل‌ها

فعل‌های متفاوت نیازمند روش‌های ریشه‌گیری متفاوتی هستند. هدف هر الگوریتم که کلمات ریشه‌ای را از فعل‌ها بیرون می‌کشد این است که که شکل سه‌حرفی اولیه را از فعل استخراج نماید. به هر حال، حذف ضمائر متصل اولین گام اساسی است.

0 بررسی ضمائر پیشوندی/پسوندی

اولین مرحله در الگوریتم ریشه‌گیری پژوهشگران یادشده این است که بررسی شود که یک کلمه فرضی دارای ضمائر متصل هست یا خیر. ضمیرها در زبان عربی به دو شکل ظاهر می‌شوند: متصل و منفصل.

0 بررسی فعل‌ها با در نظر گرفتن 5 فعل استاندارد

در زبان عربی، پنج فعل استاندارد مشهور وجود دارد که به پنج فعل شناخته می‌شوند. این فعل‌ها در شکلی خاص ظاهر می‌شوند و ویژگی‌های منحصر بفردی دارند. آن‌ها همیشه در زمان حال ظاهر می‌گردند و معمولاً با حرف «ن» پایان می‌پذیرند. همچنین، این فعل‌ها حروف اصلی و غیر اصلی دارند. نمونه‌ای از این فعل‌ها «یَفْعَلُونَ» است (آن‌ها انجام می‌دهند)، که «یَ» و «ون» غیر اصلی هستند.

0 بررسی فعل‌ها با توجه با «ده اضافات فعل»¹

در زبان عربی، هر فعل شامل فقط سه حرف است. فعل‌ها شامل بی از سه حرفاند صرفاً اشتقاقی از فعل سه کلمه‌ای اصلی‌شان هستند. اشتقاقیات هر فعل در 10 قالب خاص بروز می‌نمایند. سه تا از این قالب‌ها با اضافه کردن تنها یک حرف به فعل اصلی شکل می‌پذیرند، پنج تا از آن‌ها از طریق اضافه کردن دو حرف، و دو قالب دیگر از طریق اضافه کردن سه حرف به آن صورت می‌پذیرد. این 10 شکل به 10 اضافات فعل مشهورند.

ب) استخراج کلمات ریشه‌ای از اسم‌ها

استخراج یک کلمه ریشه‌ای از یک اسم چند گام مشابه ریشه‌گیری از فعل دارا است، اما پیچیده‌تر است. پیچیدگی ریشه‌گیری از یک اسم به دلایل متفاوتی است. اولین دلیل این است که یک اسم می‌تواند به شکل مفرد، جمع، یا دوتایی (اثان) شود. همچنین، هر یک از این سه قالب، هنگامی که یک اسم به مذکر یا مؤنث اشاره دارد، نیز فرق می‌کند. علاوه بر این، برخی استثناها برای شکل‌های دوتایی و جمع وجود دارد. به طور کلی این فرایند بدین صورت انجام می‌شود: بررسی ضمائر متصل و منفصل، بررسی اسم با توجه به اسم‌های 5 تایی، درآوردن یک اسم به شکل مفرد آن، مقایسه اسم با مشتقات رایج آن، مشتقات «م» (هنگامی که حرف اولم است)، مشتقات «ت» (وقتی حرف اول بایت شروع می‌شود)، و مشتقات گوناگون (وقتی که اولین حرف «م» یا «ت» نیست).

3. تخصیص وزن و انتخاب نمایه

الف) محاسبه وزن

¹ ten-verb-additions

در مرحله سوم روش، وزنهایی به ریشه کلمات مربوط به مدرک اختصاص می‌یابد. وزن هر کلمه بستگی به سه عامل دارد، که اهمیت یک کلمه را در مدرک حاوی آن نشان می‌دهد. اولین عامل بسامد رخداد آن کلمه در مدرک حاوی آن است. عامل دوم، شمارش کلمات ریشه‌ای برای آن کلمه است. ما عامل سوم نیز پیشنهاد می‌کنیم و آن میزان پخش¹ کلمه در کل مدرک است.

به اصطلاحات زیر که در فرمول محاسبه وزن به کار رفته‌اند توجه داشته باشید: N تعداد کلمات موجود در یک مدرک است، m تعداد حضور یک واژه معین در یک مدرک، sm تعداد کلمات ریشه‌ای برای آن کلمه به کار رفته در مدرک، و f عاملی است که بیانگر پخش کلمه مورد نظر در همان مدرک است. وزن w که با توجه به مدرک حاوی کلمه به آن اختصاص می‌یابد نشان دهنده اهمیت آن کلمه در گزینش به عنوان یک کلمه نمایه‌ای است.

فرمول کلی به صورت زیر است:

$$w = m \times sm \times f$$

برای محاسبه عامل پخش یک کلمه در یک مدرک باید به اصطلاحات زیر توجه داشت:

که d فاصله یک کلمه است و بیانگر مکان کلمه در یک مدرک است. برای نمونه، فاصله

اولین اصطلاح در یک مدرک 1 است. و فاصله آخرین کلمه در یک مدرک N است. همچنین d_i فاصله i امین کلمه است.

ad میانگین فاصله‌های هر کلمه ریشه‌ای است و به صورت زیر تعریف می‌گردد:

$$ad = \left[\frac{\sum_{i=1}^{sm} d_i}{sm} \right]$$

id فاصله ایده‌آل بین هر دو حضور کلمه ریشه‌ای است. بنابراین، در فاصله ایده‌آل، باید فاصله

بین هر دو کلمه متوالی یکسان باشد. فاصله ایده‌آل به صورت زیر محاسبه می‌شود:

$$id = \left[\frac{N}{sm + 1} \right]$$

¹ spread

rid یک فاصله ایده آل مرجع برای همه کلمات ریشه ای است. rid برای همه واژگان ریشه ای باید یکسان باشد به طوری که بتوان فاصله های ایده آل کلمات ریشه ای را با توجه به این مرجع اندازه گیری کرد.

g شکاف است، که تفاوت بین rid و ad را نشان می دهد. به عبارت دیگر:

$$g = \text{rid} - \text{ad}$$

به خاطر داشته باشید که کاهش در g بیانگر توزیع یا پخش مناسب کلمه در مدرک است.

برای محاسبه وزن، باید استفاده از f تابعی از g باشد یعنی $f(g)$.

بر اساس ملاحظاتی که در بالا یاد شد، $f(g)$ را می توان به صورت زیر تعریف کرد:

$$f(g) = \alpha + \frac{(\beta - \alpha)}{K^g}$$

که K یک مقدار ثابت است و $K \in \mathbb{R}^+; K > 1$.

4. انتخاب نمایه:

پس از تخصیص وزن به کلمات نامزد نمایه، باید به انتخاب کلمات نمایه ای پرداخت. روشی که ما به کار می بریم آن است که در مورد نمایه کتاب نیز به کار می رود. نمایه هر کتابی شامل کلیدواژه گانی است که به صورت الفبایی مرتب و با شماره صفحه در جلوی آن ها مشخص شده اند. هنگامی که واژگان موجود در یک مدرک ریشه گیری شدند، و وزن های آن ها طبق فرمول های بالا محاسبه شدند، کاربر ممکن است حد آستانه ای را برای وزن ها در نظر بگیرد و کلمات نمایه ای را بر این اساس انتخاب نماید. به این ترتیب، تنها کلماتی که وزن های آن ها بالاتر از حد آستانه است بازیابی و به عنوان کلمات نمایه ای منظور می شوند.

«پای»¹: نمایه سازی خودکار برای استخراج قطعی کلیدواژه گان از یک مدرک

«ماتسومورا»، «اوساوا»، و «ایشیزوکا» (2002) به معرفی سیستم «پای» که یک سیستم نمایه سازی ماشینی است پرداختند. این سیستم بسیار بر کلیدواژه گانی که نویسنده مدرک انتخاب کرده است تأکید دارد. ایده مبنایی «پای» این است: از آنجایی که نویسنده ای که یک مدرک را می نویسد نکات

¹ PAI

مورد توجه خویش را به خوبی بیان می‌دارد، اصطلاحات برانگیزاننده¹ که در ذهن خواننده تولید می‌شوند می‌توانند کلیدواژگان قطعی را بنمایانند. «پای» از یک «مدل فعال‌سازی پخش»² بهره می‌گیرد بدون استفاده از مجموعه‌ای از نوشتجات، اصطلاح‌نامه‌ها، تحلیل نحوی، روابط وابستگی بین اصطلاحات، یا هر دانش دیگر. در این مدل، اصطلاحات به عنوان نودهایی³ مطرح می‌گردند. و روابط بین این اصطلاحات به عنوان پیوندهای انجمنی بین نودها نمایانده می‌شوند. یکی از فرمول‌های ریاضیاتی که برای مدل فعال‌سازی پخش یا توزیع به کار می‌رود در ادامه بیان و تفسیر می‌شود:

$$A(t) = C + ((1 - \gamma)I + \alpha R) A(t - 1)$$

که $A(t)$ برداری است که فعالیت‌های نودها را در مرحله انفصال می‌نمایاند، $t = 1, 2, \dots, N$ که $A(t)_i$ فعالیت نود i را در مرحله t نشان می‌دهد. R ماتریسی است که شبکه فعال‌سازی را نشان می‌دهد. و $R_{i,j} (i \neq j)$ نشان دهنده قدرت ارتباط بین نود i و j است. و عناصر قطری (اریب) $R_{i,i} (i = j)$ که مقدار آن صفر است. C برداری است که فعالیت‌های تزریق شده به شبکه فعال‌سازی R را می‌نمایاند. و C_i نمایاننده فعالیت‌هایی است که در نود i تزریق یا پمپ می‌گردد. I یک ماتریس همانی است. $(0 < \gamma < 1)$ یک پارامتر برای کم کردن فعالیت نود، و α پارامتری برای تعیین مقدار فعالیت‌های هر نود با نودهای مجاور خویش است.

به طور کلی، الگوریتم «پای» دارای پنج مرحله اساسی است:

1. مرحله پیش‌پردازش: آماده‌سازی، حذف کلمات اضافی، پیشوندها و پسوندها، و شناسایی اصطلاحات موجود در مدرک.

2. بخش‌بخش کردن⁴: مطابق با انسجام نحوی، هر مدرک به چند قسمت تقسیم می‌شود.

$$S_t (t = 1, \dots, n)$$

¹ impressive

² spreading activation model

³ nodes

⁴ segmentation

3. شبکه فعال سازی: برای هر بخش، اصطلاحات از طریق بسامدهاشان مرتب می شوند، و برترین

اصطلاحات $N\%^2$ تفکیک می شوند با $K(t)$ به عنوان مفاهیم بنیادین. ارتباط اصطلاحات w_i و w_j این گونه تعریف می شوند:

$$assoc(w_i, w_j) = \sum_{s \in S_t} \min(|w_i|_s, |w_j|_s),$$

که $|x|_s$ تعداد x را در جمله s تفکیک می کند. جفت اصطلاحات در $K(t)$ از طریق $assoc$ مرتب می شوند.

برای پیوندهای بین w_i و w_j ، $R(t)_{i,j}$ به صورت زیر تعریف می گردد:

$$R(t)_{ij} = \frac{assoc(w_i, w_j)}{links(w_i)},$$

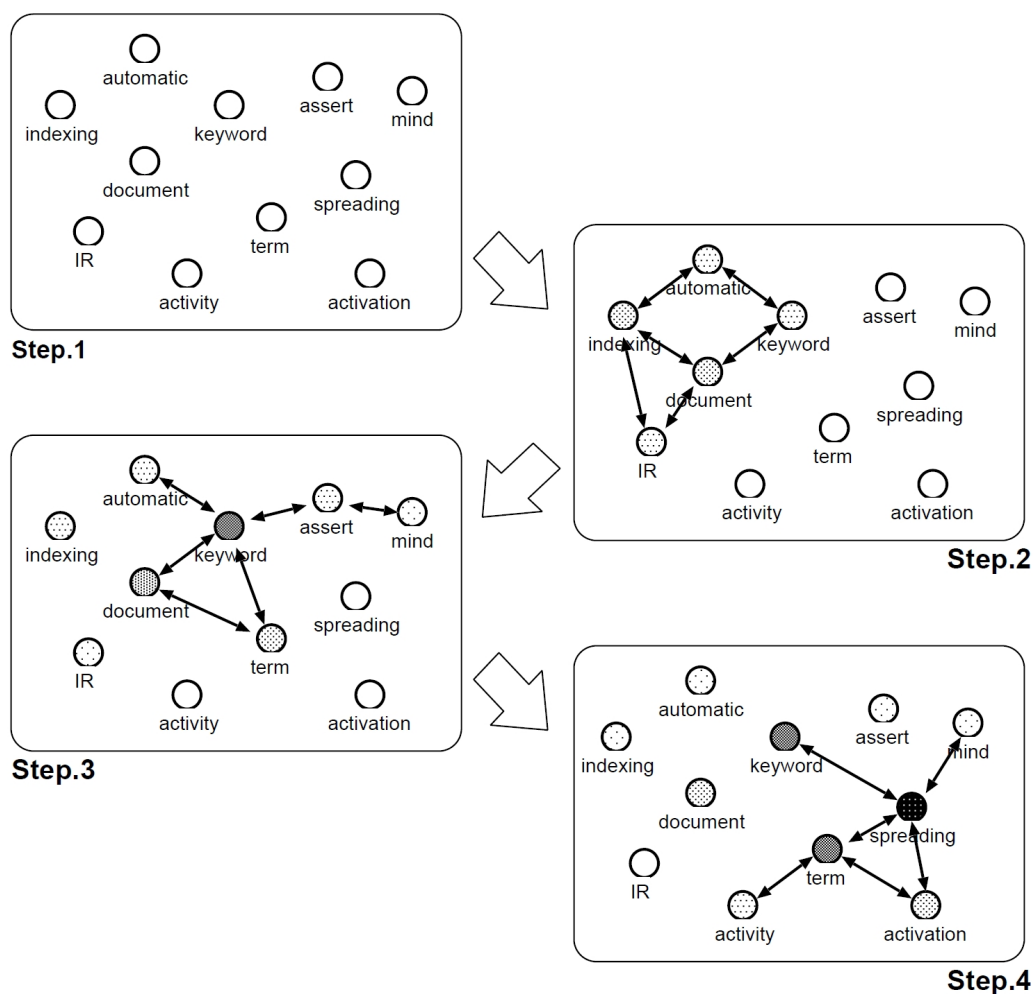
که $links(w_i)$ نشان دهنده تعداد پیوندهای ورودی به w_i است. عنصر دیگر در $R(t)$ صفر تعریف می گردد.

4. فعال سازی پخش یا توزیع:

فعالیت اصلی هر اصطلاح قبل از اجرای عمل فعال سازی پخش برابر با یک است. پارامترهای γ و α بایستی از طریق آزمون و خطا لحاظ گردد زیرا آنها بستگی به ویژگی های مدارک دارند.

5. استخراج کلیدواژگان:

پس از اعمال فعال سازی پخش بر روی همه بخش ها، اصطلاحات بسیار فعال به عنوان نکته اصلی مدنظر نویسنده لحاظ می گردند. به هر حال، اگر فعالیت چندان زیاد نباشد، اصطلاحی که مفاهیم بنیادین را ابراز می دارد نیز به عنوان نقطه نظر نویسنده تلقی می گردند. این مراحل در شکل زیر نشان داده شده اند.



شکل 4-1. فرایند نمایه سازی در «پای» (Matsumura, Ohsawa, and Ishizuka 2002)

استفاده از «ان گرامز»¹ برای بازیابی متون کره‌ای

«لی» و «آن» (Lee and Ahn 1996) به ارائه روشی برای نمایه سازی خودکار متون کره‌ای پرداختند. در زبان کره‌ای، صدها پسوند و پیشوند وجود دارد و علاوه بر این، وندهای موردنظر می‌توانند به روش‌های مختلفی ترکیب گردند و کلمات و اصطلاحات مرکب تولید نمایند. این ویژگی زبان کره‌ای، ریشه‌گیری کلمات را بسیار مشکل کرده است. بنابراین، پژوهشگران یادشده با عطف به این موضوع، از روش نمایه سازی واژک‌مدار² استفاده نمودند. این روش نمایه سازی دارای مراحل زیر است:

¹ n-Grams

² Morpheme-based

1. شناسایی کلمات متنی مفرد
2. تفکیک کلمه به همه واحدهای تشکیل دهنده یا واژکهای آن
3. انتخاب محتمل ترین واژکها
4. استخراج اسامی ساده از واژکها
5. حذف اسامی ساده موجود در فهرست حذفی
6. شکستن ریشه های معنادار کلمات به «ان گرامها»

در این روش برای وزن دهی کلمات از جدول زیر استفاده می شود:

جدول مؤلفه های وزن دهی طرح واردها

Table 2: Components of term weighting schemes

<i>Term Frequency Component</i>		
<i>b</i>	1.0	binary weight equal to 1 for terms present in a vector (term frequency is ignored)
<i>n</i>	<i>tf</i>	raw term frequency (number of times a term occurs in a document or query text)
<i>a</i>	$0.5 + 0.5 \frac{tf}{\max tf}$	augmented normalized term frequency (use maximum normalization where each <i>tf</i> is divided by maximum <i>tf</i> , and further normalize the resulting value to lie between 0.5 and 1.0)
<i>l</i>	$\ln tf + 1.0$	logarithmic term frequency which reduces the importance of raw term frequency in those collections with widely varying document length
<i>Collection Frequency Component</i>		
<i>n</i>	1.0	no change in weight; use original term frequency component (<i>b</i> , <i>n</i> , <i>a</i> , or <i>l</i>)
<i>t</i>	$\ln \frac{N}{n}$	multiply original term frequency component by an inverse document frequency factor (<i>N</i> is the total number of documents in the collection, and <i>n</i> is the number of documents to which a term is assigned)
<i>Normalization Component</i>		
<i>n</i>	1.0	no change; use factors derived from term frequency and collection frequency only (no normalization)
<i>c</i>	$\frac{1}{\sqrt{\sum_{vector} w_i^2}}$	use cosine normalization where each term weight <i>w_i</i> is divided by a factor representing Euclidian vector length

«جینکوئری: نظام نمایه سازی متون ژاپنی»

«فوجی» و «کرافت» (1993) نظامی برای نمایه سازی متون ژاپنی ابداع کردند. آنها با استفاده از

نظام «اینکوئری»¹، نظامی برای نمایه سازی زبان ژاپنی با نام «جینکوئری»¹ به وجود آوردند. نظام

¹ INQUERY

«اینکوئری» وابسته به زبان نیست. این نظام دارای دو ماژول است. یکی از این ماژول‌ها، ماژول نمایه‌سازی برای ایجاد پایگاه داده‌ای از منابع متنی است. از آنجایی که کلمات ژاپنی در متن به وسیله فاصله‌ها از هم مجزا نمی‌شوند، یک برنامه بخش‌بخش‌سازی برای شناسایی اصطلاحات نمایه‌ای از مدارک به کار می‌رود. پژوهشگران یاد شده از برنامه تفکیک «جومان»² برای نمایه‌سازی کلمه‌مدار استفاده کرده‌اند. ماژول دوم، ماژول آغازین پرسش³ برای فرموله کردن یک پرسش ساختاری با استفاده از جنبه‌های دستوری ژاپنی است. هنگامی که یک پرسش به شکل زبان طبیعی در سیستم وارد می‌شود، «جومان» به قسمت‌بندی آن می‌پردازد. به طور کلی «جینکوئری» فعالیت‌های زیر را انجام می‌دهد:

1. استخراج کلمات موجود در مدرک؛
2. حذف واژگان موجود در فهرست حذفی از مدرک؛
3. ریشه‌گیری و تولید اصطلاحات نمایه‌ای؛
4. محاسبه اطلاعات مورد نیاز (بسامد واژگان، و غیره)
5. اشاره‌گرهایی از اصطلاح به مدارک برای ساختن یک فایل مقلوب⁴.

«جیپسی»: نمایه‌سازی جغرافیایی خودکار متن مدارک

«وودروف» و «پلانت» (Woodruff and Plaunt 1994) سیستمی با نام «جیپسی»⁵ برای نمایه‌سازی خودکار ابداع نمودند. استراتژی مبنایی «جیپسی» استخراج نام‌های مکان‌ها و شاخص‌های جغرافیایی از مدارک و استفاده از این ارتباط‌های درونی برای شناسایی ناحیه‌ای که مدرک به آن اشاره می‌نماید. روش نمایه‌سازی یادشده دارای سه مرحله اصلی است:

1. تجزیه⁶: در این مرحله کلیدواژگان و اصطلاحات مرتبط استخراج می‌گردند. یک سیستم تجزیه‌گر به این منظور باید بداند که چگونه دو نوع اصطلاحات جغرافیایی را شناسایی نماید که عبارت‌اند از:

¹ JINQUERY

² JUMAN

³ Query front-end module

⁴ Inverted file

⁵ GIPSY

⁶ Parsing

0 اصطلاحاتی که با اشیاء یا ویژگی‌های موجود در مجموعه داده‌ها مطابقت دارند. این مرحله نیازمند یک اصطلاح‌نامه جامع می‌باشد که بخشی از آن به طور دستی و بخشی از آن به طور خودکار تولید شده است.

0 ساختارهای واژگانی که اطلاعات فضایی و مکانی را دربر دارد، برای نمونه، «جنوب»، «در»، «مجاور» و.... برای اجرای این مرحله، فهرستی از ساختارهای رایج بایستی ایجاد گردد.

هنگامی که اجزای تشکیل دهنده عبارت‌های جغرافیایی شناسایی شد، برخی فعالیت‌های صرف و نحوی بایستی انجام گیرد تا قسمت‌های نامربوط عبارت‌ها حذف گردند. برای نمونه در «جنوب دلتا» همه عبارت باید ذخیره گردد در حالی که در «منابع مربوط به دلتا» باید فقط «دلتا» ذخیره گردد.

2. مکان‌یابی¹ داده‌های مناسب: برون‌داد سیستم تجزیه‌گر در تابعی اعمال می‌گردد که داده‌های مناسب با اصطلاحات و عبارت‌های استخراج شده را بازیابی می‌نماید. هنگامی که مجموعه داده‌های مناسب بازیابی شد، باید به وسیله یک تابع احتمالاتی پردازش گردند تا بتوان به فهرستی چند وجهی رسید که هر یک از آن‌ها نیازمند یک استراتژی ویژه‌ای برای وزن‌دهی ربط کلمات هستند. بنابراین، وزن‌دهی ربط بر مبنای (1) اصطلاحات جغرافیایی که استخراج گشته‌اند؛ (2) فضای مدرک و تراکم این اصطلاحات در آن؛ (3) دانش مربوط به عناصر جغرافیایی در پایگاه داده و ویژگی‌های آن؛ و (4) استنتاج فضایی درباره اصطلاحات جغرافیایی موجود در عبارت‌ها مانند «کنار کالیفرنیا».

3. هم‌پوشان کردن چندوجهی‌ها²: آخرین مرحله‌ای که بایستی انجام گردد هم‌پوشی داده‌های چندوجهی است که از مرحله دو استخراج شده‌اند. این مرحله نیازمند توابع محاسباتی است که روابط درونی این چندوجهی‌ها را دریابد و نیز ارزش ربط درهم‌کردی برای هر چندوجهی ایجاد نماید.

¹ locating

² Polygons overlay

«کی گراف»^۱: نمایه سازی خودکار با استفاده از گراف هم‌رویداد^۲

«اوساوا»^۳ و همکاران (1998) الگوریتمی برای استخراج ایده اصلی از چکیده متون با نام «کی گراف» راه‌اندازی نمودند. الگوریتم پیشنهادی، بر اساس بخش‌بخش‌سازی و تفکیک گراف، به بازنمایی میزان هم‌رویدادی بین اصطلاحات موجود در یک مدرک، به صورت چند خوشه می‌پردازد. هر خوشه به یک مفهوم وابسته است که مد نظر نویسنده بوده است، و با یک تحلیل آماری، اصطلاحاتی که بالاترین جایگاه را با توجه به اندیشه نویسنده داشته‌اند به عنوان کلیدواژه گزینش می‌گردند. این استراتژی بر این اندیشه استوار است که یک مدرک مانند یک ساختمان، ساختار یافته است که بر اساس مفاهیم سنتی، ایده‌های نوینی ارائه می‌گردند. نتایج حاکی از آن است که اصطلاحات استخراج شده کاملاً با کلیدواژه‌گانی که مدنظر نویسندگان بوده است همخوانی دارند. این الگوریتم دارای سه مرحله است:

1. استخراج بنیان‌ها: که مفاهیم پایه‌ای و مقدماتی به عنوان خوشه‌های در نظر گرفته می‌شوند، و هریک از آن‌ها از طریق هم‌رویدادی بین اصطلاحات موجود در مدرک شکل داده می‌شوند.
 2. استخراج ستون‌ها^۴: که ستون‌ها (برای نمونه، روابط بین اصطلاحات در مدرک و مفاهیم پایه‌ای استخراج شده در بخش 1) به دست می‌آیند.
 3. استخراج سقف‌ها^۵: که نودها (که بیانگر اصطلاحات هستند) در پیوندگاه ستون‌های مستحکم، به عنوان سقف به حساب می‌آیند. این مرحله همان استخراج کلیدواژه‌گان است.
- نمونه‌ای از روابط معنایی استخراج شده از این روش در شکل زیر نشان داده شده است.

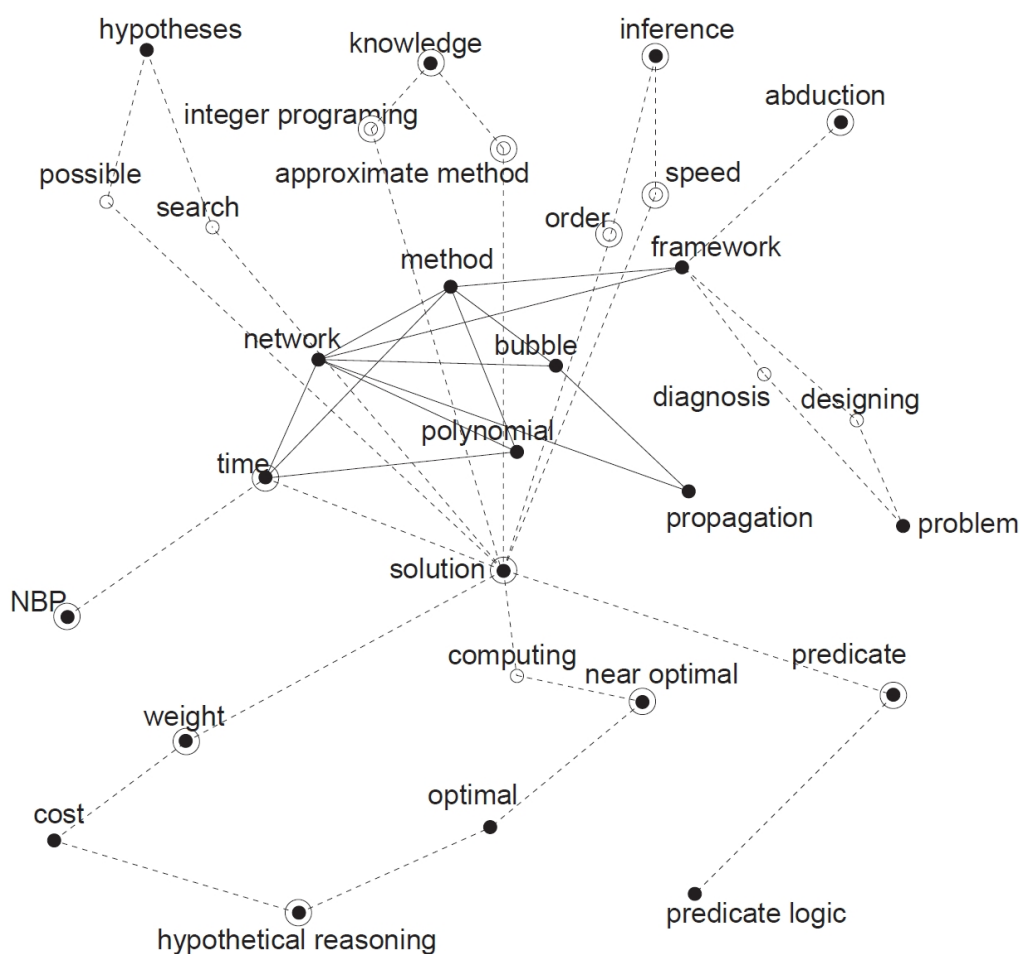
¹ KeyGraph

² Co-Occurance

³ Ohsawa

⁴ Columns

⁵ Roofs



شکل 2-4. روابط معنایی استخراج شده از روش «کی گراف»

نمایه سازی ماشینی دانشگاه «دارمشتاد»^۱

یک طرح پژوهشی که در دانشگاه «دارمشتاد» در آلمان صورت پذیرفته برای نمایه سازی ماشینی اجزای زیر را در نظر گرفته است:

1. ارتقای واژه نامه: مدارکی که به شیوه دستی نمایه سازی شده بودند باید با این هدف مورد تجزیه و تحلیل قرار گیرند که ارتباطات آماری میان فرمول بندی متن و توصیفگرهای مربوط

¹ Darmstadt University

به آن که به طور دستی برقرار شده است حفظ گردد. توصیفگرها می‌توانند همان نشانه‌های رده‌بندی یا مفاهیم اصطلاح‌نامه باشند که به صورت کنترل شده یا آزاد انتخاب شده‌اند.

2. تحلیل مدرک: مدارک باید بر این اساس مورد بررسی قرار گیرند که کدام واژه‌ها یا عبارات مربوط به نمایه‌سازی موجود در مدارک، به چه شکلی، و در چه قالب متنی وجود دارند. بنابراین، لازم است که ماشین، اطلاعات خاصی درباره قوانین نحو و ریخت‌شناسی و ساختار منطقی مدارک را بشناسد. همچنین، کیفیت بالای سیستم، به کمک واژه‌نامه‌ها و توصیفگرهای مرتبط حفظ شود.

3. عملیات نمایه‌سازی: نمایه‌سازی وظیفه بررسی این کار مهم را بر عهده دارد که با چند درصد اطمینان، توصیفگرها بر اساس اطلاعات موجود در مدرک انتخاب شده یا نشده‌اند. این کار باید در دو مرحله انجام شود. نخست باید پارامترهای عملیات نمایه‌سازی بر روی داده‌ها آموزش داده شوند، سپس عملیات نمایه‌سازی اجرا گردد (Knort 1997).

2-4. مروری بر پاره‌ای از نرم‌افزارهای موجود نمایه‌سازی ماشینی

«سیندکس»¹: نمایه‌سازی ماشینی کتاب‌ها و مجلات

«سیندکس» یکی از برنامه‌های تهیه نمایه‌هایی خودکار برای کتاب‌ها، روزنامه‌ها، و دیگر نشریات ادواری است. این برنامه ویژگی‌های نمایه‌ساز حرفه‌ای را تقلید می‌کند و کاربرد آن بسیار ساده است. از این برنامه می‌توان در ساخت اصطلاح‌نامه‌ها، یا مستندات موضوعی که بنیانی برای نمایه‌سازیهای خودکار کنونی تلقی می‌شوند نیز بهره برد.

«سیندکس» به عنوان نمایه‌ساز جانشین شما نمی‌گردد، بلکه از شما به وسیله چهریزه‌هایی که در برنامه گنجانده است پشتیبانی می‌کند و کار نمایه‌سازی را تسهیل می‌نماید. «سیندکس» عملیاتی که از سیستم نمایه‌سازی انتظار دارید را انجام می‌دهد بدون آنکه کار زیادی انجام دهید.

همه مراحل وقت‌گیر نمایه‌سازی مانند دسته‌بندی و مرتب‌کردن، بررسی ارجاعات متقابل را انجام می‌دهد و به شما امکان می‌دهد که بر روی ایده‌ها و افکاری که در متن توسعه یافته‌اند متمرکز شوید.

¹ Windex

بسیاری از ناشران از «سیندکس» برای نمایه‌سازی در محل و ویرایش و ساختار دهی نمایه‌ها از منابع بیرون از سازمان بهره می‌برند. انعطاف‌پذیری و گزینه‌های قالبی که در این سیستم وجود دارد به نمایه‌سازان امکان می‌دهد نمایه‌ها را مطابق با نیازهای ناشران مختلف تهیه نمایند. ویژگی‌های نمایه‌سازی روزنامه‌ها و مجموعه‌های خاص کتابخانه‌ها نیز در این سیستم گنجانده شده است (Indexer Research 2010).

«ماک‌رکس»¹: نمایه‌سازی در حجم وسیع

«ماک‌رکس» برنامه‌ای برای نمایه‌سازی ماشینی است که بیش از 27 سال است که در حال توسعه و بهبود می‌باشد (Calvert and Calvert 2010). بسیاری از نمایه‌سازان مشهور جهانی از این سیستم استفاده می‌کنند. این برنامه برای هر سطح و میزان از نمایه‌سازی تعریف شده است و در حجم‌های بالا بسیار خوب عمل می‌کند. «مارک‌رکس» ویرایش 8 با محیط‌های ویندوز «اکس‌پی» و «ویستا» بسیار سازگار است. از جمله ویژگی‌های مهم این سیستم می‌توان به ایجاد شناسه‌ها، فهرست مستندات، بررسی دقیق مراحل، صرفه‌جویی در زمان، استفاده از حروف خاص، بررسی مجدد نمایه‌ها و اصلاح آن‌ها پرداخت. همچنین از طریق این برنامه می‌توان نمایه‌های دیگر را وارد² یا از سیستم خارج نمود. این سیستم قابلیت نمایه‌سازی میلیون‌ها رکورد را بدون کاهش سرعت پردازش، و نیز همراه با یکپارچگی سیستم، دارا است.

«اسکای ایندکس»³

«اسکای ایندکس» در همایش سالانه انجمن نمایه‌سازی آمریکا⁴ در «پورتلند» رونمایی شد (Sky Software 2010). ویرایش 7 این نرم‌افزار نمایه‌سازی بیش از 7 سال است که توسعه، ارتقا و بهبود یافته و ویژگی‌های منحصر به فردی را دارا است. از جمله ویژگی‌های این سیستم می‌توان به موارد زیر اشاره کرد:

¹ Macrex

² import

³ Sky Index Professional

⁴ Amercian Indexing Society

1. بیش از 6 سطح برای سرعنوانهای موضوعی (6 تورفتگی)؛
2. بیش از 9 برچسب بخش برای مکان‌یاب‌ها¹ (برچسب‌های فصل/جلد)؛
3. امکان تعریف پیشوندها و پسوندهای پیش‌فرض برای همه مکان‌یاب‌ها و برای همه برچسب‌های بخش‌ها؛
4. امکان تعریف جداسازهای الفبایی رایج و تعریف حروفی که در هر بخش از نمایه نمایش داده می‌شوند.
5. تنظیم تورفتگی‌های به صورت اینچ، میلیمتر، سانتی‌متر، نقطه، و غیره.
6. کنترل بیشتر بر مرتب‌سازی (شما می‌توانید برای سیستم تعریف کنید که چه حروفی معادل فاصله است، چه حروفی نشانه محسوب می‌شوند، و چه حروفی به عنوان فهرست حذفی معین شده است).

«دکستر»^۲

- «دکستر» برای نمایه‌سازی مستقیم در واژه‌پردازهایی مانند «مایکروسافت ورد» طراحی شده است (Editurium 2010). «دکستر» محدودیت‌های برنامه‌های نمایه‌سازی دیگر و نیز برنامه‌های واژه‌پرداز را ندارد. همچنین، با این برنامه شما:
1. شما به عنوان نمایه‌ساز نیاز به ورود، ویرایش، یا توجه به شناسه‌های افزوده ندارید.
 2. نشانه‌هایی که برای شناسایی گستره وسیعی از متن‌ها نیاز است به طور خودکار در متنی که انتخاب نموده‌اید اضافه می‌شود.
 3. شناسه‌های نمایه در هر زمانی که بخواهید به صورت قابل ویرایش و به صورت جدول ترتیبی در اختیار شما است.
 4. شما به طور مستقیم در جدول‌های نمایه، نمایه‌ها را ایجاد یا ویرایش می‌کنید به جای آنکه در جدول‌های تودرتو و مدرکی که قرار است نمایه شود کار کنید.

¹ Locators

² DEXter

5. شناسه‌ها تا آنجا که همه آن‌ها ایجاد یا ویرایش نشود وارد سیستم نمی‌شوند.
6. شناسه‌های نهایی می‌توانند به صورت حرف به حرف یا کلمه به کلمه مرتب شوند.
7. می‌توان سرعنوانهای موضوعی را به صورت الفبایی یا شماره صفحات مرتب کرد.
8. می‌توان کلمات اضافی (مانند در، به، برای، که توسط شما معین شده است) را در مرتب‌سازی سرعنوانهای موضوعی نادیده گرفت.
9. کلمات یا حروفی که توسط شما تعریف می‌شوند این امکان را دارند که به عنوان حروف یا کلمات دیگر به سیستم اضافه و مرتب شوند.
10. اعدادی مانند 80806 و 8008808 به درستی مرتب می‌شوند.
11. ارجاعات متقابل به طور خودکار ترکیب می‌شوند و به صورت الفبایی دسته بندی می‌شوند.

۵. نیازمندیهای نمایه‌ساز ماشینی برای
زبان فارسی

آنچه تا به حال در مورد نمایه‌سازی ماشینی مورد بحث قرار گرفت، مروری بر تاریخچه، مفاهیم نظری و پروژه‌های موفق انجام شده در این حوزه بود. در این فصل به نیازمندی‌ها و ویژگی‌های زبان فارسی که ممکن است بر نمایه‌ساز ماشینی تأثیر بگذارد پرداخته خواهد شد. ابتدا به تفاوت‌های نمایه‌سازی به زبان فارسی با سایر زبان‌ها و ویژگی‌های نمایه‌سازی فارسی پرداخته خواهد شد و در ادامه نیازمندی‌های نمایه‌ساز ماشینی فارسی مورد توجه قرار خواهد گرفت.

5-1. تفاوت‌های نمایه‌سازی به زبان فارسی با سایر زبان‌ها

نیاکان در کتاب "نمایه‌سازی ماشینی" ویژگی‌های زبان فارسی جهت نمایه‌سازی ماشینی را در دو بخش بیان می‌دارد:

ویژگی‌هایی زبان فارسی که به سود نمایه‌سازی است:

نبودن حرف تعریف در زبان فارسی

نبودن عامل جنسیت در زبان فارسی

نبودن جمع شکسته در زبان فارسی

وجود کسره اضافه در زبان فارسی که شکل اضافه را بسیار ساده کرده است.

صفت و موصوف و عدد و معدود در فارسی با هم مطابقت ندارند. یک صفت می‌تواند به چندین موصوف برگردد.

امکان تبدیل حروف اضافه، در برخی موارد، به کسره اضافه. برای نمونه می‌توان جمله «ایران در پیش از اسلام» را تبدیل کرد به «ایران پیش از اسلام»، بی‌آنکه تغییر در مفهوم داده شود.

در برخی از موارد اعداد به جای اینکه در آغاز عبارت بیایند، در آخر عبارت واقع می‌شوند، مثل حواس ده گانه. این ویژگی سبب می‌شود که عبارت دست‌نخورده بماند و زبان طبیعی به کار گرفته شود.

وندهای فراوان فارسی که به صورتهای پیشوند، پسوند و میانوند به همراه واژه‌ها می‌آیند و معانی تازه‌ای می‌آفرینند.

از ویژگی‌های دیگر زبان فارسی، می‌توان از خاصیت ترکیب‌پذیری واژه‌ها با هم نام برد.

در ادامه کتاب، نیاکان مشکلات زبان فارسی در نمایه‌سازی را در سه بخش دستور زبان، واژه‌های زبان و صوتهای زبان تقسیم می‌کند. او بیان می‌دارد که بخش مهم که از مسئله زبان در نمایه‌سازی مطرح است دستگاه دستوری یا همان بخش دستور زبان است و دو بخش دیگر ارتباط زیادی با این بحث ندارند.

اما مشکلات دستور زبان از نگاه نیاکان در سه مورد بیان شده است.

اول اینکه میان هنجاری گفتار و نوشتار زبان فارسی تفاوت‌های زیادی وجود دارد. یعنی مردم در گفتار واژه‌ها یا الگوهای بکار می‌برند که در نوشتار نمی‌آورند.

مشکل دوم اضافات در زبان فارسی است که عبارتند از نسبت میان دو کلمه یا گروهی از کلمات با هم، که نوعی از رابطه لفظی یا معنوی را معلوم می‌دارد و در فارسی سه نوع اضافه ملکی، تخصیصی و بیانی وجود دارد.

مشکل سوم املاء و رسم‌الخط کلمات است. این مشکل شامل مواردی مختلفی است که در برخی از آنها نیز راه‌حلهایی در حوزه بازیابی و ذخیره‌سازی اطلاعات اندیشیده شده است. نیاکان این موارد و راه‌کارهای اندیشیده شده را اینگونه بیان می‌کند.

- جدا یا پیوسته‌نویسی: مانند یک‌طرفه و یک‌طرفه. حل این مشکل از طریق بی‌فاصله‌نویسی میسر شده است.

- واژه‌های ترکیبی: مانند منابع آب و مکانیک خاک. برای رفع این مشکل برخی نرم‌افزارهای فارسی در هنگام ذخیره‌سازی به ایجاد پیوندهای ساختگی میان کلمات پرداخته‌اند.

- شیوه املاء: این مورد در مورد واژه‌هایی است که به دو صورت قابل نوشتن هستند، مانند جستجو و جست و جو. این مشکل نیز می‌تواند همانند مورد قبلی با ایجاد پیوندهای ساختگی بین این دو واژه در نرم‌افزار حل شود.

- جایگاه الفبایی واژه‌ها: برخی کلمات که از زبان‌های دیگر وارد زبان فارسی شده‌اند مشکلاتی ایجاد می‌کنند. حتی در مورد زبان عربی با وجود پیوندی که با زبان فارسی دارد. به عنوان نمونه کلمه سوال که به صورت سؤال هم نوشته می‌شود.

- آواها: زبان فارسی در نشان دادن صداها ناتوان است. در این مورد می‌توان از ذخیره کلمات پربسامد و یا طراحی قواعد استفاده نمود.

داوودآبادی در پایان‌نامه کارشناسی ارشد خود با نام "پردازش معنایی جملات و اجرای دستورات صادر شده به زبان فارسی" به مسائل و پیچیدگی‌های درک متون فارسی اشاره می‌کند. با وجود اینکه هدف از پژوهش حاضر درک معنایی زبان نیست لیکن موارد اشاره شده در این مستند از برخی ابعاد می‌تواند مفید باشد.

مسائل مربوط به ریخت‌شناسی

به بیان داوودآبادی زبان فارسی از ساختار ریخت‌شناسی نسبتاً پیچیده‌ای برخوردار است این ساختار شامل تعداد زیادی پسوند و تعداد کمی پیشوند است و از یک مدل ترتیب‌دهی واحدهای کلمه تقریباً منظمی برخوردار است. ریخت‌شناسی صرفی افعال فارسی قوی و با سیستم پیچیده‌ای از ترکیبات مشخص شده است. به این دلیل که یک شکل ظاهری می‌تواند واحدهای معنی متفاوتی را نشان دهد مشکلاتی در تحلیل ریخت‌شناسی متون فارسی پیش می‌آید. به منظور ابهام‌زدایی از چنین مواردی تجزیه‌گر نیاز به استفاده از اطلاعاتی درباره اقسام کلمه و اطلاعات نحوی از واحدهای لغوی که پسوند یا پیشوند در آن ظاهر شده، دارد. همچنین تحلیل‌گر ریخت‌شناسی باید قادر باشد که همه اشکال مختلفی که وندها در جمله ظاهر می‌شوند را تشخیص دهد.

مسائل مربوط به نحو زبان فارسی

مسائل مختلفی در زمینه نحو زبان فارسی و ساختار گرامری جملات آن وجود دارد که در زیر به مهمترین آنها می‌پردازیم.

الف-بی ترتیب بودن زبان

اگرچه فارسی دارای ترتیب مرکزی فاعل-مفعول-فعل است ولی دارای استثنائات زیادی در ترتیب کلمات می‌باشد. این استثنائات در زبان گفتاری بسیار بیشترند تا جایی که عملاً فارسی را به یک زبان بی ترتیب تبدیل می‌کنند و در زبان نوشتاری نیز به شکل محدودتر وجود دارند.

ب-تعیین حدود گروههای اسمی در جمله

یکی از مسائل مهم در تحلیل محاسباتی متون نوشته شده فارسی تعیین حدود عبارات به ویژه حدود گروههای اسمی است. تعیین حدود عبارات سخت است چون فارسی زبانی است که در آن فعل در انتها می‌آید و هیچ علامتی برای تشخیص فاعل و مفعول در یک جمله وجود ندارد (بجز نشانه مفعولی را).

ج-چند معنایی و چند نقشی بودن کلمات

برخی لغات مانند کلمه "شیر" دارای چند معنی هستند که با توجه به متنی که کلمه در آن واقع می‌شود معنی آنها مشخص می‌گردد. بعضی کلمات نیز علاوه بر چند معنی دارای چند رده نحوی یا

نقش دستوری می‌باشند. به عنوان مثال کلمه "سر" در عبارت "سر کلاس" حرف اضافه و در عبارت "سرم درد گرفت" اسم است. این موارد تعیین معنی و نقش صحیح یک کلمه در جمله را با مشکل همراه می‌سازد.

د- حذف کلمات و عبارات به قرینه لفظی یا معنوی
در بسیاری موارد کلمات یا عباراتی در یک جمله حذف می‌شوند و شنونده باید با تکمیل بخش‌های حذف شده معنی عبارت را در ذهن خود بازسازی نماید. این حذف‌ها ممکن است به قرینه لفظی یا معنوی صورت گیرد.

ه- عدم وجود ابزارها و لوازم پردازش زبان طبیعی برای زبان فارسی
متأسفانه در پردازش زبان فارسی ما با کمبود دانش زبانی مدون روبرو هستیم. طبیعتاً این مسئله در مورد دانش زبانی که به صورت محاسباتی تدوین شده باشد شدیدتر است. همچنین ابزارهای موجود برای پردازش زبان عموماً برای فارسی (مانند تجزیه گره‌های نحوی) بر روی زیر مجموعه کوچکی از زبان و تحت شرایط خاص عمل می‌کنند. لذا در پردازش زبان فارسی مشکل عمده فقدان لوازم و دانش‌های اولیه مانند دستور زبان محاسباتی، قواعد ساختارهای مدون، الگوهای نحوی و معنایی زبان، تجزیه گر کامل و عمومی برای تحلیل نحوی جملات زبان و واژگان حاوی کلمات و عبارات زبان فارسی می‌باشد. به خصوص یک مشکل اساسی در تجزیه جملات فارسی عدم ارائه گرامر مستقل از متن بدون ابهام برای فارسی است.
همچنین داوودآبادی به مسائل مربوط به تحلیل معنایی جملات اشاره می‌کند که از دامنه این پژوهش بیرون است.

2-5. نیازمندی‌های نمایه‌ساز ماشینی به زبان فارسی

طبیعی است که تمامی مواردی که پیشتر در مورد تفاوت‌های زبان فارسی و نمایه‌سازی به زبان فارسی به آنها اشاره شد نکاتی هستند که می‌بایست در طراحی یک نمایه‌ساز ماشینی مورد توجه قرار گیرند. اما با توجه به اینکه هدف از این پژوهش طراحی مدل مفهومی است برخی از جزئیات تاثیر در اجزاء و فرایندهای کلی این سیستم نخواهد گذاشت و صرفاً در لایه‌های جزئی‌تر طراحی می‌بایست مورد توجه قرار گیرند. در ادامه مواردی را که بر اجزاء اصلی و فرایندهای سیستم نمایه‌سازی به زبان فارسی به زبان فارسی موثر است مورد بررسی قرار خواهد گرفت.

همانطور که پیشتر در فصل دوم ارائه گردید فرایند متداول انتخاب اصطلاحات نمایه‌ای زبان

طبیعی از متون که همه محتوای آن را باز می‌نماید دارای مراحل زیر است (Salton 1989, 303):

1. شناسایی کلمات مفرد متن، که تحلیل واژگانی¹ نامیده می‌شود.
2. جابجایی یا حذف کلمات اضافی و اصطلاحاتی که به کرات در متن تکرار شده‌اند.
3. کاهش اختیاری کلمات باقی مانده به شکل ریشه‌ای آنها، که فرایند ریشه‌گیری² نام دارد.
4. شکل‌دهی اختیاری عبارتها به عنوان اصطلاحات نمایه‌ای.
5. جانشینی اختیاری کلمات، ریشه‌های کلمات، یا عبارتهای مربوط به آنها.
6. محاسبه وزن هر کلمه یا ریشه آن، یا اصطلاحات موجود در رده مورد نظر در اصطلاح‌نامه، یا اصطلاح عبارتی.

بررسی‌های انجام شده در فصل سوم و مرور پروژه‌های موفق نشان داد که در غالب موارد سیستم‌های تجاری موجود از تمامی این قدم‌ها یا برخی از آنها بهره می‌برند. به عنوان مثال در کی‌پی‌ماینر شناسایی اصطلاحات کلیدی نامزد و اعمال لیست حذفی همان تحلیل واژگانی و حذف کلمات است. البته در این روش اشاره‌ای به جایگزین اصطلاحات با ریشه‌ی آنها نشده ولی وزن‌دهی بر اساس فراوانی اصطلاح در مدرک مورد نظر و فراوانی آن در کل مدارک محاسبه می‌گردد. علاوه بر این اشاره شد که سفارشی نمودن این سیستم برای زبان عربی با جایگزینی فهرست حذفی و ریشه‌گیر به زبان عربی ممکن شد. با توجه به این مورد می‌توان نتیجه گرفت که فرایند کلی و یا مراحل اصلی نمایه‌سازی ماشینی در تمامی زبان‌ها یکی است و صرفاً می‌بایست زیرسیستم‌ها برای آن زبان سفارشی‌سازی شوند. البته کارایی سیستم می‌بایست پس از اعمال تغییرات بررسی شده و در صورت نیاز سیستم تنظیم³ گردد.

در اینجا این سوال مطرح می‌گردد که با وجود سیستم‌های متعددی که در فصل سوم به آنها پرداخته شد، آیا نیازی به ایجاد یک سیستم نمایه‌ساز ماشینی برای زبان فارسی وجود دارد؟ برای پاسخ به این سوال می‌توان مسئله را از ابعاد گوناگونی بررسی کرد.

- امکان سفارشی‌سازی سیستم‌های موجود. برای اینکه بتوان سیستم‌های نمایه‌سازی فعلی را برای زبان فارسی سفارشی کرد اولین شرط باز بودن این سیستم‌هاست بدین معنی که قابلیت ایجاد تغییرات در سیستم‌های آنها وجود داشته باشد. در مورد سیستم‌های تجاری نیز قطعاً

¹ lexical analysis

² stemming

³ Tuning

باید توافقاتی با تولید کننده صورت بگیرد که در این مورد فاکتور هزینه اهمیت خواهد یافت.

- انعطاف سیستم‌های موجود برای تطابق با زبان فارسی. نکته بعدی برای پاسخ به سوال بالا انعطاف این سیستم‌ها برای تطابق با شرایط خاص زبان فارسی است. از آنجا که نمایه‌ساز ماشینی سیستمی است که می‌بایست به مرور زمان بالغ شود و این بلوغ وابسته به تنظیم و تغییر تمامی زیرسیستم‌هاست، صرف قابلیت سفارشی‌سازی زبان برای یک سیستم کافی نیست.
- قابلیت توسعه‌های آتی. علاوه بر این ممکن است برای ارتقاء سیستم نیاز به اضافه کردن یک زیر سیستم جدید یا پیاده‌سازی الگوریتم‌هایی باشد که تا به حال مورد استفاده قرار نگرفته است. به عنوان مثال تکنیک‌های هوش مصنوعی که به سرعت در حال فراگیر شدن و تکامل می‌باشند از این مواردند. استفاده از این سیستم‌ها ممکن است باعث ایجاد قابلیت یادگیری در سیستم نمایه‌ساز ماشینی در آینده گردد.

با توجه به دلایل اشاره شد ممکن است ایجاد یک نمایه‌سازی ماشینی بر اساس سیستم‌های موجود به عنوان یک راه حل مقطعی قابل ارائه باشد لیکن بسیاری از قابلیت‌های متصور بر یک نمایه‌ساز ماشینی بالغ را محدود خواهد کرد. بنابراین توسعه یک سیستم نمایه‌سازی ماشینی مختص زبان فارسی با معماری دقیق و زیرسیستم‌های قابل تغییر و توسعه قطعاً روش مناسب‌تری خواهد بود.

5-3. اجزاء نمایه‌ساز ماشینی فارسی

مواردی را که سالتون به عنوان اجزاء اصلی یک نمایه‌ساز ماشینی در سال 1989 مطرح کرد مبنایی برای سیستم‌های تجاری امروزی قرار گرفته و بررسی آنها نشان می‌دهد که این چارچوب برای یک مدل کلی نمایه‌سازی ماشینی مناسب است. البته اینکه ماهیت این اجزاء چه خواهد بود کاملاً به خصوصیات زبان مرجع وابسته خواهد بود.

بر همین اساس مبنای مدل مفهومی نمایه‌ساز ماشینی زبان فارسی را اجزاء ارائه شده توسط سالتون قرار داده و در ادامه خصوصیات هر جزء و فرایندهای بین آنها تشریح خواهد شد.

5-3-1. زیرسیستم مبدل دیجیتال

این زیر سیستم در واقع ورودی نمایه‌ساز ماشینی است. در صورتی که اسناد از ابتدا دیجیتالی باشند که مستقیماً وارد مراحل بعدی می‌شوند و در غیر این صورت ابتدا دیجیتالی شده و سپس برای تجزیه کلمات و مراحل بعدی وارد نمایه‌ساز می‌گردند.

5-3-2. زیرسیستم تحلیل واژگانی

این زیرسیستم متن را به کلمات تفکیک نموده و ماهیت هر کلمه را تشخیص خواهد داد. زیرسیستم شامل تشخیص نوع کلمه، شناسایی افعال، الفاظ و اصطلاحات می‌باشد. در ابتدا واژه شکن زنجیره کلمات را تفکیک کرده و نوع آن را تشخیص خواهد داد. در حالات خاص مثل اختصارات و یا اسامی خاص (مانند اسامی سازمان‌ها)، سیستم کلمه مورد نظر را با جداول مربوطه تطبیق خواهد داد. همچنین از آنجا که یکی از منابع خوب برای انتخاب کلمات کلیدی پیشنهادات نویسنده است، سیستم تحلیل واژگانی کلمات پیشنهادی نویسنده را نیز به عنوان نامزد کلمه کلیدی به لیست کلمات اضافه خواهد نمود.

5-3-3. زیرسیستم گزینش عبارت‌ها و اصطلاحات و نرمال‌سازی

این زیرسیستم رابط بین نمایه‌ساز ماشینی و اصطلاحنامه است. بدین معنی که اصطلاحات و واژه‌های استخراج شده از متن با واژه‌های استاندارد مطابقت داده شده و واژه نرمال جایگزین خواهد شد. علاوه بر این ارتباط بین نمایه‌ساز ماشینی و اصطلاحنامه یکطرفه نیست و در مواردی اصطلاحاتی که در حال حاضر در اصطلاحنامه وجود ندارد شناسایی شده و برای اضافه نمودن توسط سیستم پیشنهاد می‌گردد.

5-3-4. زیرسیستم فهرست حذفی

همانطور که پیشتر اشاره شد کلمات پرکاربرد و وجود دارند که نمی‌توانند جزء کلمات کلیدی انتخاب شوند (مانند حروف اضافه). لیست این کلمات (فهرست حذفی) در این زیرسیستم وجود دارد و زیرسیستم آنها را از کلمات نامزد کلیدواژه حذف خواهد کرد. علاوه بر این بر اساس نظر بالرینی¹ کلمات که عدد آستانه مربوط به حرف را دارا می‌باشند حذف می‌گردند و در این حالت برای از دست ندادن کلمات کوتاه که ارزش نمایه شدن را دارند از لیست کلمات ضدحذفی استفاده می‌گردد. دلیل انتخاب این روش آن است که این فرایند حجم لیست نامزد نمایه را به صورت قابل توجهی کاهش خواهد داد.

5-3-5. زیرسیستم ریشه‌گیری کلمات

همانطور که اشاره شد یکی از تکنیک‌هایی که کیفیت نمایه‌سازی را بالا می‌برد ریشه‌گیری کلمات است. در این زیرسیستم کلمات به دو نوع فعل و لفظ یا واژه تقسیم می‌گردند. ریشه‌گیری در پی این هدف است که تطابق بین اصطلاحات نمایه‌ای پرسش و متن مدرک را بهبود بخشد. شانس تطابق

¹ Ballerini

هنگامی افزایش می‌یابد که اصطلاحات نمایه‌ای به ریشه‌هایشان تجزیه شوند. به عقیده سالتون ریشه-گیری جامعیت را افزایش داده و دامنه اصطلاحات کلیدی را در متن گسترش خواهد داد. همچنین ابزار توسعه‌ای به نام تجزیه و تحلیل معناشناختی برای این زیرسیستم در نظر گرفته شده تا کلمات با چندمعنی و یا جملات بدون بارمعنایی شناخته شوند. همچنین بررسی معناشناختی می‌تواند متون با معانی مشترک را گروه‌بندی کند تا با استفاده از این ارتباط معنایی بتوان آنها را بازیابی کرد.

5-3-6. زیر سیستم وزن‌دهی به اصطلاحات

پس از اینکه اصطلاحات و کلمات نامزد نمایه انتخاب شدند، طبیعی است که نمی‌توان همه آنها را به عنوان نمایه‌های نهایی انتخاب کرد. همچنین همه آنها نمی‌توانند نمایه‌هایی گویا باشند. از اینرو مکانیزمی مورد نیاز است که نمایه‌ها را اولویت‌بندی کرده و تعدادی از آنها را به عنوان کلیدواژه‌های نهایی انتخاب کند. این زیرسیستم به هر اصطلاح یک وزن عددی تخصیص می‌دهد که وابسته به فراوانی اصطلاح در مدرک و همچنین وابسته فراوانی این اصطلاح در سایر مدارک است. شاید بتوان گفت مهمترین بخش یک نمایه‌ساز ماشینی بخش وزن‌دهی است که می‌بایست به صورت رفت و برگشتی تنظیم شود. این به این معنا است که پس از استفاده از یک الگوریتم وزن‌دهی کیفیت آن بر اساس مدارک بازیابی شده تنظیم شود. در صورتی که الگوریتم جوابگوی نیاز سیستم نیست تغییر کند و یا مقادیر فاکتورهای آن تنظیم شوند. الگوریتم‌های هوش مصنوعی و یادگیری ماشینی بیشترین نقش را می‌توانند در این زیرسیستم بازی کنند.

5-4. تاثیرات زبان فارسی در اجزاء نمایه‌ساز فارسی

همانطور که اشاره شد فرایندهای حاکم بر نمایه‌سازی ماشینی و حتی بسیاری از زیرسیستم‌ها و الگوریتم‌ها از یک الگوی نسبتاً ثابت پیروی می‌کنند و خصوصیات زبانی بیشتر بر جزئیات و یا داده‌های ورودی به این الگوریتم‌ها و یا طراحی قواعد تاثیرات خود را اعمال می‌کنند. زبان فارسی نیز از این قاعده مستثنا نیست و تفاوت‌های زبانی که پیشتر بر آنها اشاره شد بر ساختار و معماری یک نمایه‌ساز ماشینی تاثیر ندارد بلکه جزئیات، ورودی‌ها و تنظیمات را تغییر خواهد داد. همچنین ممکن است توالی فرایندها تغییراتی بکند. از آنجا که هدف این پژوهش ورود به جزئیات اجزاء یک نمایه‌ساز ماشینی نیست، نمی‌توان تمامی این تفاوت‌ها و تاثیرات آن بر نمایه‌ساز ماشینی را در نظر گرفت لیکن به برخی از آنها مختصراً اشاره می‌گردد.

5-4-1. مسائل ریخت‌شناسی

همانطور که اشاره شد زبان فارسی از ساختار ریخت‌شناسی نسبتاً پیچیده‌ای برخوردار است. تعداد زیاد پسوند و تعداد کم پیشوندها، سیستم پیچیده ترکیبات، تفاوت در واحدهای معنایی از جمله خصوصیات این ساختار است. مهمترین اثر این ساختار بر زیرسیستم تحلیل واژگانی و زیرسیستم ریشه‌گیر کلمات خواهد بود.

5-4-2. مسائل نحوی

مواردی همچون بی‌ترتیب بودن، چند معنایی بودن و حذف به قرینه لفظی و معنوی از جمله خصوصیات نحوی زبان فارسی است که بر سیستم نمایه‌ساز ماشینی تاثیر خواهد گذاشت. تاثیر اصلی این تفاوت‌ها باز هم بر زیرسیستم‌های تحلیل واژگانی و ریشه‌گیر و در برخی موارد بر زیرسیستم گزینش عبارت‌ها و اصطلاحات و نرمال‌سازی خواهد بود. زیرسیستم فهرست حذفی می‌بایست کاملاً بر اساس زبان فارسی طراحی گردد و زیرسیستم وزن‌دهی نیز می‌بایست بر اساس خصوصیات زبانی طراحی و تنظیم گردد.

ع. نایه سازی ماشین:

طراحی مدل مفهومی

6-1. متدولوژی و معماری طراحی سیستم

یکی از مهم‌ترین فاکتورهای موفقیت توسعه پروژه‌های پیاده‌سازی سیستم‌های نرم‌افزاری انتخاب صحیح متدولوژی تحلیل و طراحی آنهاست. نمایه‌ساز ماشینی که در واقع یک پروژه توسعه نرم‌افزار با پشتوانه علمی بالاست نیز از این قاعده مستثنی نیست. علاوه بر این نمایه‌ساز ماشینی به غیر از فاکتورهای معمول پروژه‌های نرم‌افزاری خصوصیات منحصر به فرد دیگری هم دارد. از جمله این خصوصیات می‌توان به توسعه تدریجی این سیستم اشاره کرد. همانطور که در بخش‌های پیشین به آن اشاره شد و در نمونه‌های موفق نمایه‌ساز ماشینی نیز بررسی شد، یک نمایه‌ساز ماشینی ممکن است از زیرسیستم‌های متفاوتی تشکیل شده باشد که وجود برخی از آنها برای سیستم الزامی و باقی زیرسیستم‌های توسعه‌ای می‌باشند. در واقع می‌توان گفت یک نمایه‌ساز ماشینی ممکن است از سطوح بلوغ متفاوتی برخوردار باشد و یا در طول زمان بالغ شود.

این خصوصیت باعث می‌شود که فرایند تحلیل به صورت یک فاز از پروژه تعریف نگردد، بلکه تحلیل، فرآیندی پیوسته خواهد بود که در تمامی مراحل بلوغ سیستم همراه مراحل دیگر است. البته این روش یعنی همراه بودن فاز تحلیل به صورت پیوسته مختص نمایه‌ساز ماشینی نیست و موضوعی است که همیشه در دنیای تجزیه، تحلیل و طراحی سیستم‌های نرم‌افزاری مطرح بوده است.

با توجه به آنچه بحث شد شاید بتوان گفت بهترین متدولوژی که می‌تواند از فرآیند توسعه سیستم نمایه‌ساز ماشینی پشتیبانی کند، متدولوژی RUP¹ باشد. دلیل این گفته را می‌توان در موارد زیر خلاصه کرد:

- در متدولوژی RUP توسعه نرم‌افزار یک پروسه پیوسته و ادامه‌دار است.
- RUP علاوه بر اینکه یک متدولوژی برای توسعه نرم‌افزار می‌باشد، یک معماری نرم-افزاری نیز هست که به اجزای اصلی به خوبی می‌پردازد.
- فرآیندهای RUP به گونه‌ای طراحی شده است که مدیریت پروژه نرم‌افزاری را تسهیل کرده و تمامی مراحل آنرا پشتیبانی می‌کند.

¹ Rational Unified Process

- متدولوژی RUP زبان مدلسازی استاندارد UML^۱ را به عنوان زبان مرجع خود معرفی کرده که مدلسازی فرآیندهای مدیریت تجزیه، تحلیل، طراحی و پیاده سازی نرم افزار را تمامی مراحل پشتیبانی می کند.

- یکی دیگر از خواص متدولوژی RUP و زبان مدلسازی UML پشتیبانی از معماری مدل رانه^۲ است. این بدین معناست که اسکلت نرم افزار، پایگاه داده ها، کلاس های برنامه نویسی و بخش قابل توجهی از نرم افزار از طریق تفسیر مدل های تولید شده به صورت خود کار تولید می شود.

با توجه به خواصی که در مورد این متدولوژی به آن اشاره برای طراحی مدل مفهومی نمایه ساز ماشینی نیز از زبان مدلسازی UML در قالب معماری RUP استفاده گردید، که در ادامه به آن پرداخته خواهد شد.

نکته ای که در مورد دیاگرام ها و مدل های مورد استفاده قابل اشاره است، ماهیت رفتاری آن- هاست. از آنجا که در اینجا ساخت مدل مفهومی و طراحی اسکلت کلی سیستم نمایه ساز ماشینی مدنظر بود، از مجموعه دیاگرام هایی استفاده شد که رفتار سیستم را مدل می کردند و در موارد اندکی به جزئیات ساختار سیستم و در واقع مدل های استاتیک و نزدیک به ساختار واقعی و نهایی نرم افزار پرداخته شده است. لیکن همانطور که گفته شد در معماری مورد استفاده این مدل ها مبنای اصلی توسعه سیستم قرار خواهند گرفت و تحلیلگر و توسعه سیستم آنرا مانند نقشه ای مبنای کار خود قرار خواهند داد.

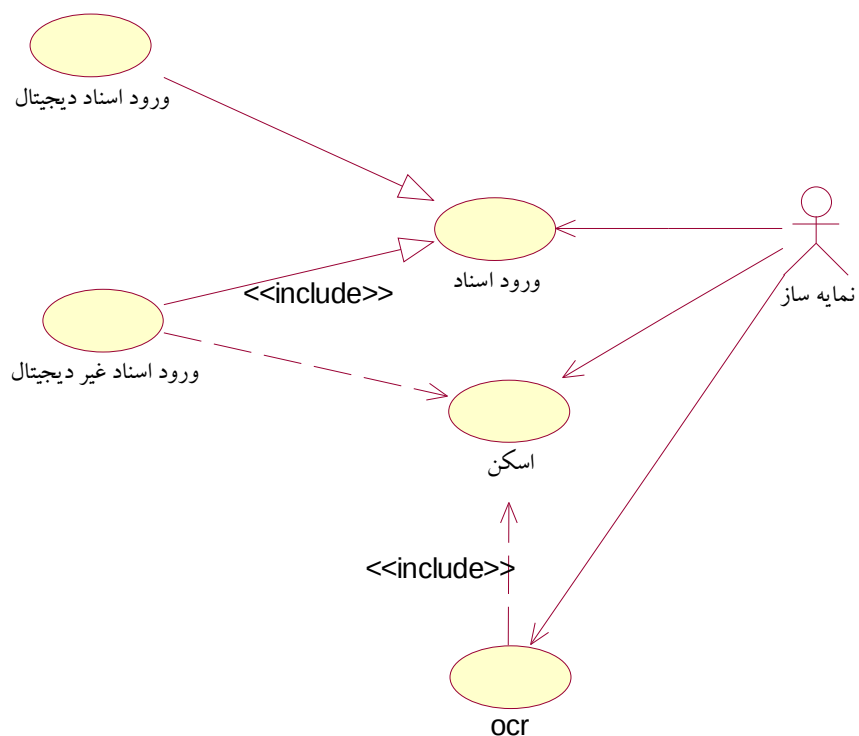
با وجود ابزارهای متعددی که از متدولوژی RUP پشتیبانی می کنند، در اینجا از نرم افزار معروف Rational Rose ساخت شرکت IBM نسخه 2003 استفاده شد. دلیل این انتخاب سادگی کار با این نرم افزار، پشتیبانی از زبان فارسی، آشنایی غالب طراحان سیستم با آن و همخوانی این نرم افزار با زبان ها و محیط های برنامه نویسی شرکت مایکروسافت همچون .NET بود، که از رایج ترین ابزارهای توسعه نرم افزار می باشد و اغلب برنامه نویسان با آن ها آشنایی کامل داشته و معمولاً برنامه های سازمانی به وسیله آن ها توسعه می یابد.

^۱ Unified Modeling Language
^۲ Model Driven Architecture

2-6. دیدگاه مورد کاربرد¹

این دیدگاه در واقع مفهومی است که به درک تحلیلگر از چگونگی رفتار سیستم کمک می‌کند. همچنین در این نما ارتباط بین کاربران و سیستم روشن می‌گردد و تا حد نیاز زیرسیستم‌ها نیز تحلیل می‌گردند.

شکل 6-1. دیاگرام مورد کاربرد² مبدل دیجیتال

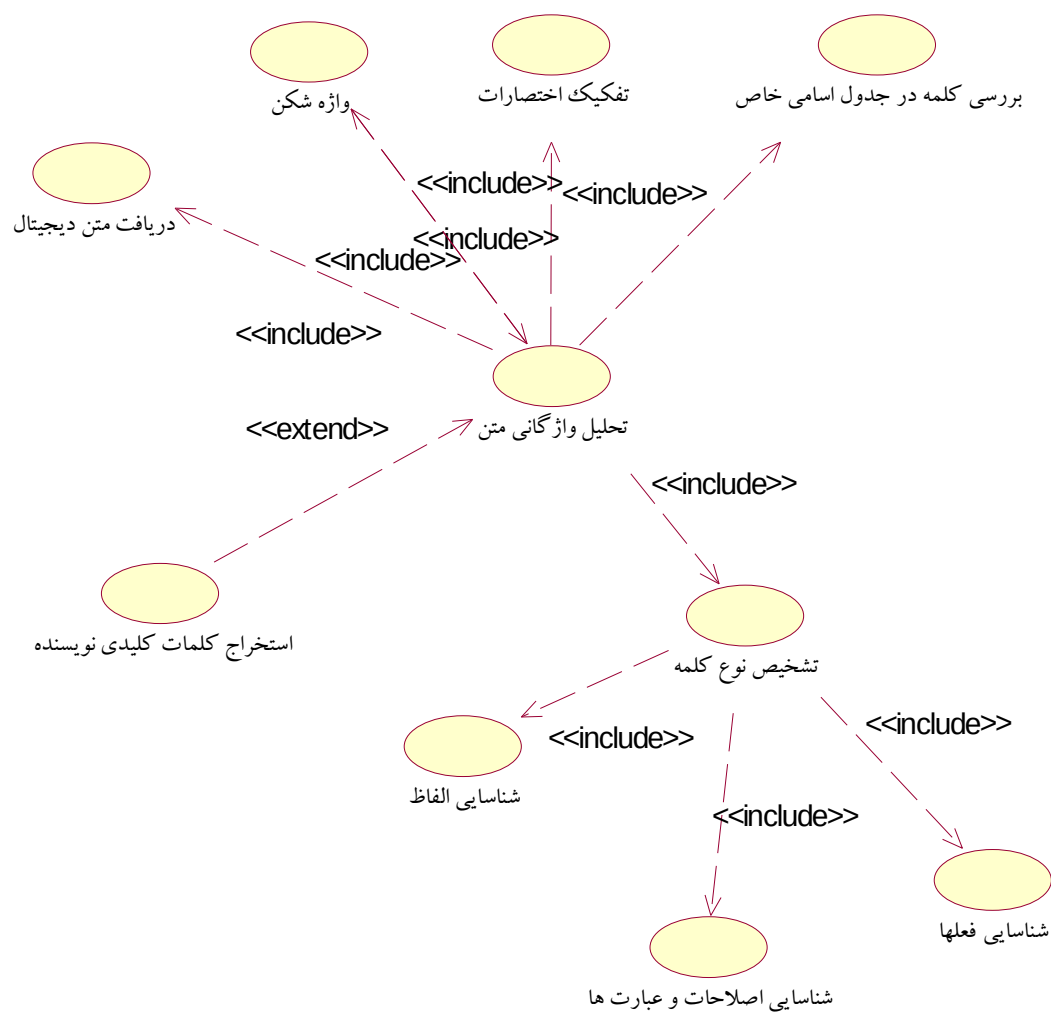


¹ UseCase View
² UseCaseDiagram

جدول 6-1. موارد کاربرد مبدل دیجیتال	
ورود اسناد	به منظور ورود مدارک به سامانه نمایه سازی ماشینی مورد استفاده فوق مورد استفاده قرار می گیرد. نمایه ساز مدارک مورد نظر خود را با استفاده از آن وارد می نماید. با توجه به نوع اسناد نمایه ساز فرآیند ورود مناسب مدارک را شناسایی نموده و مدارک را وارد سامانه می نماید.
ورود اسناد دیجیتال	در صورت الکترونیکی بودن مدارک به راحتی توسط نمایه ساز بررسی شده و وارد سامانه می گردد. فرمت های الکترونیکی باید استاندارد و در قالب متن باشد.
ورود اسناد غیر دیجیتال	در صورت غیر الکترونیکی بودن مدارک سامانه پس از اسکن ¹ نمودن تصاویر مدرک و اجرای فرآیند تشخیص متن ² مدارک را به صورت الکترونیکی تبدیل خواهد نمود.
آکتورها ³	
نمایه ساز	متصدی بررسی و ورود مدارک به سامانه نمایه سازی ماشینی می باشد.

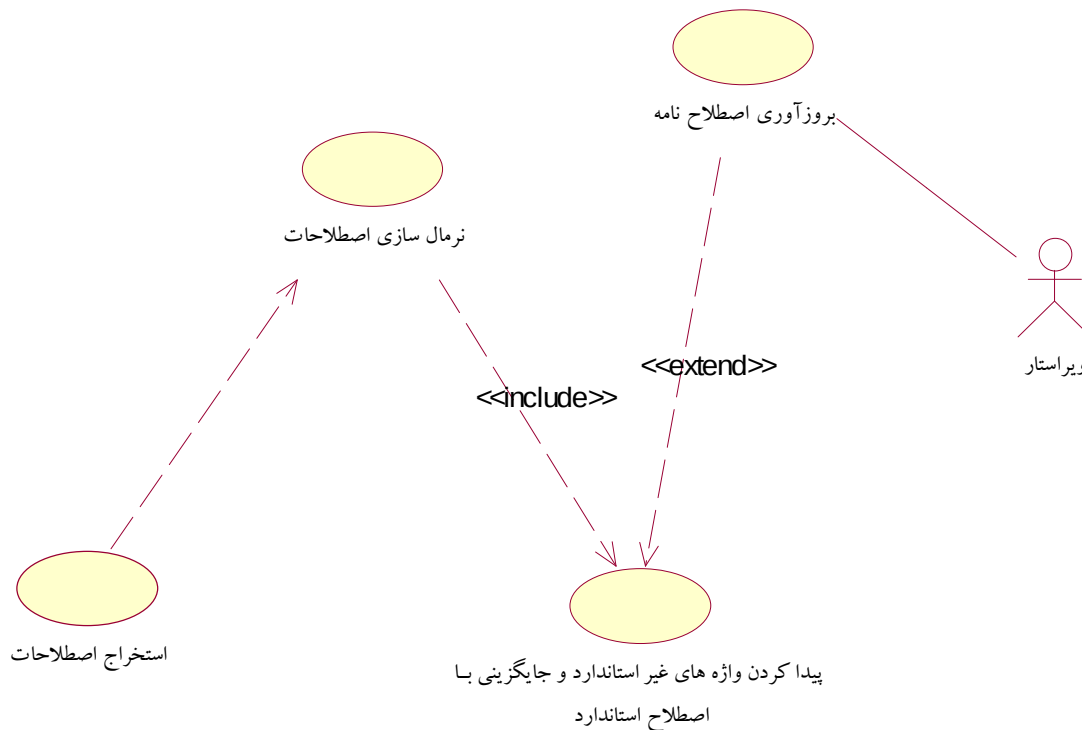
¹ Scan
² OCR
³ Actors

شکل 6-2. دیاگرام موارد کاربرد تحلیل واژگانی



جدول 6-2. موارد کاربرد تحلیل واژگانی	
تحلیل واژگانی متن	جریان ورودی مدارک را دریافت نموده و عملیات لازم به منظور تفکیک متن به کلمه‌ها را فراهم می‌نماید. یک کلمه یا اشتقاق آن به عنوان زنجیره‌ای از حروف تعریف می‌شود که با یک فاصله سفید یا / و علامت سجاوندی از هم مجزا می‌شوند.
تشخیص نوع کلمه	به منظور شناسایی و بررسی نوع کلمه مورد استفاده قرار می‌گیرد. به عنوان مثال یک کلمه می‌تواند لفظ، فعل، یا اختصار باشد. در این مرحله طبقه‌بندی مطابق با نتایج بررسی انجام می‌گردد.
شناسایی فعل‌ها	به منظور تفکیک و شناسایی کلمه‌های فعل از جریان ورودی کلمه‌ها مورد استفاده قرار می‌گیرد.
شناسایی الفاظ	به منظور شناسایی الفاظ از جریان ورودی مورد استفاده قرار می‌گیرد.
شناسایی اصطلاحات	در اینجا اصطلاحات بر اساس جدول اصطلاحات تفکیک شده و جهت استاندارد شدن در مراحل بعدی ذخیره می‌گردد.
استخراج کلمات کلیدی نویسنده	هر مدرک می‌تواند دارای یک دسته از کلیدواژه‌هایی باشد که توسط نویسنده مشخص می‌شوند که در بخش واژه‌شکنی باید وارد سامانه گردند.
دریافت متن دیجیتال	به منظور دریافت متن مورد استفاده قرار می‌گیرد. کلیه مدارک الکترونیکی توسط این بخش دریافت شده و در فرآیند تحلیل واژگانی به عنوان ورودی مورد استفاده قرار می‌گیرند.
واژه‌شکن	به منظور تفکیک متن به کلمات مورد استفاده قرار می‌گیرد.
تفکیک اختصارات	اختصارات (برای نمونه، «نک») برخی مواقع گیج‌کننده می‌شوند، اگر این «نک» در آخر جمله و جزیی از یک کلمه باشد، تمیز آن از اختصارات امری ساده نیست. مورد کاری فوق برای شناسایی اختصارات بکار گرفته می‌شود.
بررسی کلمه در جدول اسامی خاص	جدول اسامی خاص، شامل اسامی سازمان‌ها، موسسات و نهادهایی می‌باشد که در نمایه‌سازی موثرند.

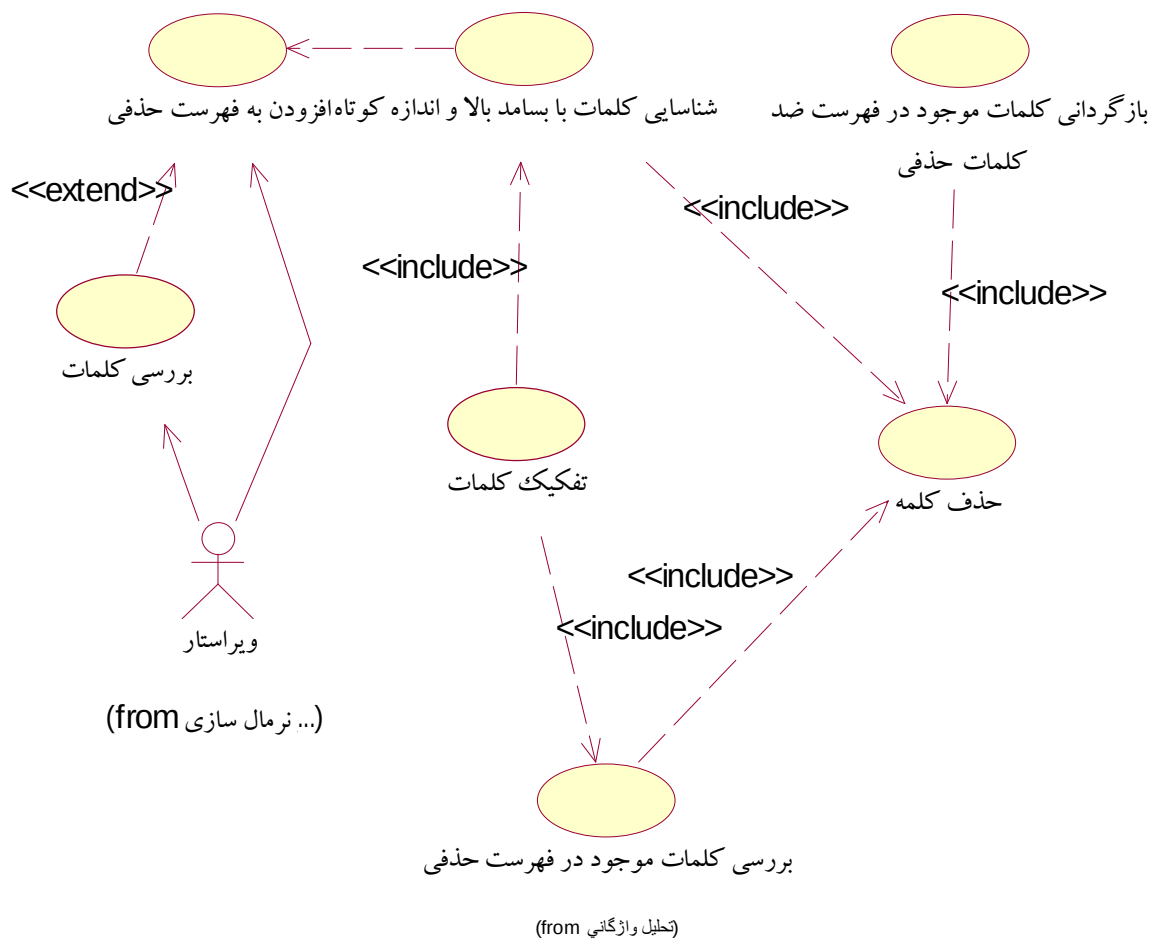
شکل 6-3. دیاگرام موارد کاربرد گزینش عبارت‌ها و اصطلاحات و نرمال‌سازی



جدول 6-3. موارد کاربرد گزینش عبارت‌ها و اصطلاحات و نرمال‌سازی	
اصطلاح نامه یک فهرست و در واقع یک پایگاه داده پویا است که از طرق این پروسه توسط ویراستار بروزآوری می‌گردد.	بروزآوری اصطلاح نامه
نرمال‌سازی اصطلاحات به معنی بررسی اصطلاحات غیراستاندارد مورد استفاده در متن و جایگزینی آن‌ها با اصطلاحات استاندارد اصطلاح نامه است. مثلاً به جای اصطلاح "حاملگی" می‌بایست از "بارداری" استفاده شود.	نرمال‌سازی اصطلاحات
این پروسه با تطبیق اصطلاحات اصطلاح نامه با کلمات متن، اصطلاحات استاندارد و غیراستاندارد موجود در اصطلاح نامه را از درون متن استخراج می‌نماید.	استخراج کلمات
این پروسه درون اصطلاح نامه صورت می‌گیرد. یعنی اصطلاح غیراستاندارد شناسایی و اصطلاح استاندارد جایگزین آن شده و به عنوان نامزد نمایه‌ها استخراج می‌گردد.	جایگزینی واژه‌های غیراستاندارد و با استاندارد

آکتورها	
ویراستار	اصطلاح‌نامه‌ها توسط ویراستار بروزآوری می‌شوند.

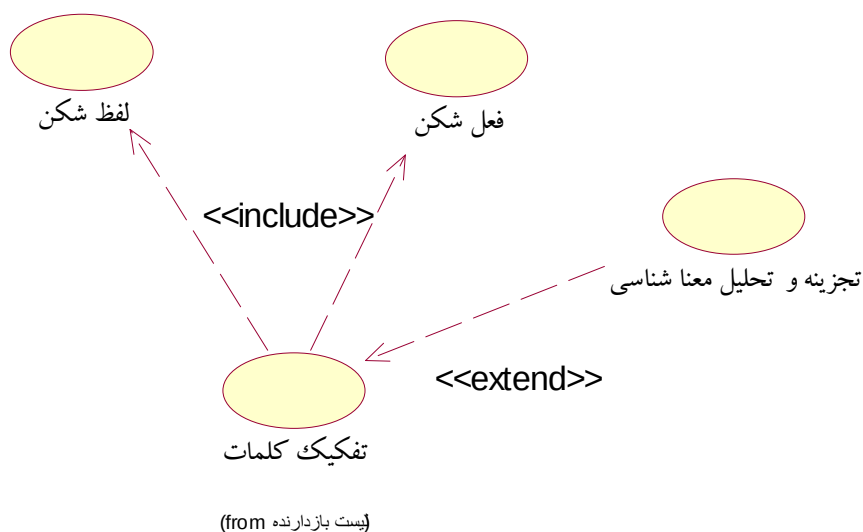
شکل 6-4. دیاگرام موارد کاربرد فهرست حذفی



جدول 6-4. موارد کاربرد فهرست حذفی	
تفکیک کلمات	در این پروسه کلماتی که در فهرست حفظی قرار دارند و یا قواعد کلنات قابل حذف در مورد آنها صدق می کند، شناسایی و تفکیک می گردند.

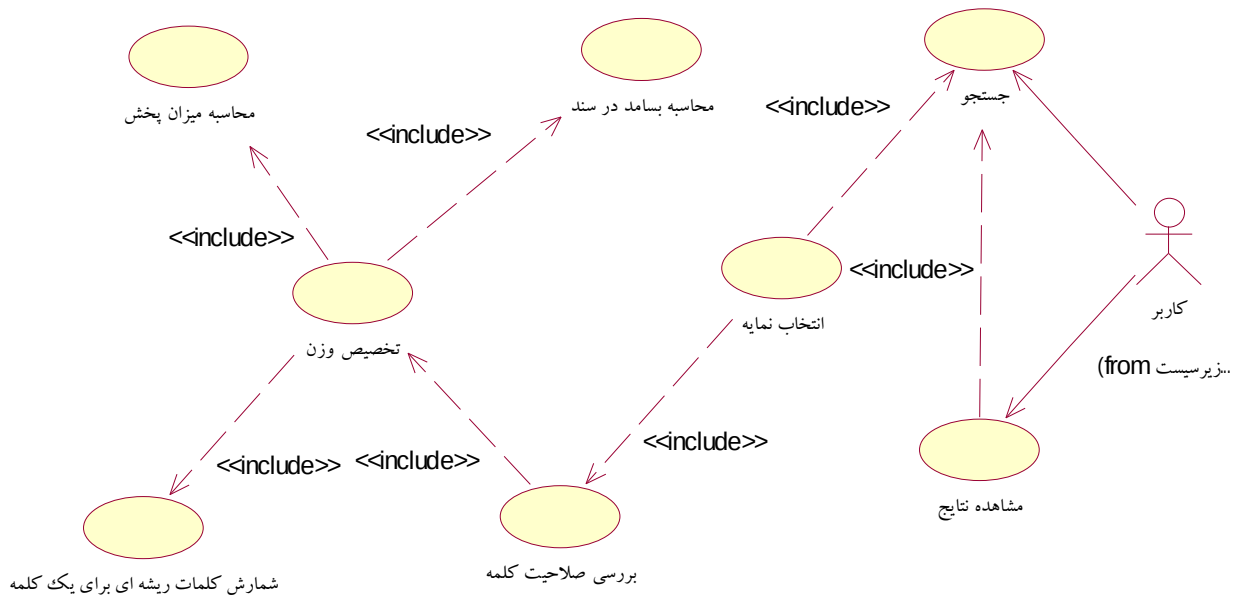
حذف کلمه	در این مرحله اصطلاحات حذفی، حذف می گردند.
بررسی کلمات	سیستم کلمات کوتاهی که نامزد فهرست حذفی هستند و در حال حاضر در این فهرست قرار ندارند را در اختیار ویراستار قرار می دهد تا در صورت تایید به این فهرست اضافه گردند.
بازگردانی کلمات موجود در فهرست ضد کلمات حذفی	به دلیل آنکه کلمات وابسته معمولاً کوتاه (دارای تنها چند حرف) هستند، گاهی همه کلمات کوتاهی که عدد ارزش آستانه مربوط به حروف را دارا می باشند از متن حذف می گردند. در این حالت، استفاده از یک فهرست «ضد کلمات حذفی»، از حذف کلمات کوتاه مهم جلوگیری می نماید. پس از انجام عملیات حذف توسط این مورد استفاده اقدامات لازم پیرامون بازگردانی کلمات عنوان شده، انجام می گیرد.
شناسایی کلمات با بسامد بالا و اندازه کوتاه	در بسامدگیری، ابتدا فهرستی از واژه های غیرمجاز تهیه و در برنامه گنجانده می شود. سامانه، متن مدارک را با این فهرست تطبیق داده واژه های غیرمجاز را حذف و بسامد واژه های باقیمانده را تعیین می کند، واژه های با بسامد بسیار بالا و بسیار کم، حذف و بقیه به عنوان موضوع تنظیم و همراه با جای نما در قالب مدخل ها ارائه می شوند.
افزودن به فهرست حذفی	یک فهرست حذفی یا واژه نامه منفی فهرستی از واژه هایی را داراست که نمی توان آن ها را به عنوان اصطلاحات نمایه ای انتخاب کرد. مورد استفاده فوق، این دسته از کلمات را توسط ویراستار به فهرست حذفی افزوده می گردند. کلماتی با بیشترین بسامد محاسبه شده در متن در فهرست حذفی گنجانده می شوند.
بررسی کلمات موجود در فهرست حذفی	فهرست حذفی شامل اصطلاحات کلمات زائدی می باشد که به معنای جملات هیچ کمکی نمی نمایند بلکه فقط در شکل دهی جملات درست به کار می روند. در این بررسی نسبت به شناسایی کلمات فوق اقدام می گردد.
آکتورها	
ویراستار	اقدامات لازم پیرامون ورود کلمات لازم به فهرست حذفی را انجام می دهد.

شکل 5-6. دیاگرام موارد کاربرد ریشه‌گیری کلمات



جدول 5-6. موارد کاربرد ریشه‌گیری کلمات	
فعل شکن	مورد کاربردی فوق در شناسایی فعل و تشخیص ریشه فعل کاربرد دارد و در مورد استفاده تفکیک کلمات مورد استفاده قرار می‌گیرد.
لفظ شکن	مورد کاربردی فوق در شناسایی الفاظ و تشخیص نوع لفظ در عملیات تفکیک کلمات مورد استفاده قرار می‌گیرد.
تجزیه و تحلیل معناشناسی	تجزیه و تحلیل معناشناسی در سطح مدارک مشترک در یک پایگاه صورت می‌گیرد. در این مورد کاربرد، ارتباطات معنایی موجود در مدارک شناسایی شده تا متون مشترک بتوانند به صورت واحدهای هم معنی تجزیه شوند. بدون دانش معناشناختی امکان ندارد کلماتی که دارای چند معنی مختلف هستند و یا جملات بدون بارمعنایی، شناسایی شوند.
تفکیک کلمات	در مورد کاربردی فوق نوع کلمه مشخص شده و کلمه براساس آن طبقه‌بندی خواهد شد. بدین منظور از دور مورد استفاده فعل شکن و لفظ شکن می‌شود و در صورت نیاز از تجزیه و تحلیل معناشناسی استفاده می‌گردد.

شکل 6-6. دیاگرام موارد کاربرد وزن‌دهی به اصطلاح نمایه

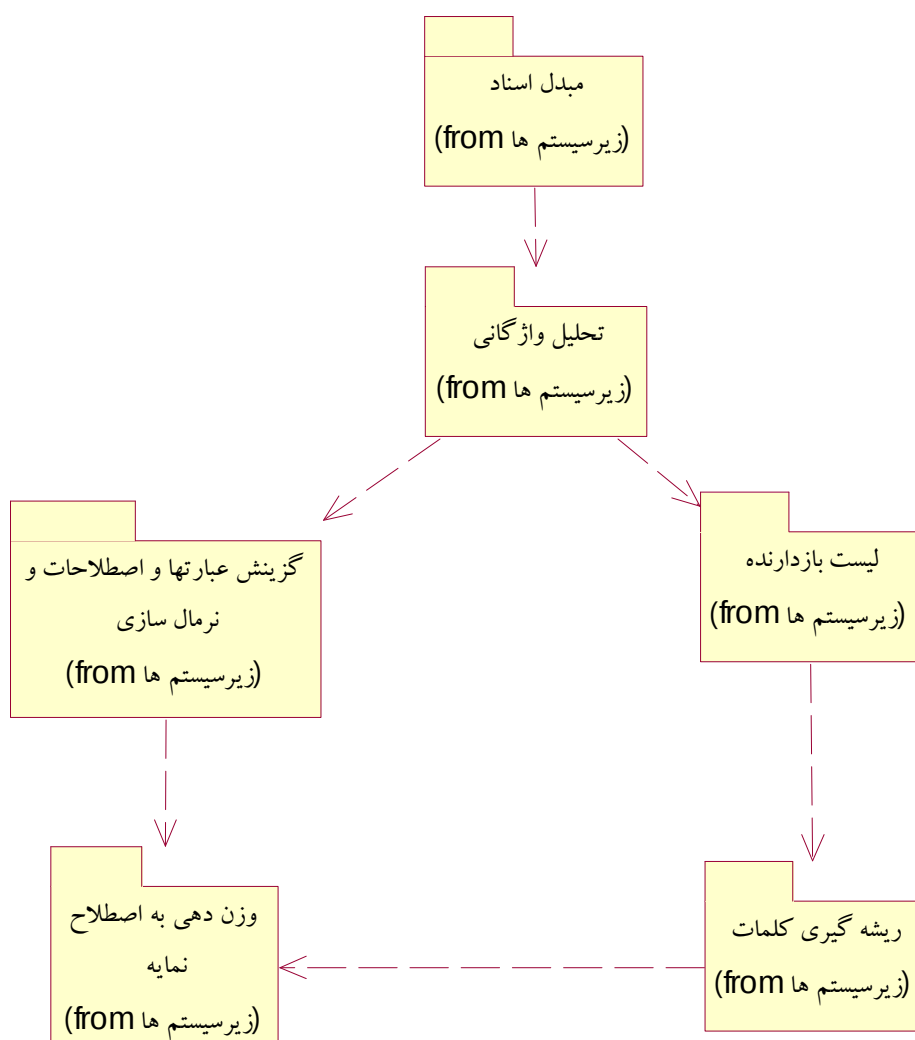


جدول 6-6. موارد کاربرد وزن‌دهی به اصطلاح نمایه	
بررسی صلاحیت کلمه	همه کلمات موجود در مدرک اصطلاحات نمایه‌ای مناسبی نیستند و کلماتی که اصطلاحات نمایه‌ای مناسب قلمداد می‌گردند نیز در تعریف محتوای یک متن نقش هم‌اندازه‌ای ندارند. با توجه به روش‌های آماری اتخاذ شده در شناسایی نمایه پس از محاسبه وزن کلمات بر مبنای محاسبه بسامد، میزان پخش و شمارش کلمات ریشه‌ای نسبت به صحت انتخاب صورت پذیرفته اقدام می‌گردد.
شمارش کلمات ریشه-ای برای یک کلمه	به منظور شمارش کلمات ریشه‌ای در یک مدرک پس از عملیات ریشه‌گیری و استخراج ریشه در متن مدرک ریشه‌شناسایی شده شمارش شده و نتایج در سامانه ثبت خواهد شد. شمارش کلمات ریشه در تخصیص وزن به کلمه موثر خواهد بود.
تخصیص وزن	وزن‌دهی به اصطلاح، برای انتخاب کلمات نمایه‌ای مناسب جهت بازنمایی متن یا

	بهبود تطابق پرسش و جامعیت در محیط بازیابی بسیار ضروری است. وزن اصطلاح نمایه‌ای معمولاً دارای یک ارزش عددی است وزن‌های اصطلاح دارای ارزش صفر یا بیشتر هستند و در مورد وزن‌های نرمال شده بین صفر و یک متغیر است. که وزن یک به اصطلاحات بسیار مهم و وزن صفر به اصطلاحا بسیار ضعیف تخصیص می‌یابد. ارزش صفر به این معنا نیز هست که اصطلاح مورد نظر دارای هیچ ارزش محتوایی نیست.
انتخاب نمایه	مورد کاربردی انتخاب نمایه، با استفاده از روش‌های احتمالی نسبت به تعیین نمایه اقدام می‌نماید. در صورت انتخاب یک کلمه به عنوان نمایه، کلمه فوق در جدول مربوط به نمایه‌های مدارک قرار خواهد گرفت.
جستجو	کاربران برای پیدانمودن مدارک مورد نیاز خود با داشتن کلیدواژه‌های مورد نظر خود در سامانه وارد شده و نسبت به استفاده از امکان جستجو اقدام می‌نمایند. با توجه به ورودی کاربر این مورد استفاده بر مبنای نمایه‌های استخراج شده از مدارک نتایج را استخراج نموده و در اختیار کاربر قرار می‌دهد.
مشاهده نتایج	در صورت هر بار استفاده از مورد کاربرد جستجو نتایج حاصل از جستجو در نمایه‌ها در اختیار کاربر قرار می‌گیرد. این فهرست شامل نمایه‌های پیشنهادی و راهنمای ارجاع به مدارک حاوی نمایه‌ها خواهد بود.
محاسبه بسامد در سند	از جمله عوامل موثر در تخصیص وزن کلمه می‌باشد. مورد کاربرد فوق، به منظور محاسبه بسامد بکار گرفته می‌شود. معمولاً واژه‌های بایساند بسیار بالا یا کم حذف می‌گردند.
محاسبه میزان پخش	از جمله عوامل موثر در محاسبه وزن کلمه می‌باشد. مورد استفاده فوق میزان پخش یک کلمه در مدرک را مورد بررسی قرارداده و نتایج را در جدول مربوط به مدرک ذخیره می‌نماید.
آکتورها	
کاربران	استفاده‌کنندگان اصلی از سامانه می‌باشند که با استفاده از فرم‌های مربوط به جستجو نسبت به ورود لغات مورد نظر خود جهت بررسی در نمایه‌های اسناد اقدام نموده و سامانه پس از بررسی در نمایه‌ها نتایج را در اختیار کاربران قرار می‌دهد.

3-6. دیدگاه منطقی¹

در این دیدگاه با توجه به دیاگرام‌ها و تحلیل‌های انجام شده در دیدگاه شماتیک موارد کاربرد، اجزای اصلی سیستم شناسایی شده و در قالب دیاگرام‌های کلاس² نمایش داده می‌شوند. در هر مرحله دیاگرام توالی³ که نحوه ارتباط بین کلاس‌ها و دیاگرام همکاری که نمایش متفاوتی از همان دیاگرام توالی است نیز ترسیم شده است. شکل کلی زیر نحوه تعامل و توالی کلی زیرسیستم‌های طراحی شده در فاز قبل را نشان می‌دهد.



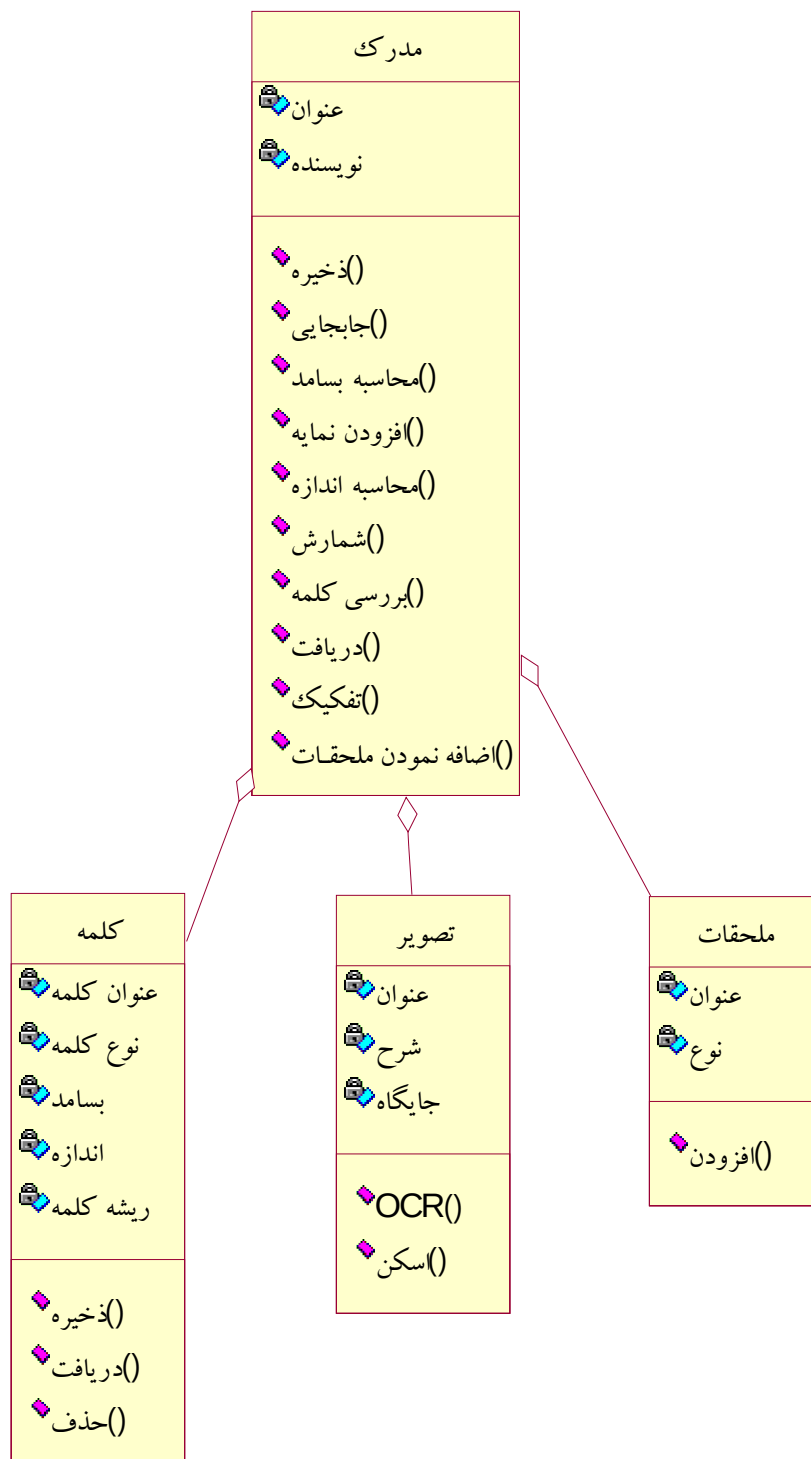
¹ Logical View

² Class diagram

³ Sequence diagram

6-3-1. زیرسیستم مبدل اسناد

شکل 6-7. دیاگرام کلاس¹ زیرسیستم مبدل اسناد

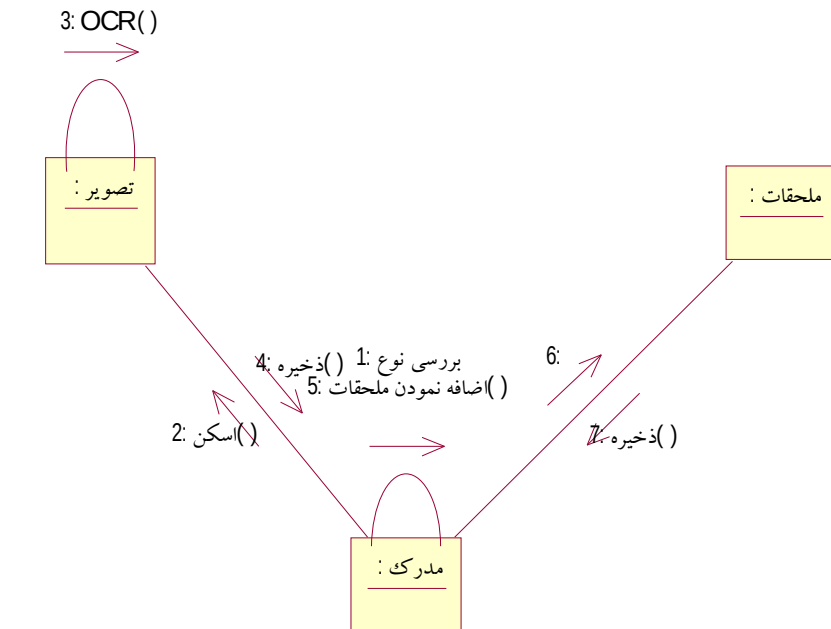


¹ Class diagram

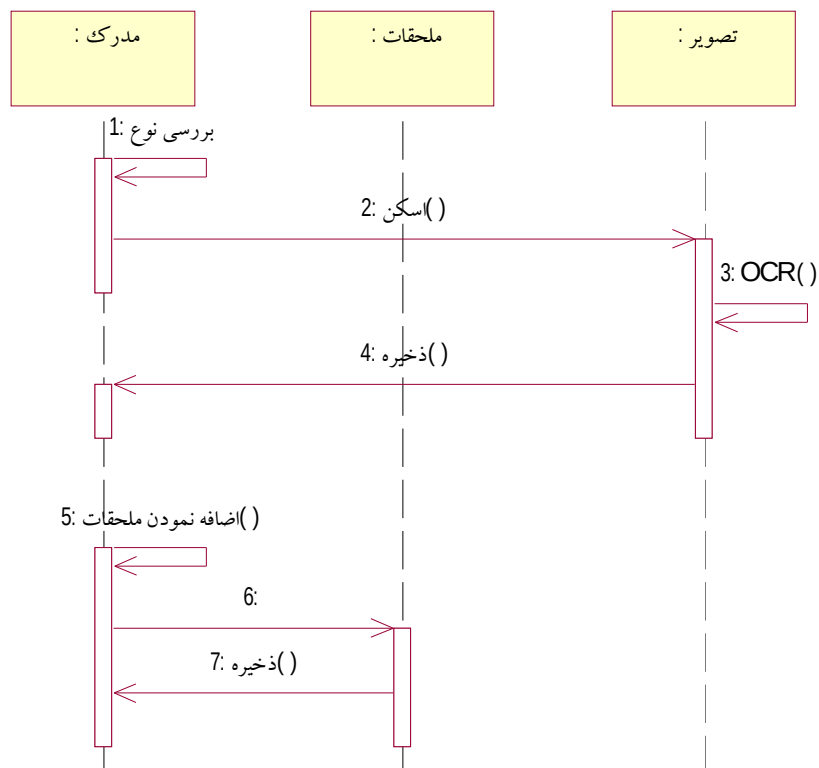
جدول 6-7. شرح اشیاء ¹ در زیرسیستم مبدل اسناد	
مدرک	اسناد، مدارک علمی، مقالات، پایان نامه‌ها، کتب و غیره به عنوان ورودی سیستم در قالب مدرک می‌باشند که کلیه خصوصیات ² و عملکردهای ³ مربوط به مدارک درشی مدرک نگهداری می‌شوند. به عنوان مثال می‌توان به عنوان مدرک، نام نویسنده، محاسبه بسامد، محاسبه اندازه و شمارش و سایر موارد اشاره نمود.
کلمه	متن هر مدرک حداقل از یک یا چند کلمه تشکیل یافته است لذا بین کلمه و مدرک رابطه اجتماع ⁴ برقرار می‌باشد. کلیه خصوصیات و عملکردهای مرتبط با کلمه در اینشی قرار دارد.
تصویر	هر مدرک می‌تواند شامل یک یا چند تصویر باشد. شی تصویر با مدرک دارای رابطه اجتماع می‌باشد.
ملحقات	به طور معمول هر مدرک می‌تواند شامل انواع گوناگونی از پیوست‌ها باشد. شی ملحقات دارای رابطه اجتماع با مدرک می‌باشد. ملحقات می‌توانند تصویر، فیلم، صوت، نرم‌افزار و یا سایر موارد مرتبط با مدرک باشند.

¹ objects
² attributes
³ operations
⁴ aggregation

شکل 6-8. دیاگرام همکاری¹ زیرسیستم مبدل اسناد



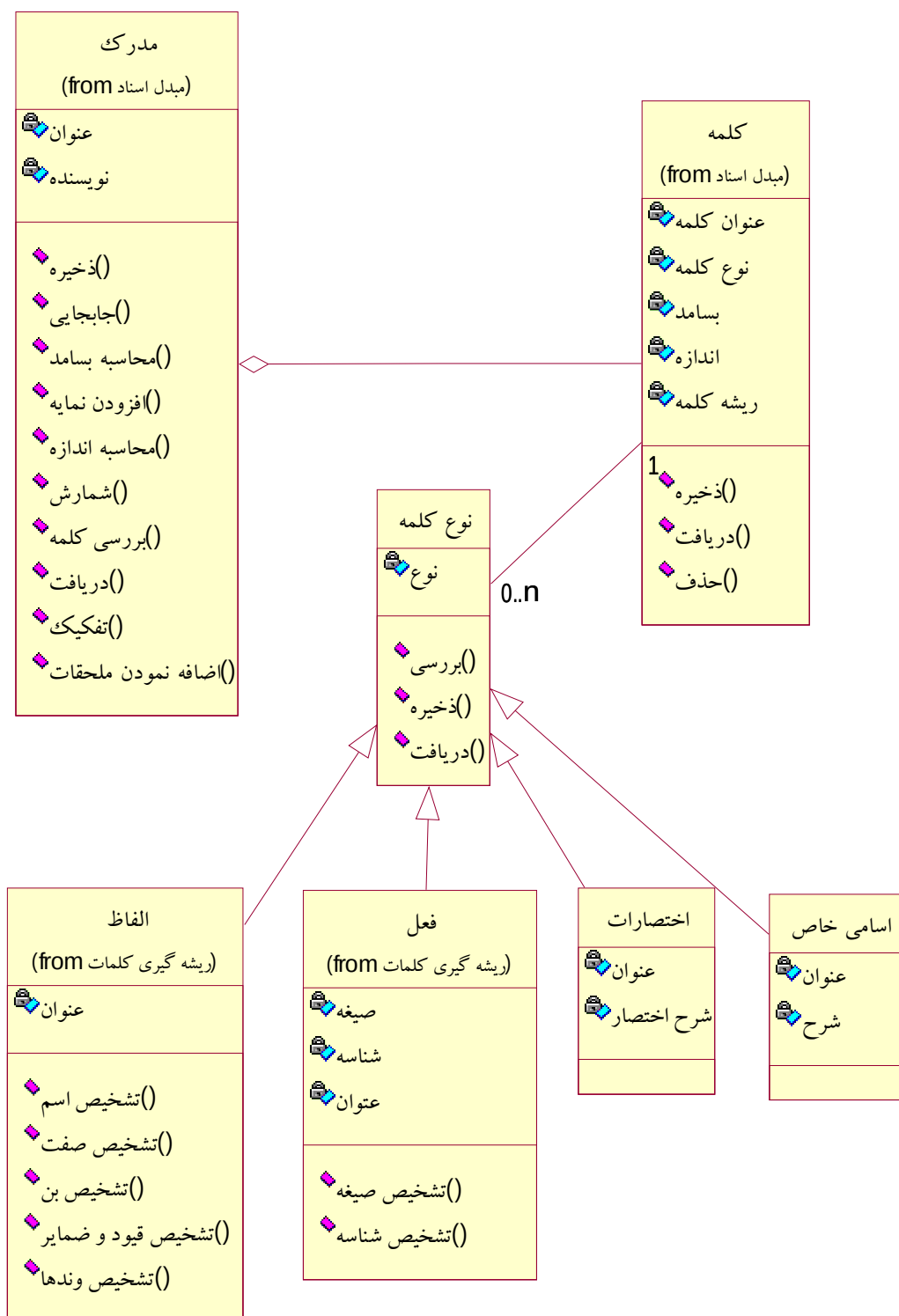
شکل 6-9. دیاگرام توالی² زیرسیستم مبدل اسناد



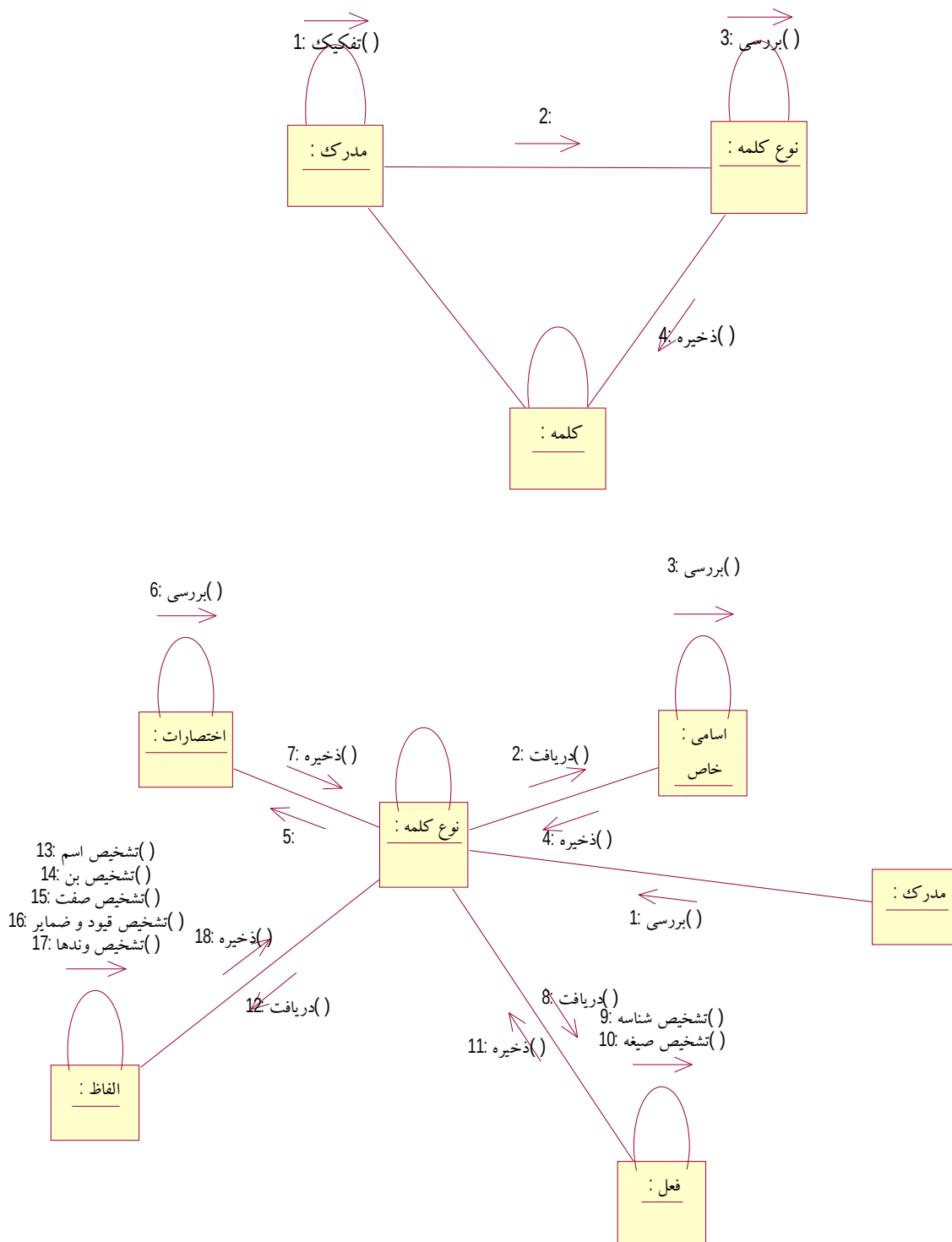
¹ Collaboration diagram
² Sequence Diagram

2-3-6. زیرسیستم تحلیل واژگانی

شکل 6-10. نمودار کلاس زیرسیستم تحلیل واژگانی



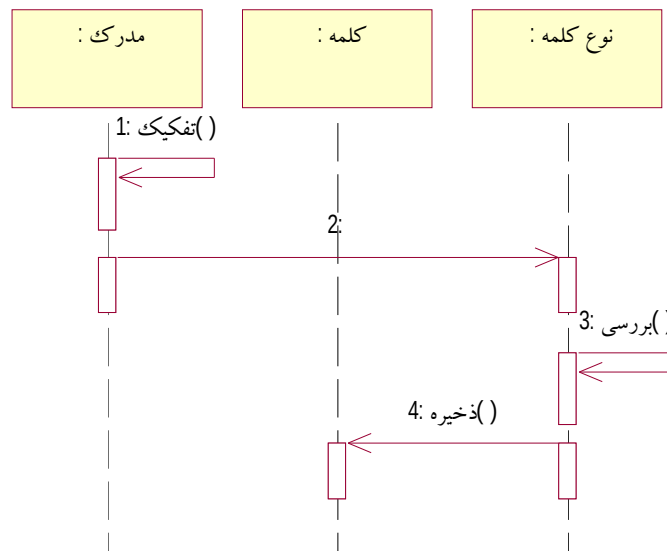
شکل 6-11. نمودار همکاری زیرسیستم تحلیل واژگانی

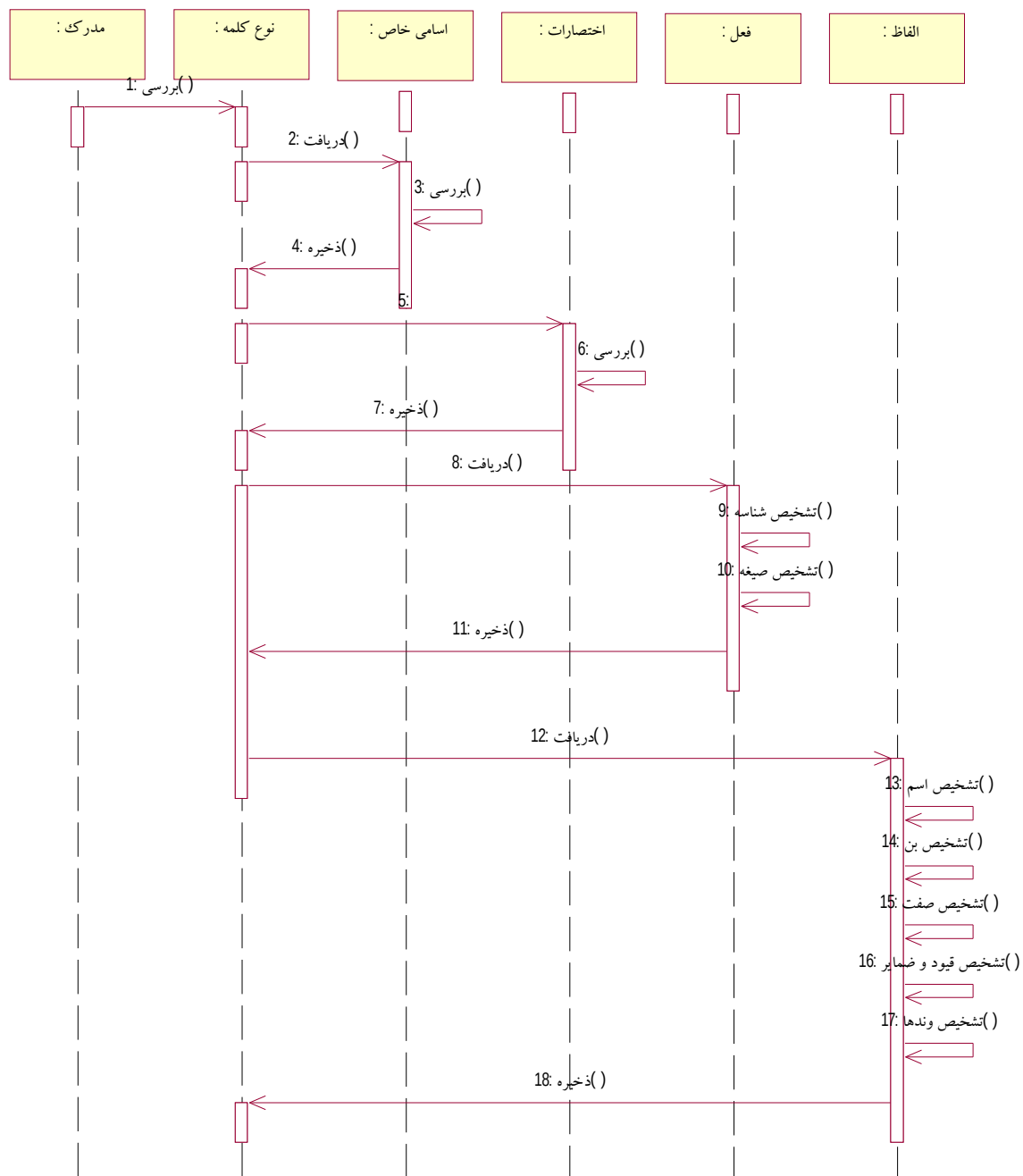


جدول 6-8. شرح اشیاء در زیرسیستم تحلیل واژگانی	
مدرك	اسناد، مدارك علمي، مقالات، پايان نامه‌ها، كتب و غيره به عنوان ورودی سیستم در قالب مدرك می‌باشند كه کلیه خصوصیات و عملکردهای مربوط به مدارك درشی مدرك نگهداری می‌شوند. به عنوان مثال می‌توان به عنوان مدرك، نام نویسنده، محاسبه بسامد، محاسبه اندازه و شمارش و سایر موارد اشاره نمود.
كلمه	متن هر مدرك حداقل از يك یا چند كله تشكيل یافته است لذا بین كلمه و مدرك رابطه اجتماع برقرار می‌باشد. کلیه خصوصیات و عملکردهای مرتبط با كلمه در اینشی قرار دارد.
نوع كلمه	يك فرزند ¹ ازشی نوع كلمه می‌باشد كه منعكس كننده ویژگی‌های افعال می‌باشد.
الفاظ	يك فرزند از كلاس نوع كلمه می‌باشد كه منعكس كننده‌ی ویژگی‌های اسم می‌باشد.
اختصارات	يك فرزند از كلاس نوع كلمه می‌باشد كه اشاره به كلماتی با حروف مختصر دارد كه به صورت مخفف مورد استفاده قرار می‌گیرند.
اسامی خاص	يك فرزند از كلاس نوع كلمه می‌باشد كه اشاره به اسامی سازمان‌ها، نهادها، و غيره را دارد.

¹ child

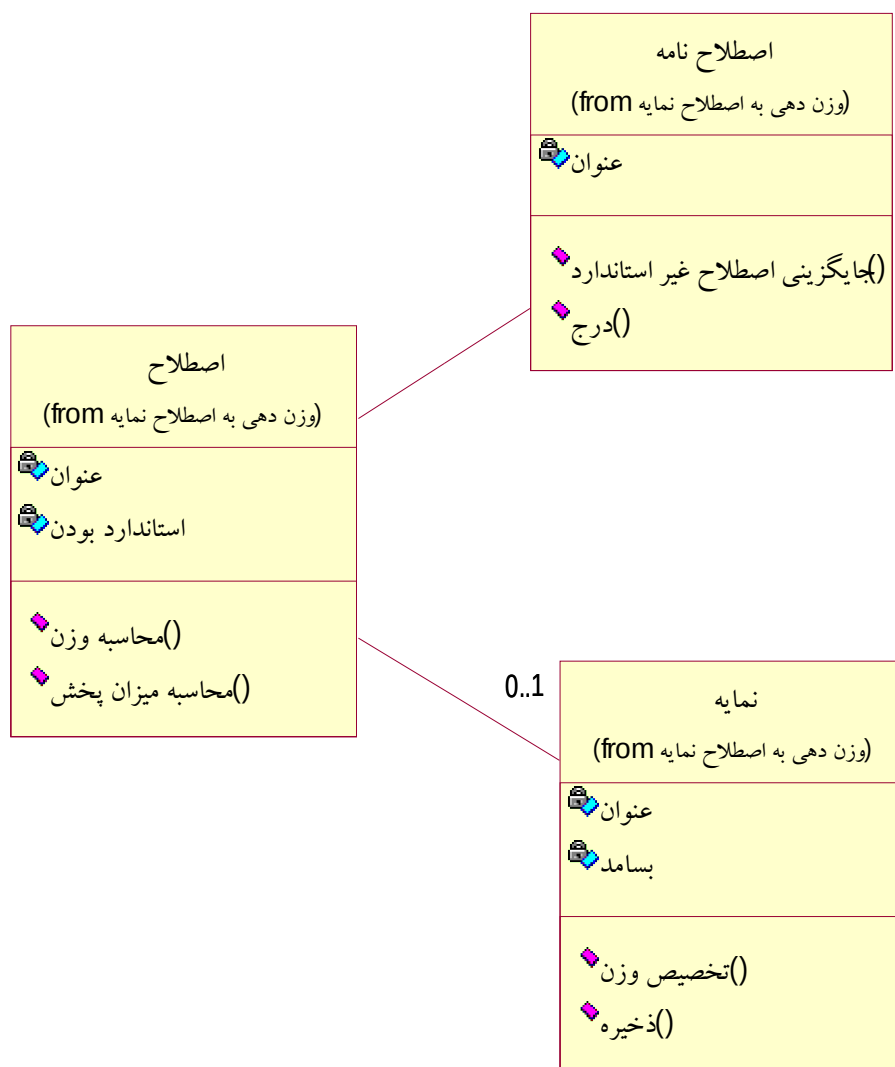
شکل 6-12. نمودار توالی زیرسیستم تحلیل واژگانی





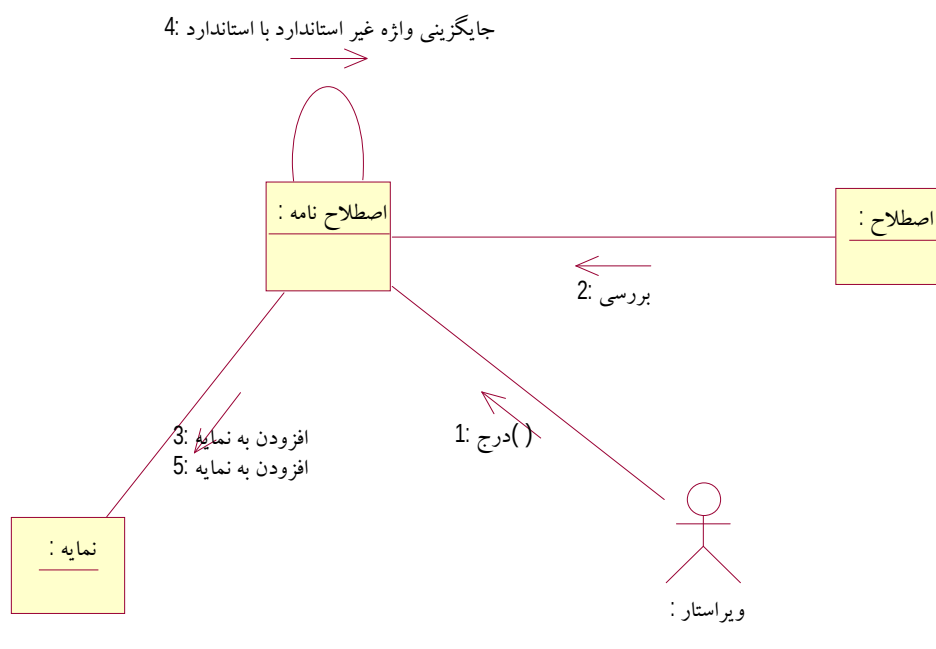
3-3-6. زیرسیستم گزینش اصطلاحات و نرمال سازی

شکل 6-13. دیاگرام کلاس زیرسیستم گزینش اصطلاحات و نرمال سازی

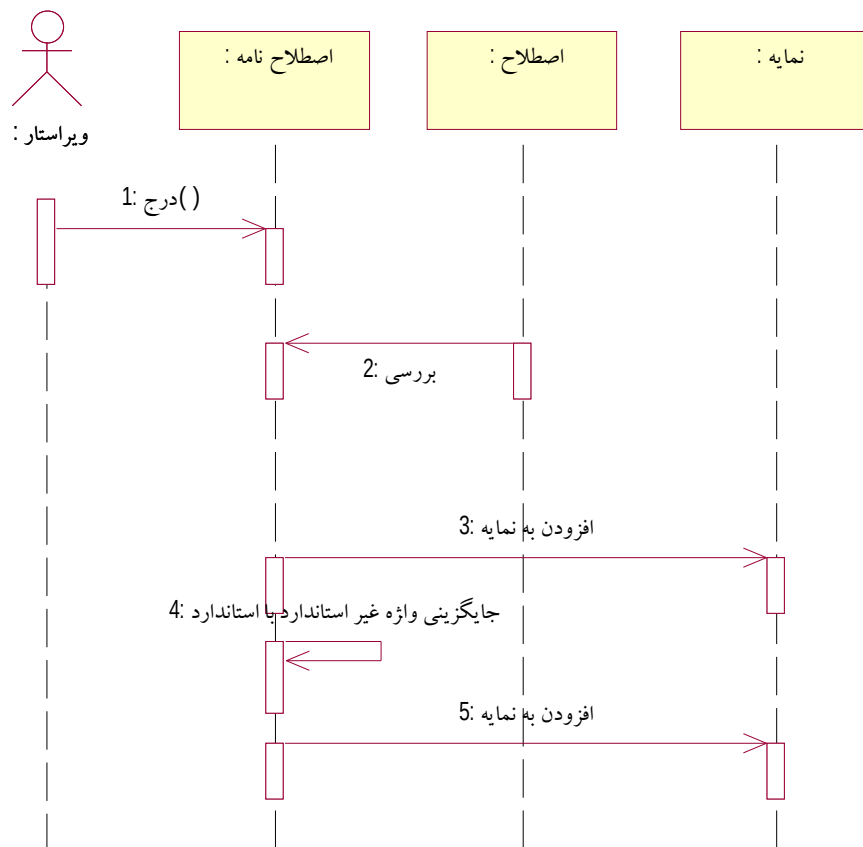


جدول 6-9. شرح اشیاء در زیرسیستم گزینش اصطلاحات و نرمال سازی	
نمایه	هر متن از اجتماع چندین کلمه تشکیل شده است که هر کلمه می تواند به عنوان نمایه ای از مدرک مورد بررسی قرار گیرد.
اصطلاح	یکشی نمایه می تواند با یک یا چند اصطلاح ارتباط داشته باشد اما هر شی اصطلاح الزما دارای ارتباط بایشی نمایه نمی باشد.
اصطلاح نامه	حاوی خصوصیت و رویدادهایی پیرامون اصطلاح نامه ها می باشد.

شکل 6-14. دیاگرام همکاری زیرسیستم گزینش اصطلاحات و نرمال سازی

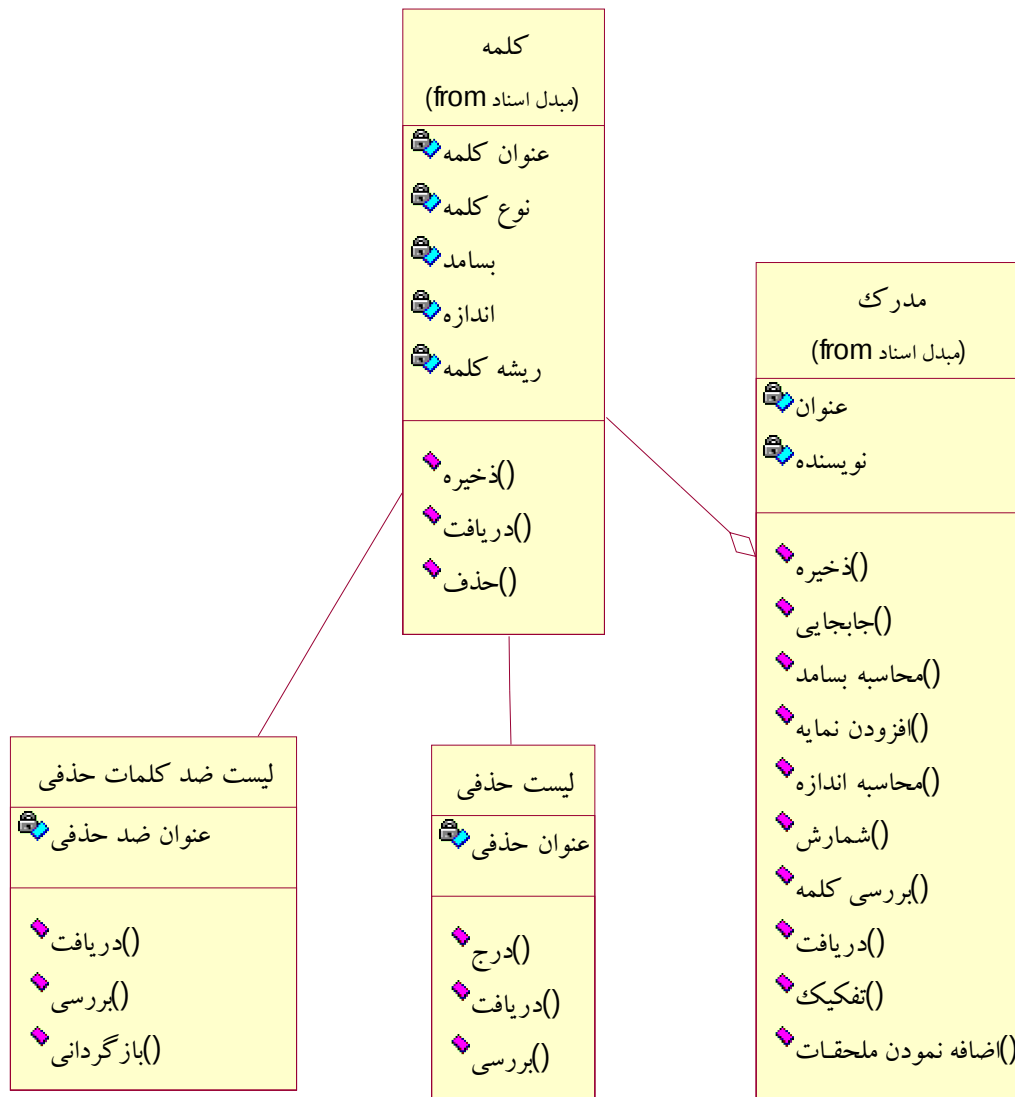


شکل 6-15. دیاگرام توالی



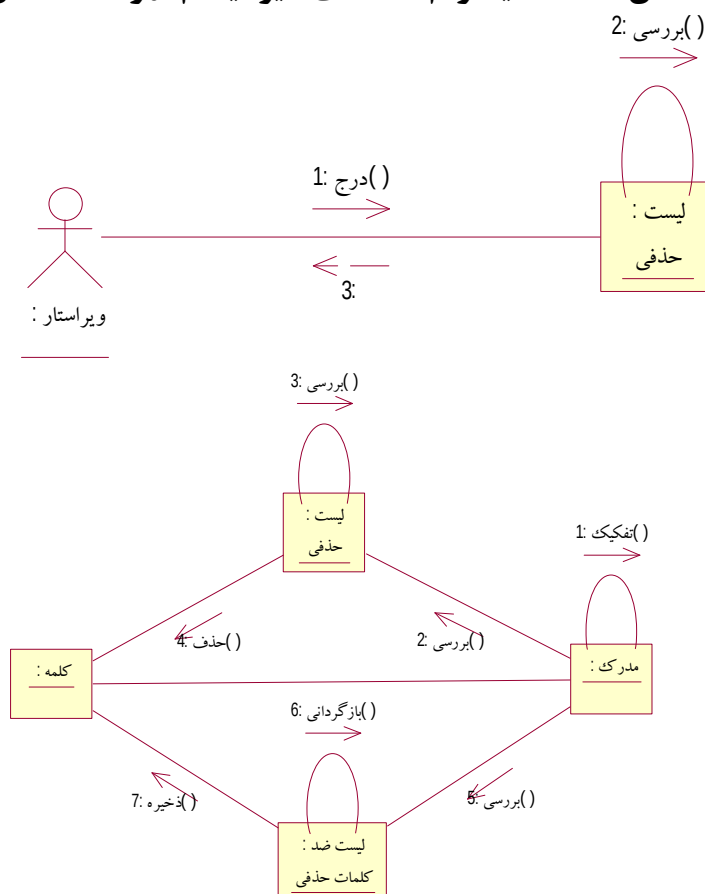
4-3-6. زیرسیستم فهرست حذفی

شکل 6-16. دیاگرام کلاس زیرسیستم فهرست حذفی

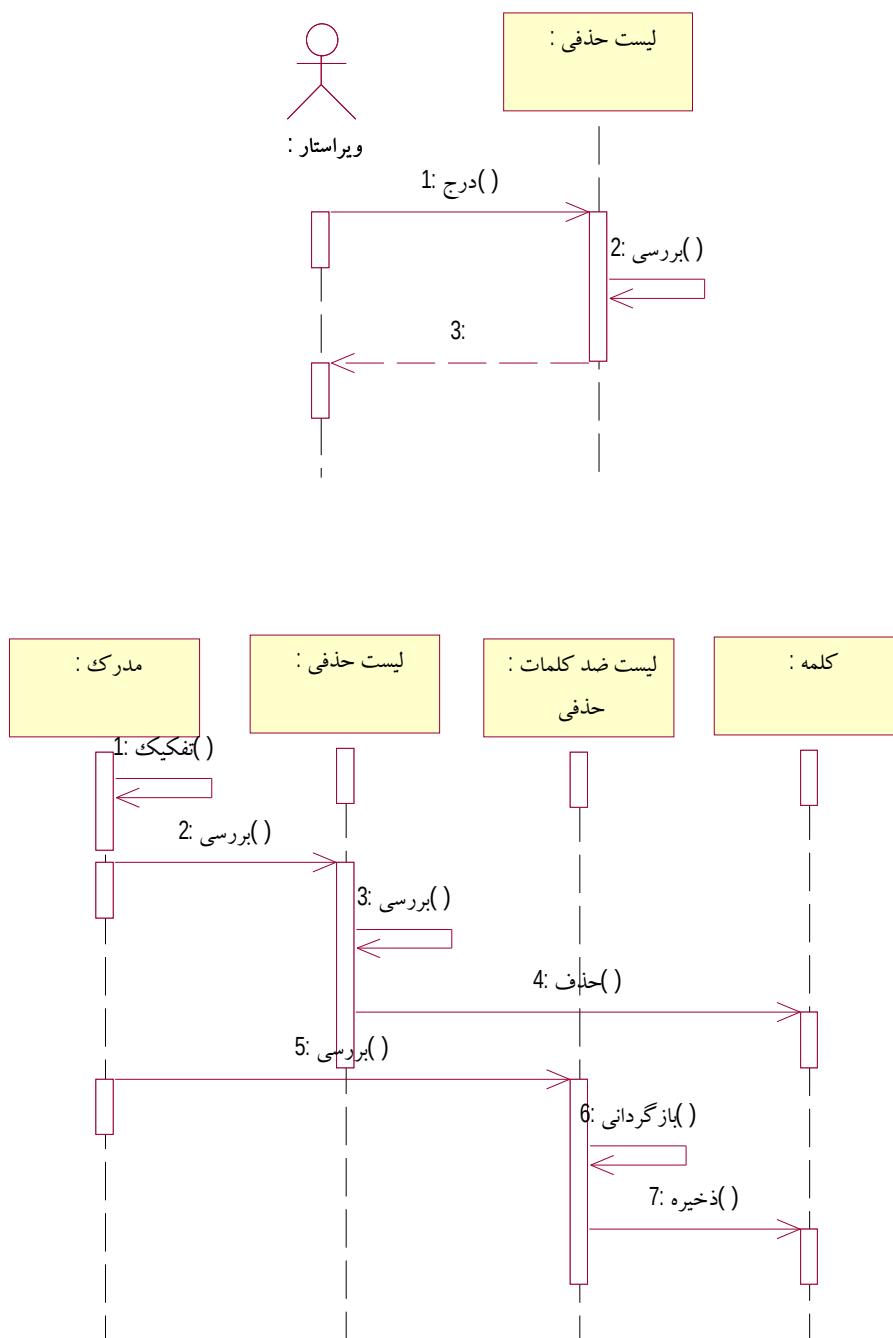


جدول 6-10. شرح اشیا زیرسیستم فهرست حذفی	
کلمه	متن هر مدرک حداقل از یک یا چند کلمه تشکیل یافته است لذا بین کلمه و مدرک رابطه اجتماع برقرار می‌باشد. کلیه خصوصیات و عملکردهای مرتبط با کلمه در اینشی قرار دارد.
نوع کلمه	شی نوع کلمه مشخص کننده‌ی انواع مختلفی می‌باشد که یک کلمه می‌تواند اتخاذ نماید و دارای رابطه‌ی یک به چند با کلمه می‌باشد. به عنوان مثال یک کلمه می‌تواند انواع مختلفی را در مدرک داشته باشد.
لیست حذفی	هر شی لیست حذفی دارای ارتباط با درجه یک به چند با کلمه می‌باشد.
لیست ضد کلمات حذفی	هر شی لیست ضد کلمات حذفی دارای ارتباط با درجه یک به چند با کلمه می‌باشد.

شکل 6-17. دیاگرام همکاری زیرسیستم فهرست حذفی

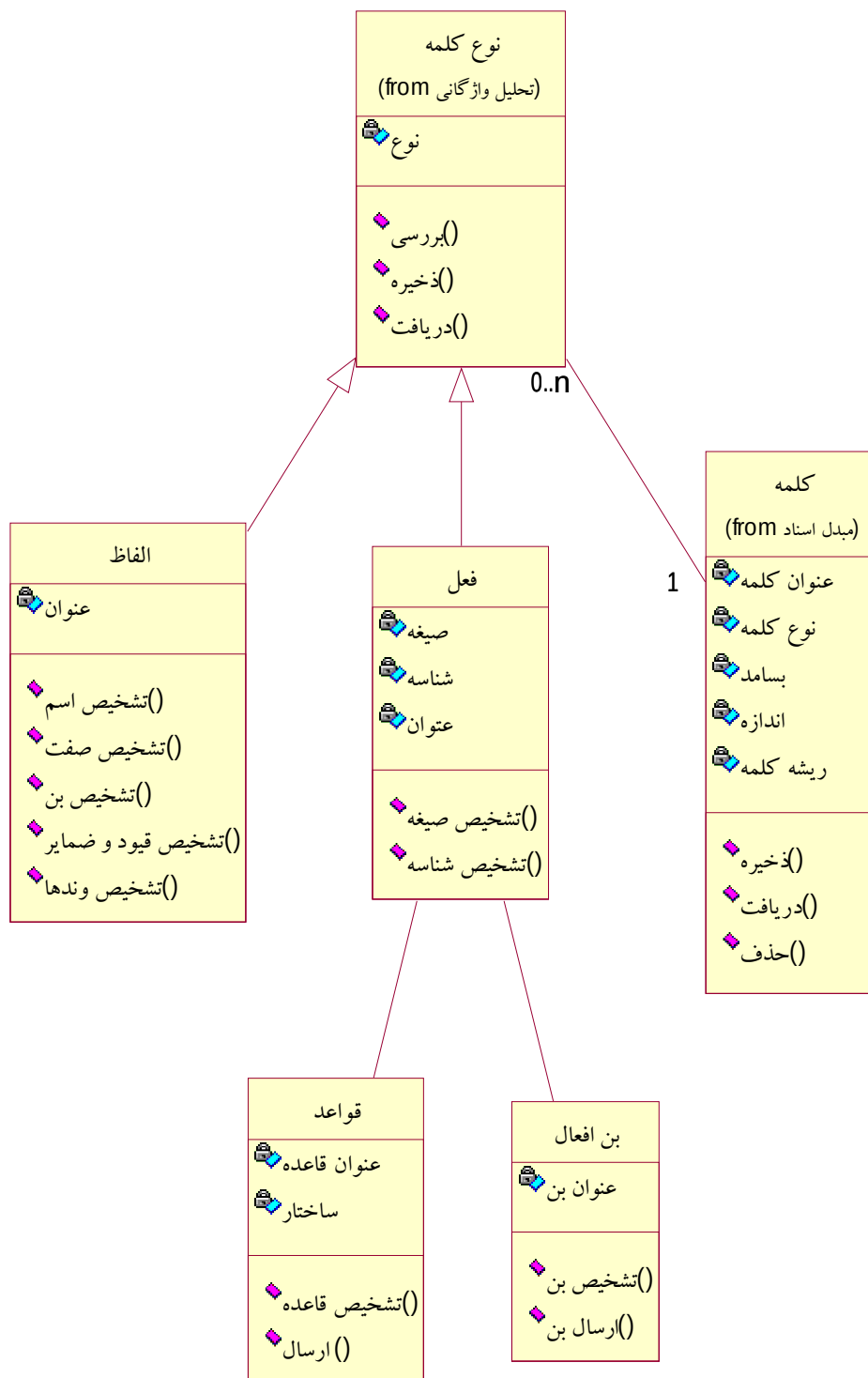


شکل 6-18. دیاگرام توالی زیرسیستم فهرست حذفی

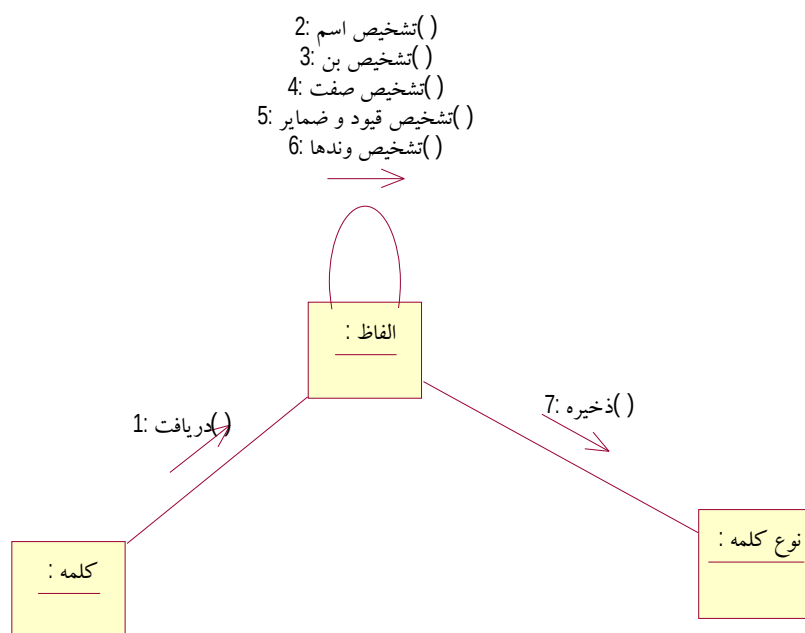


5-3-6. زیرسیستم ریشه‌گیری کلمات

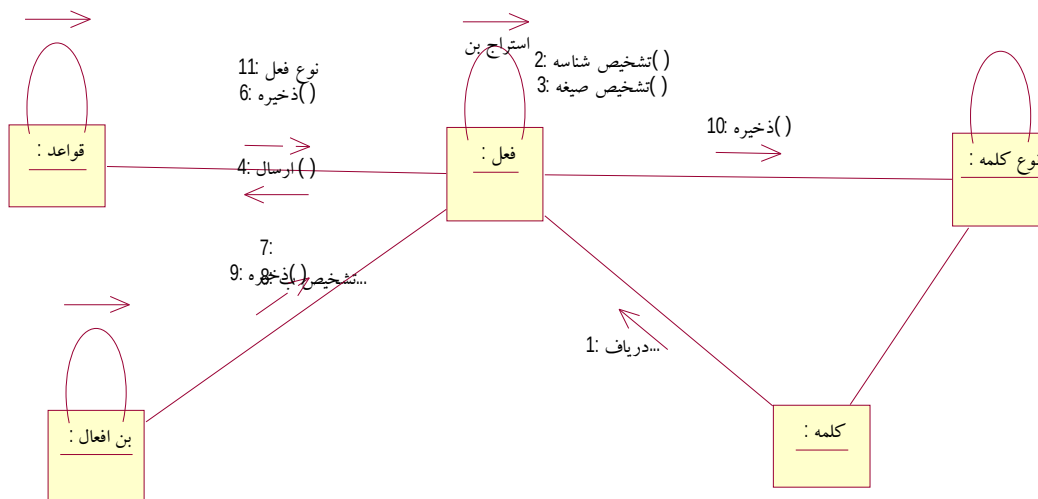
شکل 6-19. نمودار کلاس زیرسیستم ریشه‌گیری کلمات



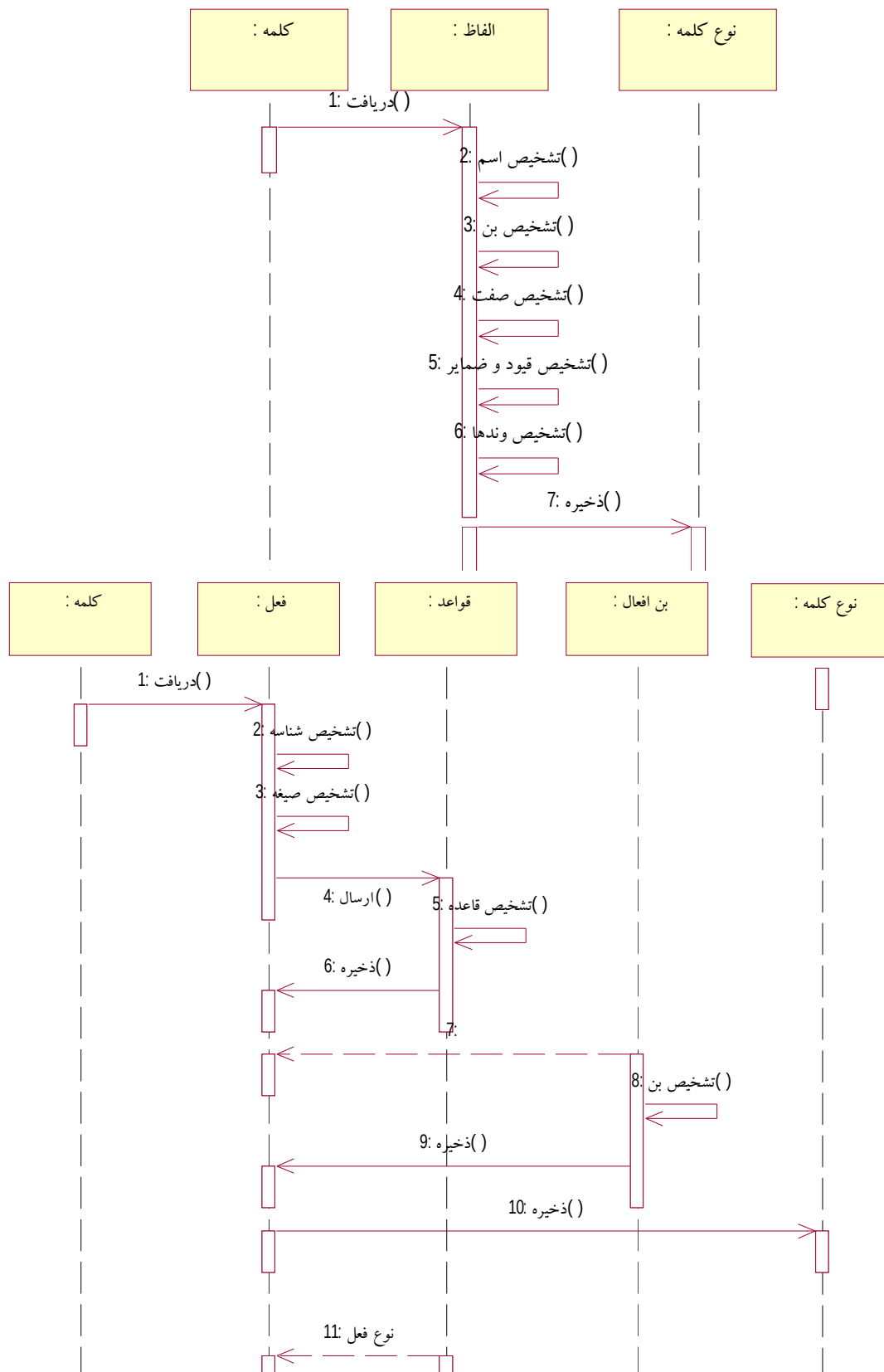
شکل 6-20. نمودار همکاری زیر سیستم ریشه گیری کلمات



(تشخیص قاعده: 5)



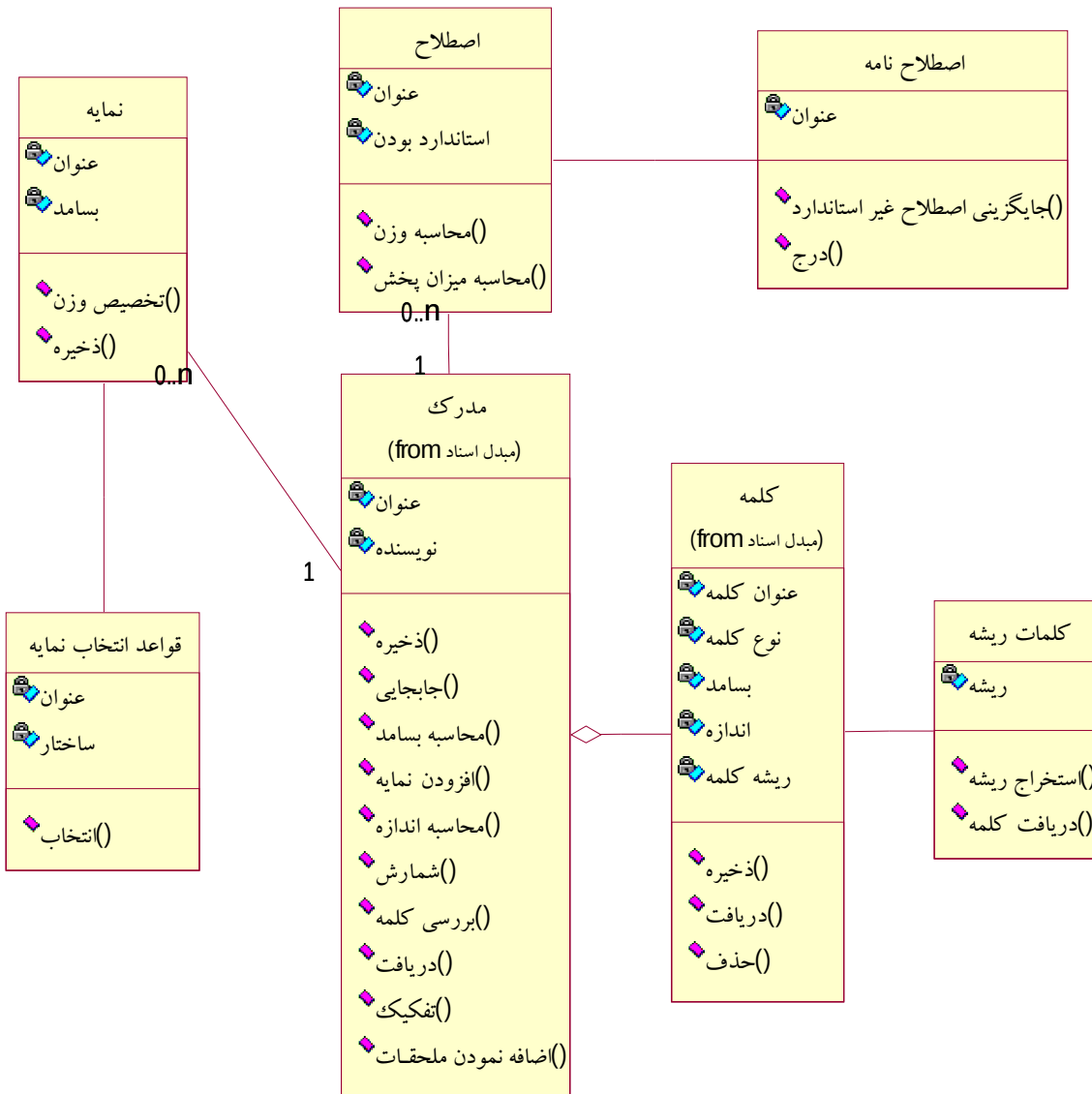
شکل 6-21. نمودار توالی زیرسیستم ریشه‌گیری کلمات



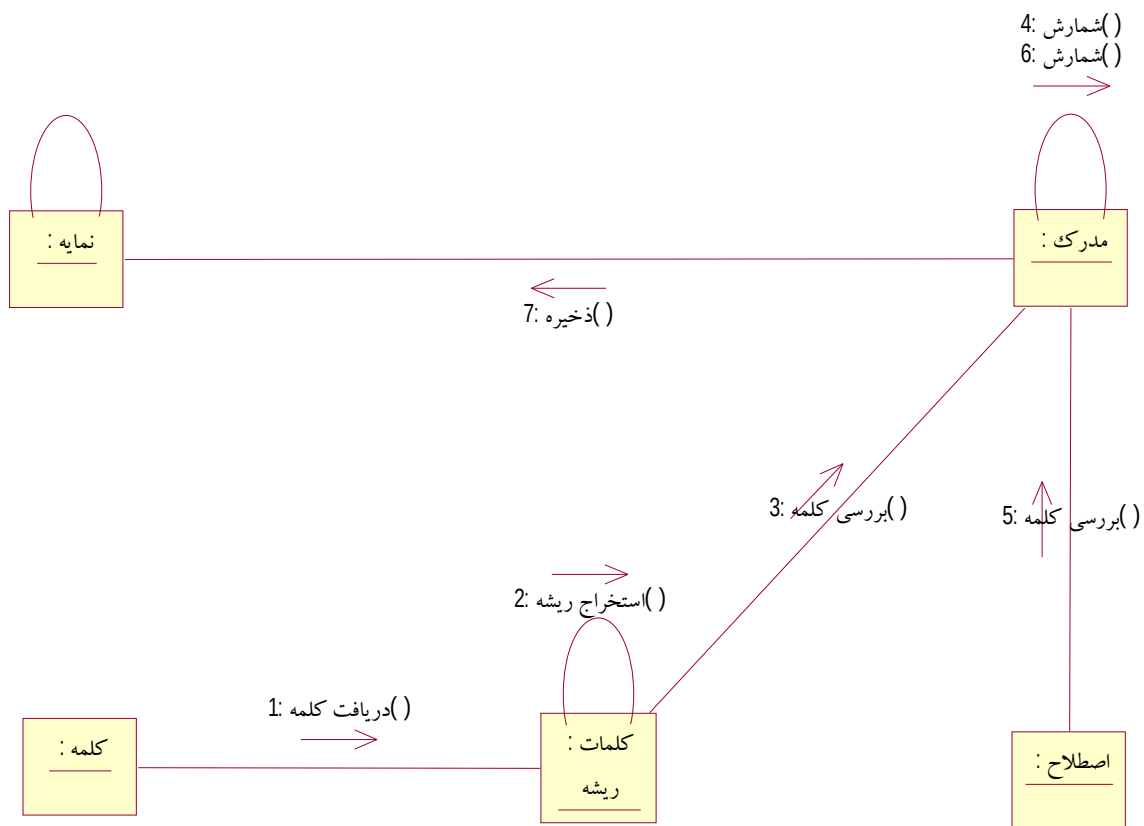
جدول 6-11. شرح اشیا زیر سیستم ریشه گیری کلمات	
کلمه	متن هر مدرک حداقل از یک یا چند کلمه تشکیل یافته است لذا بین کلمه و مدرک رابطه اجتماع برقرار می باشد. کلیه خصوصیات و عملکردهای مرتبط با کلمه در اینشی قرار دارد.
فعل	یک فرزند ازشی نوع کلمه می باشد که منعکس کننده ویژگی های افعال می باشد.
الفاظ	یک فرزند از کلاس نوع کلمه می باشد که منعکس کننده ی ویژگی های اسم می باشد.
بن فعل	هر فعل می تواند دارای بن های مختلفی در یک مدرک باشد. به منظور نگاهداری بن های مختلف افعال ازشی بن فعل استفاده شده است که دارای رابطه یک به چند بایشی فعل می باشد.
قواعد	هر فعل می تواند از طریق قواعد مختلفی ایجاد گردد. به منظور نگاهداری قواعد مختلفی که فعل می تواند داشته باشد شی فوق ایجاد شده است.

6-3-6. زیرسیستم وزن دهی به اصطلاح نمایه

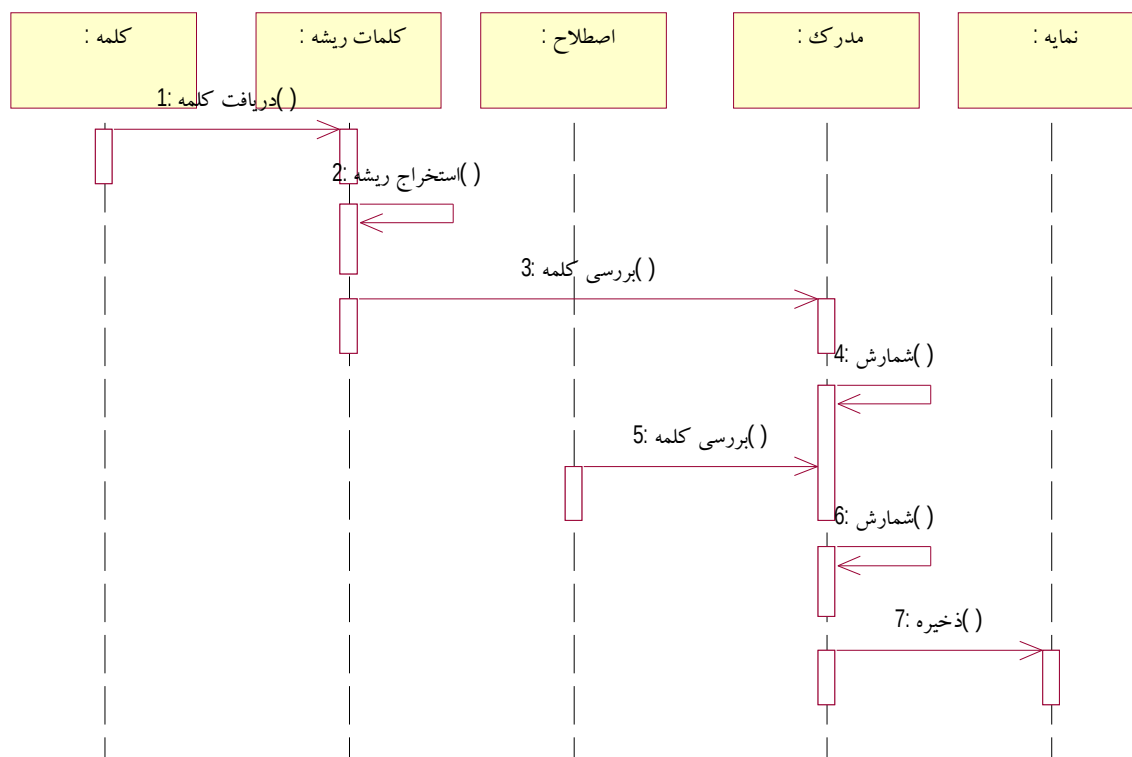
شکل 6-22. نمودار کلاس زیرسیستم وزن دهی به اصطلاح نمایه



شکل 6-23. نمودار همکاری زیرسیستم وزن‌دهی به اصطلاح نمایه



شکل 6-24. نمودار توالی زیرسیستم وزن‌دهی به اصطلاح نمایه



جدول 6-12. شرح اشیا زیرسیستم وزن‌دهی به اصطلاح نمایه	
مدرک	اسناد، مدارک علمی، مقالات پایان‌نامه‌ها، کتب و غیره به عنوان ورودی سیستم در قالب مدرک می‌باشند که کلیه خصوصیات و عملکردهای مربوط به مدارک درشی مدرک نگهداری می‌شوند. به عنوان مثال می‌توان به عنوان مدرک، نام نویسنده، محاسبه بسامد، محاسبه اندازه و شمارش و ساسیر موارد اشاره نمود.
کلمه	متن هر مدرک حداقل از یک یا چند کلمه تشکیل یافته است لذا بین کلمه و مدرک رابطه اجتماع برقرار می‌باشد. کلیه خصوصیات و عملکردهای مرتبط با کلمه در اینشی قرار دارد.
نمایه	هر متن از اجتماع چندین کلمه تشکیل شده است که هر کلمه می‌تواند به عنوان نمایه‌ای از مدرک مورد بررسی قرار گیرد.

اصطلاح	یکشی نمایه می‌تواند با یک یا چند اصطلاح ارتباط داشته باشد اما هرشی اصطلاح الزاما دارای ارتباط بایشی نمایه نمی‌باشد.
اصطلاح‌نامه	حاوی خصوصیت و رویدادهایی پیرامون اصطلاح‌نامه‌ها می‌باشد.
کلمات ریشه	هر کلمه دارای ریشه می‌باشد که ویژگی‌ها و رویدادهای آن درشی کلمات ریشه قرار داد و دارای ارتباط بایشی کلمه می‌باشد.
قواعد انتخاب نمایه	با توجه به راهکارهای مختلف موجود در شناسایی نمایه، هر کدام از روش‌ها در قالبی قواعد انتخاب نمایه قرار می‌گیرد که شامل مشخصاتی نظیر پارامترها و ساختار قواعد می‌باشد.

۷. نمایه‌سازی ماشین:

نیازمندیهای ایجاد سیستم در پژوهشگاه

پژوهشگاه علوم و فناوری اطلاعات^۱ از دیرباز در حوزه نمایه‌سازی به صورت تخصصی فعالیت داشته و مسئولیت نمایه‌سازی بسیاری از مدارک علمی از جمله پایان‌نامه‌ها، گزارشات دولتی و نشریات در کشور بر عهده این سازمان بوده است. همچنین پایگاه وب پژوهشگاه نیز از گذشته در گاه مهمی برای دسترسی به مدارک علمی به خصوص پایان‌نامه‌ها بوده و از اینرو اهتمام به سازماندهی، اشاعه و تحلیل اطلاعات از نیازهای اصلی این پژوهشگاه می‌باشد.

از طرف دیگر افزایش قابل توجه حجم اطلاعات تولید شده در کشور و در پی آن پایان‌نامه‌ها و سایر مدارک مرتبط با مأموریت پژوهشگاه مشکلات اساسی برای پژوهشگاه در راستای مأموریت خود ایجاد کرده است که از آن جمله می‌توان به موارد زیر اشاره کرد:

- نیروی انسانی پژوهشگاه شامل پرسنل متخصص اعم از نمایه‌ساز و ویراستار جوابگوی نیازهای نمایه‌سازی پژوهشگاه نبوده، لذا نمایه‌سازی اطلاعات در پژوهشگاه به‌روز نبوده و باعث شده که کاربران به اطلاعات بروز پایان‌نامه‌ها و سایر مدارک علمی موجود در پژوهشگاه دسترسی نداشته باشند.

- ابلاغ وزیر در سال ۱۳۸۷ مبنی بر الزام فرستادن پایان‌نامه‌ها به پژوهشگاه و تایید پژوهشگاه جهت فارغ‌التحصیلی دانشجویان کارشناسی ارشد و دکتری باعث افزایش دفعتی حجم ورود اطلاعات به پژوهشگاه شده و این روند هنوز در حال رشد می‌باشد.

طی دو سال اخیر پژوهشگاه حرکات خوبی در زمینه بهبود پروسه فراهم‌آوری انجام داده که از جمله سیستم دریافت پایان‌نامه به صورت برخط و دیجیتالی می‌باشد. به هر حال با توجه به موارد مذکور لزوم وجود سیستمی که فرآیند نمایه‌سازی را تسهیل کند انکارناپذیر است.

در گذشته فعالیت‌های غیرمتمرکزی در حوزه نمایه‌سازی ماشینی در پژوهشگاه صورت گرفته که می‌توانند مبنایی برای ایجاد سیستم متمرکز نمایه‌ساز ماشینی باشند. از جمله مهم‌ترین این فعالیت‌ها می‌توان به پروژه‌های واژه‌شکن، فعل‌شکن، اصطلاح‌نامه جامع و OCR اشاره کرد. در ادامه به صورت

^۱ نام این پژوهشگاه در گذشته پژوهشگاه اطلاعات و مدارک علمی ایران بوده و از سال ۱۳۸۸ تغییر نام داده است

مختصر به پروژه‌های مرتبط با طرح مذکور اشاره خواهد شد و نحوه استفاده از آن‌ها در نمایه‌ساز ماشینی نیز مورد بررسی قرار خواهد گرفت.

7-1. واژه‌شکن

پروژه واژه‌شکن فارسی در شهریورماه سال 1385 توسط دکتر سمائی در پژوهشگاه به اتمام رسید. این پروژه که به گفته مجری در حدود 3 سال به طول انجامیده، به وسیله تیمی متشکل از اساتید حوزه زبان‌شناسی و نرم‌افزار در این پژوهشگاه صورت گرفته است.

واژه‌شکن که در واقع یک نرم‌افزار است، از زبان مجری چنین عمل می‌کند.

"واژه‌شکن فارسی برنامه‌ای رایانه‌ای است که کلمات پیچیده زبان فارسی را به اجزاء 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 قرار دارد. نرم‌افزار با عرضه کلمه مفروض به آن، ابتدا صورت کامل کلمه را در حافظه جستجو و مقوله دستوری و معنایی آن را تعیین می‌کند. چنانچه صورت کامل کلمه در حافظه وجود نداشت در پی تجزیه آن بر می‌آید. و اگر اجزاء کلمه طبق قواعد داده شده در کنار هم آمده باشند مقوله دستوری و معنایی آن را به کاربر عرضه می‌کند. چنانچه اجزاء مطابق قواعد نباشد صرفاً به تجزیه کلمه پیچیده اکتفا می‌کند. واژه‌شکن فارسی هم کلمات مشتق و هم کلمات مرکب و مرکب بن ساختی را تجزیه می‌کند و هم قواعد ترکیب حوزه اشتقاقی و حوزه کلمات مرکب و کلمات بن ساختی را می‌شناسد."

با توجه به قابلیت واژه‌شکن که در واقع بخشی از پروژه ناتمام مترجم فارسی بوده است، می‌توان گفت که این نرم‌افزار می‌تواند یکی از اجزای اصلی نمایه‌ساز ماشینی باشد. به طور مثال همانطور که اشاره شد شناسایی ریشه کلمات جهت شمارش کلمات هم ریشه یا هم‌خانواده آن درون متن یکی از مهم‌ترین فاکتورها برای ام‌واژه به عنوان نمایه است، که واژه‌شکن این کار را انجام می‌دهد.

7-2. فعل‌شکن فارسی

پروژه فعل‌شکن فارسی نیز همانند پروژه واژه‌شکن در خرداد 1388 توسط دکتر مهدی سمائی در این پژوهشگاه انجام گرفته، ولی هنوز به نرم‌افزار تبدیل نشده و صرفاً یک طرح پژوهشی است. هدف

از این طرح ساخت نرم‌افزاری برای استخراج قواعد برای تشخیص و تجزیه‌ی فعل‌های ساده است. به گفته‌ی نویسنده در این طرح طرح فعل‌های مرکب و پیشوندی بررسی نشده‌اند. پروژه فعل‌شکن در واقع یک پروژه نیمه تمام است که اگر قرار به استفاده از آن در نمایه‌ساز ماشینی باشد، نیازمند تکمیل و تبدیل به یک مدول نرم‌افزاری است.

3-7. اصطلاح‌نامه جامع

شاید بتوان گفت، سابقه این پژوهشگاه در حوزه اصطلاح‌نامه از تمامی سازمان‌های مشابه بیشتر و تجربه و دانش سازمانی آن در حال حاضر فراتر از هر سازمان دیگری باشد. اصطلاح‌نامه جامع طرحی است که بر اساس گفته مجری آن خانم دکتر ملوک‌السادات بهشتی برای اهداف زیر در پژوهشگاه انجام گرفته است.

- فراهم نمودن مجموعه اطلاعات واژگانی غنی جهت استفاده نمایه‌سازان مراکز اطلاع‌رسانی کشور.
- مشخص نمودن استراتژی‌های جستجو با استفاده از اصطلاح‌نامه‌های کامپیوتری.
- تلاش در تثبیت اصطلاحات فارسی و کاربرد صحیح به عنوان زبان علم.
- مکانیزه کردن نمایه‌سازی مدارک علمی کشور.
- افزایش بهره‌وری اطلاع‌رسانی در سطح کشور.
- هماهنگی بین مراکز اطلاع‌رسانی کشور در انتخاب معادل‌های فارسی.
- کمک به برنامه‌ریزی علمی در بخش‌های مختلف کشور با استفاده از بازیابی اطلاعات مربوط به پژوهش‌های انجام یافته.

این اصطلاح‌نامه در سال 1386 در 7 زمینه «زیست‌شناسی»، «زمین‌شناسی»، «شیمی»، «فنی - مهندسی»، «فیزیک»، «ریاضی» و «کشاورزی» تکمیل و از طریق پایگاه وب پژوهشگاه برای نمایه‌سازان درون پژوهشگاه و متخصصان خارج از پژوهشگاه قابل دسترسی شد. در حال حاضر تعداد اصطلاح‌نامه‌های موجود در این سیستم به 10 مورد افزایش یافته است. شکل زیر صفحه اصلی این اصطلاح‌نامه را که از طریق پایگاه وب پژوهشگاه در اسفندماه سال 1388 گرفته شده نشان می‌دهد.

☐ AGROVOC ☐ Engineering ☐ Mathematics ☒ فنی و مهندسی ☒ علوم زمین ☒ ارتقای بهداشت
☐ Biological Sciences ☐ Geosciences ☐ Physics ☒ فیزیک ☒ علوم زیستی ☒ جامعه شناسی
☐ Chemistry ☐ Health Promotion ☐ Sociology ☒ مدیریت بحران ☒ علوم کشاورزی ☒ ریاضیات
☐ Crisis Management ☒ شبیهی

SEARCH: اصلاح MATCH: ☒ Any part of term ☐ From start ☐ Whole term PAGE SIZE: 10 [Start Search](#)

	Thesaurus	Term	Thesauri List
1.	● علوم زمین	اصلاح	→
2.	● فنی و مهندسی	اصلاح	→
3.	● جامعه شناسی	اصلاح	→
4.	● جامعه شناسی	اصلاح اجتماعی	→
5.	علوم کشاورزی	اصلاح اراضی کشاورزی	→
6.	● جامعه شناسی	اصلاح آموزش	→
7.	● علوم کشاورزی	اصلاح چراگاه	→
8.	● علوم کشاورزی	اصلاح خاک	→
9.	فنی و مهندسی	اصلاح دند	→
10.	● جامعه شناسی	اصلاح رفاهی	→

1 2 3 4
 Total Results: 34

©2009 Thesaurus Builder™ - All rights reserved.
 Powered by [ThesaurusBuilder.com](#)

7-4. نرم افزار OCR¹

این طرح که در دو مرحله توسط مهندس اسماعیل فرامرزی در پژوهشگاه انجام گرفته، به گفته مجری دارای قابلیت های زیر می باشد:

شناسایی بخش چکیده پایان نامه ها به صورت خود کار و تبدیل آنها به متن با استفاده از SDK های نرم افزارهای OCR موجود.

تشخیص شماره صفحات مربوط به بخش عناوین، فهرست مطالب و شماره صفحات اصلی پایان نامه ها و ایجاد ارتباط بین آنها و متن به صورت لینک.

این نرم افزار هم می تواند بر روی پایگاه های داده درون پژوهشگاه نصب شده و هم به صورت مجزا قابل استفاده می باشد.

¹ Optical Character Recognition

5-7. ارزیابی وضعیت فعلی پژوهشگاه برای ایجاد نمایه‌ساز ماشینی

با توجه به مطالبی که درخصوص تجربه‌ها و پروژه‌های صورت گرفته در پژوهشگاه بیان شد، شاید بتوان گفت پژوهشگاه علوم و فناوری اطلاعات ایران، بهترین گزینه برای طراحی و توسعه سیستم نمایه‌ساز ماشینی در کشور باشد. ولی این بدین معنی نیست که ساخت نمایه‌ساز ماشینی نیز هزینه و زمان کمی نیاز داشته باشد.

بر اساس مدل مفهومی که در بخش قبلی طراحی و مورد بررسی قرار گرفت، یک سیستم نمایه‌ساز ماشینی مدول‌های متعددی را شامل می‌شود که برخی از آن‌ها مثل اصطلاح‌نامه‌ها و یا بررسی نحوی کلمات خود زیرسیستم بسیار بزرگی را شامل می‌شود.

نکته مهمی که در خصوص سیستم نمایه‌ساز ماشینی در حال بلوغ می‌توان گفت بزرگ بودن این پروژه است. بدین معنی که اگر پژوهشگاه قصد کند که این سیستم را از ابتدا در حالت بهینه خود طراحی کند می‌بایست هزینه و زمان بسیار زیادی را صرف آن نماید و نهایتاً احتمال عدم موفقیت در آن وجود دارد. به علاوه بخش مهمی از کارایی این سیستم به داده‌هایی ربط خواهد داشت که در آینده وارد آن شده و باعث افزایش خروجی‌های آن می‌گردد. بخش‌های مهمی از سیستم نمایه‌ساز ماشینی شامل الگوریتم‌های ریاضی و آماری است که به مرور زمان می‌بایست تنظیم¹ گردند و به حالت بهینه خود برسند.

با توجه به این نکات جدول بعدی وضعیت پژوهشگاه را در مورد میزان بلوغ در خصوص زیرسیستم‌هایی که در بخش طراحی بررسی شدند نشان می‌دهد. شایان ذکر است که این ارزیابی از طریق مصاحبه و گفتگو با متخصصان پژوهشگاه و مجریان پروژه‌های مذکور به دست آمده و توسط مجری دسته‌بندی گردیده است.

جدول 13-4. ارزیابی وضعیت پژوهشگاه در مورد زیرسیستم‌های نمایه‌ساز ماشینی	
عنوان زیر سیستم	تشریح وضعیت پژوهشگاه
مبدل اسناد	مهم‌ترین بخش در این زیر سیستم تبدیل اسناد غیر دیجیتال به دیجیتال

¹ Tune

	می‌باشد، که قسمت اعظم این فعالیت مربوط به مدول OCR است. در خصوص این زیرسیستم می‌توان گفت پژوهشگاه در حال حاضر تجربه کافی برای ایجاد این زیرسیستم را دارا می‌باشد، لیکن نرم‌افزار موجود برای نیازهای سیستم نمایه‌ساز ماشینی کافی نیست.
تحلیل واژگانی	در مورد این زیر سیستم دو پروژه فعل‌شکن و واژه شکن کمک شایانی خواهند بود، لیکن این دو مدول کفایت تحلیل نحوی متن را نکرده و می‌بایست مدول قوی‌تری از ترکیب این دو، که یکی منجر به نرم‌افزار شده و دیگری هنوز کامپیوتری نشده است، ایجاد گردد. مدول جدید می‌بایست قوانین نحوی مربوط به جملات را در حد نیاز در خود داشته باشد، که قطعاً نیازمند پروژه مجزایی است. از طرف دیگر در زیر سیستم تحلیل واژگانی جدول اسامی خاص وجود دارد که در حال حاضر در پژوهشگاه به صورت غیر نرم‌افزاری موجود و در حال بروزرسانی است، لیکن تلاش‌هایی در جهت نرم‌افزاری نمودن آن شده است. در مورد تشخیص اختصارات علاوه بر این که بسیاری از اختصارات در جدول اسامی خاص موجود می‌باشند لیکن، می‌بایست یک پایگاه قواعد برای تشخیص اختصارات جدید تشکیل شود.
فهرست حذفی	تاکنون فعالیت مشخصی در این خصوص در پژوهشگاه صورت نگرفته، به غیر از اینکه دانش مربوط به این فهرست و نحوه تشکیل آن در ذهن متخصصان نمایه‌سازی و ویراستاران در پژوهشگاه نهادینه شده است و توسعه این زیرسیستم در واقع نیازمند تولید نرم‌افزار بر اساس این دانش خواهد بود.
اصطلاح‌نامه	حاوی خصوصیت و رویدادهایی پیرامون اصطلاح‌نامه‌ها می‌باشد.
گزینش عبارت‌ها و اصطلاحات و نرمال‌سازی	این زیر سیستم از نمایه‌ساز ماشینی که اصطلاح‌نامه هسته اصلی آن است، مهم‌ترین بخش نمایه‌ساز ماشینی را تشکیل خواهد داد، چراکه غالب اصطلاحات منتخب به عنوان نمایه از طریق این زیرسیستم شناسایی خواهند شد. در این مورد تجربیات و نرم‌افزارهای مربوط به اصطلاح‌نامه‌های برخط پژوهشگاه مبنای خوبی را تشکیل می‌دهند لیکن یک نمایه‌ساز ماشینی برای

رسیدن به کارکرد واقعی به اصطلاح‌نامه‌های بسیار بیشتر و قوی‌تر و همچنین یکپارچگی بیشتری احتیاج دارد. از سوی دیگر همانطور که بیان شد شاید وضعیت موجود اصطلاح‌نامه کفایت نسخه‌های اول اصطلاح‌نامه ماشینی را که به احتمال زیاد کاملاً خودکار نیستند، بکند، لیکن در مورد سیستم‌های بالغ‌تر می‌بایست این مجموعه‌ها و ارتباطات بین آن‌ها بسیار قوی‌تر شده و مکانیزم‌های یادگیرنده نیز به آن‌ها اضافه گردد. که این مهم قطعاً هزینه‌های بسیار زیادی را به دوش پژوهشگاه تحمیل خواهد کرد.	
در مورد ریشه‌گیری کلمات شاید با وجود نرم‌افزار واژه‌شکن بتوان گفت بخشی مهمی از این زیرسیستم موجود است، چرا که در این مرحله تحلیل نحوی جمله نیز مورد نظر نیست، لیکن این مدول می‌بایست تکمیل‌تر گردد.	ریشه‌گیری کلمات
در مورد این زیرسیستم تا به حال هیچ پروژه‌ای در پژوهشگاه صورت نگرفته و تجربه خاصی هم وجود ندارد. این زیرسیستم شامل پایگاهی از قواعد آماری و داده‌کاوی برای تعیین نمایه‌های نهایی از بین کلمات نامزد نمایه شدن است که از مراحل قبلی وارد این زیرسیستم شده‌اند. این کلمات شامل اصطلاحات استاندارد اصطلاح‌نامه‌ها، اسامی خاص، اختصارات و برخی کلمات دیگر می‌باشند. به هر حال کار بر روی این زیرسیستم نیازمند تجربه‌های آماری و نرم‌افزاری در حوزه الگوریتم‌های جستجو و بازیابی است.	وزن‌دهی به اصطلاح نمایه

۸. نتیجه‌گیری

هدف اصلی طرح پیش‌رو طراحی مدلی مفهومی برای روشن شدن ابعاد مفهومی و نرم‌افزاری نمایه‌ساز ماشینی به عنوان یک سیستم و همچنین ایجاد مبنایی برای مطالعات و پروژه‌های عملیاتی‌تر در حوزه ایجاد نمایه‌ساز ماشینی در پژوهشگاه علوم و فناوری اطلاعات ایران بود. در این راستا سعی شد برای مرور ادبیات موضوع از منابع متعدد استفاده شده و در مورد طراحی مفهومی سیستم نیز از معماری و زبان مدلسازی استاندارد استفاده گردد، تا بیش از هر چیز برای تحلیل‌گران و توسعه‌دهندگان سیستم قابل فهم بوده و برخی مشکلات متخصصان حوزه‌ی نمایه‌سازی در مورد انتقال دانش به این گروه رفع کند.

اما آنچه می‌توان در انتها به عنوان نتیجه‌ای مهم در مورد نمایه‌ساز ماشینی و ایجاد آن در پژوهشگاه اشاره کرد، قدم‌های متفاوتی است که در مورد آن قابل اجراست. همانطور که قبلاً هم به آن اشاره شد نمایه‌ساز ماشینی به معنای بالغ خود یک پروژه بسیار بزرگ در سطح ملی است، که توسعه آن توسط هر سازمانی من جمله پژوهشگاه نیازمند پشتیبانی‌های مالی و دانشی فراوانی است. لیکن پژوهشگاه می‌تواند این مهم را در زمان طولانی‌تر و قدم‌های کوتاه‌تری محقق نماید. چه اینکه انتخاب متدولوژی RUP نیز به همین منظور بود. در ادامه به صورت مختصر می‌توان به برخی از این قدم‌ها به عنوان‌های قدم‌های عملیاتی‌تر اشاره کرد. البته این قدم‌ها پیشنهادی بوده و قطعاً با نظر متخصصین و کارشناسان می‌بایست بهینه گردند:

- ممکن است در قدم‌های اولیه نمایه‌ساز ماشینی سیستمی باشد که دستیار نمایه‌ساز قرار گیرد و در این حالت سیستم جستجوی اصطلاحات موجود در متن را از طریق اصطلاح‌نامه‌های فعلی انجام داده و آن‌ها را در اختیار نمایه‌ساز قرار می‌دهد و یا اطلاعات مربوط به بسامد کلمات را در اختیار نمایه‌ساز قرار می‌دهد.
- در قدم بعدی نمایه‌ساز به عنوان ناظر سیستم قرار خواهد گرفت و نمایه‌های استخراج شده توسط سیستم در اختیار نمایه‌ساز قرار می‌گیرد تا آن‌ها را تصحیح و نهایی کند.
- آخرین مرحله استقلال سیستم نمایه‌ساز ماشینی و بازبینی دوره‌ای و یا تصادفی نمایه‌ها است. در این مرحله اشکالات موجود در نمایه‌ساز ماشینی رفع شده، پایگاه داده مربوط به

اصطلاح‌نامه‌ها، فهرست حذفی و پایگاه قواعد نحوی و فرمول‌های انتخاب نمایه تکمیل و تنظیم شده‌اند. طبیعی است دستیابی به این سطح از نمایه‌سازی ماشینی بسیار زمانبر می‌باشد.

بهشتی، ملوک السادات. 1378. اصطلاحنامه جامع: گزارش طرح پژوهشی. تهران: پژوهشگاه اطلاعات و مدارک علمی ایران

دیانی، محمدحسین. 1379. مواد و خدمات مرجع: فشرده درس‌ها. تهران: انتشارات کتابخانه رایانه‌ای.

سلطانی، پوری و فرودین راستین. 1381. اصطلاحنامه کتابداری. تهران: فرهنگ معاصر.

سماعی، سید مهدی. 1385. واژه شکن فارسی: گزارش طرح پژوهشی. تهران: پژوهشگاه اطلاعات و مدارک علمی ایران

سماعی، سید مهدی. 1388. فعل شکن فارسی: گزارش طرح پژوهشی. تهران: پژوهشگاه اطلاعات و مدارک علمی ایران

نوروزی، علیرضا. 1380. نمایه سازی کتاب: راهنمای برای ناشران، نمایه سازان، کتابداران، مولفان و مترجمان. تهران: چاپار.

نیاکان، شهرزاد. 1383. بررسی کاربرد نمایه سازی ماشینی در کتابخانه‌ها. علوم اطلاع‌رسانی 16(3-4): 49-55.

Adamson, G.W. and J. Boreham. 1974. The use of an association measure based on character structure to identify semantically related pairs of words and document titles. *Information Storage and Retrieval* 10, 253-260.

Amati, G. and F. Crestani. 1999. Probabilistic learning for selective dissemination of information. *Information Processing and Management* 35, 633-654.

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.

Bookstein, A. 1983a. Information Retrieval: A Sequential Learning Process. *Journal of the American Society for Information Science* 34, 331-342.

Bookstein, A. 1983b. Outline of a General Probabilistic Retrieval Model. *Journal of Documentation* 39(2): 63-72.

Bookstein, A. 1985. Probability and Fuzzy Set Applications to Information Retrieval. *Annual Review of Information Science and Technology* 20, 117-151.

Booth, Pat F. 2001. *Indexing: The manual of good practice*. Munchen: K. G. Saur.

Borko, H., and M. Bernick. 1963. Automatic document classification. *Journal of the Association for Computing Machinery* 10, 151-162.

Buckley, C. 1993. The importance of proper weighting methods. *Human Language Technology: Proceedings of a Workshop Held at Plainsboro, New Jersey, March 21-24, 1993* (pp. 349-352). Sanfrancisco: Morgan Kaufmann.

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

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

- Cooper, W. 1991. Some inconsistencies and misnomers in probabilistic IR. In: *Proceedings of the Fourteenth Annual International ACM SIGIR Conferences on Research and Development in Information Retrieval*, 57-61. ACM, New York.
- Cooper, W. 1995. Some inconsistencies and misidentified modelling assumptions in probabilistic information retrieval. *ACM Transactions on Information Systems* 13, 100-111.
- Cooper, W., A. Chen, and F. Gey. 1994. Full text retrieval based on probabilistic equations with coefficients fitted by logistic regression. *The Second Text REtrieval Conference (TREC-4)*, D. K. Harman (Ed.), Special Publication 500-215 (p. 57-66). National Institute of Standards and Technology, Gaithersburg, MD.
- Crestani, F., M. Lalmas, C. J. van Rijsbergen, and I. C. Campbell. 1998. Is This Document Relevant?... Probably: A survey of probabilistic models in information retrieval. *ACM Computing Surveys* 30(4): 528-550.
- Croft, W. B. 1998. Boolean Queries and Term Dependencies in Probabilistic Retrieval Models. *Journal of the American Society for Information Science* 37(2): 71-77.
- Croft, W. B., and D. J. Harper, 1979. Using probabilistic models of document retrieval without relevance information. *Journal of Documentation* 35, 285-295.
- Croft, W.B., H. R. Turtle, and D. D. Lewis. 1991. The use of phrases and structured queries in information retrieval. In A. Bookstein, Y. Chiaramella, G. Salton, & V.V. Raghaven (Eds.), *Proceedings of the Fourteenth SIGIR Conference* (pp. 32-45). New York: ACM.
- Croghan, Anthony. 2002. The process of indexing. *Catalogue & Index* (143): 8-10.
- Dumais, S. 1997. Using Latent Semantic Indexing (LSI) for information retrieval, information filtering and other things. *Cognitive Technology conference*, April 4.
- Earl, L.L. 1970. Experiments in automatic extracting and indexing. *Information Storage and Retrieval* 6 (6): 313-334.
- 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.
- Fagan, J. L. 1989. The effectiveness of a nonsyntactic approach to automatic phrase indexing for document retrieval. *JASIS* 40 (2): 115-1 32.
- Fox, C. 1992. Lexical analysis and stop lists. In W.B. Fakes & R. Baez-Yates (Eds.), *Information Retrieval: Data Structures & Algorithms* (pp. 102-1 30). Englewood Cliffs, NJ: Prentice Hall.
- Fakes, W. B. 1992. Introduction to information storage and retrieval systems. In W.B. Fakes & R. Baez-Yates (Eds.), *Information Retrieval: Data Structures & Algorithms* (pp. 1-12). Englewood Cliffs, NJ: Prentice Hall.
- Fuhr, N. 1989a. Models for Retrieval with Probabilistic Indexing. *Information Processing and Management* 25(1): 55-72.
- Fuhr, N. 1989b. Optimum Polynomial Retrieval Functions Based on the Probability Ranking Principle. *ACM Transactions on Information Systems* 7(3): 083-204.
- Fuhr, N. 1990. A Probabilistic Framework for Vague Queries and Imprecise Information in Databases. In *Proceedings of the 16th International Conference on Very Large Databases*, 696-707. California: Morgan Kaufman.

- Fuhr, N., C. Buckley. 1991. A probabilistic learning approach for document indexing. *ACM Transactions on Information Systems* 9 (3): 223-248.
- Fuji, H., and W. B. Croft. 1993. A Comparison of Indexing Techniques for Japanese Text Retrieval. *ACM-SIGIR'93*, Pittsburg, PA, USA.
- Gingrande, Arthur. 2004. The value of indexing. *AIIM E-Doc Magazine* 18(2): 9.
- Hater, M.A., and S. F. Weiss. 1974. Word segmentation by letter successor varieties. *Information Storage and Retrieval* 10:371-385.
- Harter, S. P. 1975a. A probabilistic approach to automatic keyword indexing. Parts 1 and 2. *Journal of the American Society for Information Science* 26:197-206.
- Harter, S. P. 1975b. A probabilistic approach to automatic keyword indexing. Parts 1 and 2. *Journal of the American Society for Information Science* 26:280-289.
- Hoekstra, D. 2000. A probabilistic justification for using tf.idf term weighting in information retrieval. *International Journal on Digital Libraries* 3 (2): 131-139.
- Jacobs, P.S. 1993. Using statistical methods to improve knowledge-based news categorization. *IEEE expert*(April): 13-23.
- Knort, G. et al. 1997. Adaptive automatic indexing for complex documents. *LDV-Forum* 7:98-102.
- Koll. M. 1979. WEIRD: An approach to concept-based information retrieval. *SIGIR Forum* 13:32-50.
- Krovetz, R. 1993. Viewing morphology as an inference process. In R. Korfhage, E. Rasmussen, & P. Willett (Eds.), *Proceedings of the Sixteenth SIGIR Conference* (pp. 191- 202). New York: ACM.
- Kuhr, P. 2003. *Abstracting and Indexing*. World Encyclopedia of Library and Information Sciences, 1-5.
- Kwok, K. L. 1995. A network approach to probabilistic information retrieval. *ACM Transactions on Information Systems* 13, 325-353.
- Lancaster, F. W. 1991. *Indexing and abstracting in theory and practice*. Champaign: University of Illinois Graduate School of Library and Information Science.
- Lee, J. H. and J. S. Ahn. 1996. *Using n-Grams for Korean Text Retrieval*. SIGIR 96, Zurich, Switzerland.
- Letsche, Todd A. 1996. *Toward Large-Scale information retrieval using Latent Semantic Indexing*. A thesis presented for the Master of Science Degree, The University of Tennessee, Knoxville.
- Lovins, J. B. 1968. Development of a stemming algorithm. *Mechanical Translation and Computational Linguistics* 11 (1-2): 22-31.
- 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.
- Mani, I., and T.K. MacMillan. 1996. Identifying proper names in newswire text. In B. Boguraev & J. Pustejovsky (Eds.), *Corpus Processing for Lexical Acquisition* (pp. 41-59). Cambridge, MA: The MIT Press.
- Mansour, N., R. A. Haraty, W. Daher, and M. Houri. 2008. An auto-indexing method for Arabic text. *Information Processing and Management* 44:1538–1545.
- Maron, M. D., and J. L. Kuhns. 1960. On relevance, probabilistic indexing and information retrieval. *Journal of the ACM* 7: 216-244.
- 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 21 November 2009).

- National Information Standards Organization (NISO). 1997. Guidelines for indexes and related information retrieval devices: A technical report. Bethesda: NISO Press.
- Noreault, T., M. McGill, and M.B. Koll. 1981. A performance evaluation of similarity measures, document term weighting schemes and representations in a Boolean environment. In R.N. Oddy, S.E. Robertson, C. J. van Rijsbergen, & P. W. Williams (Eds.), *Information Retrieval Research* (pp. 57-76). London: Butterworth & Co.
- Ohsawa, Y., N. E. Benson, and M. Yachida. 1998. KeyGraph: Automatic Indexing by Co-occurrence Graph based on Building Construction Metaphor. Proceedings of the Advances in Digital Libraries Conference, P. 12, <http://portal.acm.org/citation.cfm?id=785950> (accessed 16 June 2010).
- Online Dictionary for Library and Information Science (ODLIS). 2009. Automatic Indexing. <http://lu.com/odlis> (accessed 28 September 2009).
- Prasher, Ram Gopal. 1989. *Index and indexing*. Delhi: Medallion press.
- Rau, L.F. 1992. Extracting company names from text. In *Seventh IEEE AI Applications Conference*, 189-194.
- Ro, J.S. 1988. An evaluation of the applicability of ranking algorithms to improve the effectiveness of full-text retrieval. II. On the effectiveness of ranking algorithms on full text retrieval. *JASIS* 39 (3): 147-160.
- Robertson, S. E. 1977. The probability ranking principle in IR. *Journal of Documentation* 33:294-304.
- Robertson, S. E. 1990. On term selection for query expansion. *Journal of Documentation* 46:359-364.
- Robertson, S. E., and J. D. Bovey. 1982. Statistical problems in the application of probabilistic models to information retrieval. UK: City University.
- Robertson, S. E., and S.Walker, 1994. Some simple effective approximations to the 2 Poisson model for probabilistic weighted retrieval. In *Proceedings of the Seventeenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 232-241). New York: Association for Computing Machinery.
- Robertson, S. E., and S. Walker. 2000. Threshold setting in adaptive filtering. *Journal of Documentation* 56 (3): 312-331.
- Robertson, S., E., Maron, and W. S. Cooper. 1982. Probability of relevance: a unification of two competing models for document retrieval. *Information Technology: Research and Development* 1: 1-21.
- Salton, G. 1971. A new comparison between conventional indexing (MEDLARS) and automatic text processing (SMART). *Technical Report*, 71-115.
- Salton, G. 1975. *A Theory of Indexing*. Bristol, UK: J.W. Arrowsmith.
- 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*. Reading: Addison-Wesley.
- Salton, G. and M. Lesk. 1965. The SMART automatic document retrieval system- an illustration. *Communications of the ACM* 8(6): 391-398.
- Salton, G., and C. Buckley. 1988. Term-weighting approaches in automatic text retrieval. *IP&M* 24 (5): 513-523.

- Salton, G., and M.J. McGill. 1983. *Introduction to Modern Information Retrieval*. New York: McGraw-Hill.
- Salton, G., and C. Buckley. 1990. Improving retrieval performance by relevance feedback. *Journal of the American Society for Information Science* 41(4): 288-297.
- Salton, G., C. Buckley, and M. Smith. 1990. On the application of syntactic methodologies in automatic text analysis. *IP&M* 26 (1): 73-92.
- Salton, G., C.S. Yang, and C.T. Yu. 1975. A theory of term importance in automatic text analysis. *JASIS* 26 (1): 33-44.
- Šauperl, Alenka. 2004. Catalogers' common ground and shared knowledge. *Journal of the American Society for Information Science and Technology* 55(1): 55-63.
- Smeaton, A. F. 1992. Progress in the application of natural language processing to information retrieval tasks. *Computer Journal* 35 (3): 268-278.
- Smeaton, A.F., and P. Sheridan. 1991. Using morpho-syntactic language analysis in phrase matching. In *RIAO 91 Conference Proceedings Intelligent Text and Image Handling* (pp. 414-429). Paris: C.I.D.-C.A.S.I.S.
- Sparck Jones, K. 1973. Index term weighting. *Information Storage and Retrieval* 9:619-633.
- Sparck Jones, K., and J. R. Galliers. 1996. *Evaluating Natural Language Processing: An analysis and review*. New York: Springer.
- Sparck Jones, K., S. Walker, and S. E. Robertson. 1998. A probabilistic model of information retrieval: development and status. <http://www.cl.cam.ac.uk/> (21 September 2009).
- Sparck Jones, K., S. Walker, and S. E. Robertson. 2000a. A probabilistic model of information retrieval: development and comparative experiments, Part 1. *Information Processing and Management* 36:779-808.
- Sparck Jones, K., S. Walker, and S. E. Robertson. 2000b. A probabilistic model of information retrieval: development and comparative experiments, Part 2. *Information Processing and Management* 36:809-840.
- Strzalkowski, T., L. Guthrie, J. Karlgren, J. Leistensnider, F. Lin, J. Perez-Carballo, T. Straszheim, J. Wang, and J. Wilding. 1997. Natural language information retrieval: TREC-5 report. In E. Voorhees & D.K. Harman (Eds.), *Information Technology: The Fifth Text REtrieval Conference (TREC-5)* (pp. 291-313). Gaithersburg, MD: NIST SP 500-238.
- Tenopir, C., J.S. Ro, and S. Harter. 1991. *Full Text Databases*. London: Greenwood Press.
- Van Rijsbergen, C. J. 1977. A theoretical basis for the use of co-occurrence data in information retrieval. *Journal of Documentation* 33:106-119.
- Van Rijsbergen, C. J. 1979. *Information Retrieval*. 2nd ed. London: Butterworths.
- 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.
- Xu, J., and W.B. Croft. 1998. Corpus-based stemming using co occurrence of word variants. *ACM Transactions on Information Systems* 16 (1): 61 -81.
- Indexer Research. 2010. *Cindex Indexing Software for Windows and Macintosh*. <http://indexres.com/soft.php> (accessed 16 June 2010).
- Sky Software. 2010. Sky Index Professional. http://www.sky-software.com/contact_info/ContactInfo.htm (accessed 16 June 2010).

- Calvert, Drusilla and Hilary Calvert. 2010. Macrex, MACREX VERSION 8.
<http://www.macrex.com/> (accessed 16 June 2010).
- Editorium, The. 2010. DEXter for Microsoft Word.
<http://www.editorium.com/DEXter.htm> (accessed 16 June 2010).
- Hofmann, Thomas. 1999. Probabilistic Latent Semantic Indexing. *Proceedings of the Twenty-Second Annual International SIGIR Conference on Research and Development in Information Retrieval*, August 15-19, Berkeley, CA, USA.
www.cs.brown.edu/~th/papers/Hofmann-SIGIR99.pdf (accessed 17 June 2009).
- Hong, J. 2000. *An overview of latent semantic indexing*.
http://www.contentanalyst.com/images/images/Overview_LSI.pdf (accessed 19 June 2010).
- Coyle, Karen. 2008. Machine Indexing. *Journal of Academic Librarianship* 34(6).
http://www.kcoyle.net/jal_34_6.html (accessed 15 June 2010).
- Himwich, Williamina A., Eugene Garfield , Helen O. Field, John M. Whittock, and Sanford V. Larkey. 1955. Welch Medical Library Indexing Project. Sponsored by the Armed Forces Medical Library, The Johns Hopkins University, Baltimore, Maryland. <http://www.garfield.library.upenn.edu/papers/26.html> (accessed 20 July 2009).