مقاله

فاصلهٔ خالی میان واژه‌ها در ذخیره و بازیابی رایانه‌ای اطلاعات

سعید اکبری‌نژاد[1]

چکیده: تمام نظام‌های رایانه‌ایِ ذخیره و بازیابی اطلاعات کتابشناختی به زبان فارسی از نظر فاصلهٔ میان واژه‌ها و عبارت‌ها مشکل دارند. این مشکل ضمن جست‌وجو بر جامعیّت مطلوب یا سرعت جست‌وجو تأثیر منفی دارد. مشکل نوع فاصله را مراکزی مانند فرهنگستان زبان و ادب فارسی باید حل کنند ولی تا آن زمان به دلیل اینکه از زبان کنترل شده استفاده می‌کنیم، کتابداران از استانداردها و مستندها تبعیت می‌کنند، و اطلاعات کتابشناختی به دلیل نوع آن مشکلات زبان طبیعی را ندارد می‌توان ضوابطی برای فاصلهٔ خالی میان واژه‌ها و عبارت‌ها تعیین کرد. منابع و مراکزی که مسئولیت مستندسازی را دارند، بهتر است نوع فاصلهٔ میان واژه‌ها و اصطلاح‌ها را نیز مستند کنند.

یک تجربه

ضمن بازدید از یکی از پایگاه‌های اطلاعاتی داخل کشور جست‌وجویی دربارهٔ "منابع طبیعی" درخواست شد. این پرسش نه به منظور شناسایی منابع و مدارک داخل پایگاه اطلاعاتی بلکه برای آگاهی از سرعت نرم‌افزار درخواست شد. جست‌وجوگر پس از تایپ "منابع طبیعی" و طی مراحل گوناگون اعلام کرد که دربارهٔ این موضوع اطلاعات کتابشناختی ۳۵ مدرک در حافظه موجود است. با توجه به حجم اطلاعات پایگاه و محدودهٔ موضوعی آن به نظر می‌رسید که تعداد مدارک بیش از این تعداد است. پرسیده شد: فقط ۳۵ مدرک؟ جست‌وجوگر پس از مکثی

۱. کارشناس ارشد کتابداری و اطلاع‌رسانی معاونت پژوهشی کتابخانهٔ ملی جمهوری اسلامی ایران

عبارت "منابع × طبیعی" را به رایانه داد[۱]. پس از پایان این مرحلهٔ جست‌وجو عدد ۷۲ روی صفحهٔ نمایش ظاهر شد. بنابراین در یک نظام محلی و در یک رایانه ۱۰۷ عنوان مدرک در مورد "منابع طبیعی" یافت شد. ولی متأسفانه عبارت منابع طبیعی به دو صورت "منابع‌طبیعی" بدون فاصلهٔ خالی میان دو واژه، و "منابع طبیعی"، با یک فاصلهٔ خالی میان دو واژه ضبط شده بود. از مسئول پایگاه اطلاعاتی سؤال شد: مشکل چیست؟ ایشان گفتند: اشکال از پانچ است میان منابع و طبیعی نباید فاصلهٔ خالی درج کرد!

مشکلِ بیان شده، که به نظر بسیار پیش پا افتاده است، مشکل تمام پایگاه‌های اطلاعاتی فارسی است. کافی است با کمی دقت اولین پایگاه اطلاعاتی فارسی را که در دسترس است مورد جست‌وجو قرار داد و به چشم دید که گاه حتی یک رکورد اطلاعاتی صحیح را نمی‌توان در آن یافت. بنابراین در این مقاله به بررسی فضای خالی میان واژه‌ها و عبارت‌ها و راه‌حل‌های احتمالی آن پرداخته شود.

مشکل چیست؟

نرم‌افزارهای موجود در بازار که بیشترین استفاده‌کننده را دارند از تک‌تک کلمات و واژه‌ها ایندکس تهیه می‌کنند و با استفاده از این ایندکس‌هاست که جست‌وجو انجام می‌شود[۲]. مبنای شناسایی واژه برای رایانه نیز فضای خالی میان واژه‌ها و عبارت‌هاست. به این ترتیب برای این نرم‌افزارها عبارتِ "به‌فروش می‌رسد" دارای سه واژه است، همین نرم‌افزارها "به‌فروش می‌رسد" و "بفروش می‌رسد" را دو واژه می‌دانند، و باز هم همین نرم‌افزارها "به‌فروش می‌رسد"، "به فروش می‌رسد" و ... را یک واژه می‌داند. مطمئناً برای این نرم‌افزارها عبارت "منابع طبیعی ایران در قرن بیستم" شش واژه دارد ولی "منابع‌طبیعی ایران در قرن بیستم" یک واژه است.

بهتر است با ذکر مثالی مشکلاتی را که این گونه فاصله‌های خالی ضمن جست‌وجو به وجود می‌آورند روشن‌تر کرد. فرض کنیم به جای بانک اطلاعاتی با بیست هزار رکورد بانک اطلاعاتی‌یی با هزار رکورد در اختیار ماست و به جای ۱۰۷ رکورد جست‌وجو شده در مورد منابع طبیعی فقط ۶ رکورد در حافظه موجود است. از این ۶ رکورد، ۲ رکورد، مثلاً رکوردهای ۸ و

طبیعی	۴، ۹، ۱۵، ۶۵
منابع	۴، ۹، ۱۵، ۶۵
منابع طبیعی	۸، ۱۳

تصویر ۱. ایندکس رایانه برای ۶ رکورد فرضی

۱۳، منابع طبیعی را بدون فاصلهٔ خالی به صورت "منابع‌طبیعی" و ۴ رکورد، مثلاً رکوردهای ۴،۹، ۱۵، و ۶۵، با فاصلهٔ خالی میان دو واژه، به صورت "منابع طبیعی" درج کرده‌اند. با توجه به فرض‌های بیان شده ایندکس رایانه‌ای به صورت تصویر ۱ درخواهد آمد.

این ایندکس به ظاهر ساده و این مشکل بسیار کوچک مشکلات پیچیده‌تر زیر را باعث می‌شود:

۱. با توجه به تصویر ۱ اگر جست‌وجوگر "منابع‌طبیعی" را تایپ کند فقط به منابع ۸ و ۱۳ و اگر "منابع × طبیعی" تایپ کند فقط به منابع ۴، ۹، ۱۵، و ۶۵ خواهد رسید، که در هر صورت جامعیّت را از دست داده است. بنابراین می‌توان نتیجه گرفت که چنین نظام‌هایی به دلیل مشکل فضای خالی میان عبارت‌ها همیشه از جامعیّت مطلوب برخوردار نخواهد بود. البته رسیدن به جامعیّت مطلوب را عوامل متعددی نقض می‌کند ولی عامل فضای خالی عاملی افزودن بر آن عامل‌هاست.

۲. جست‌وجوگر به تجربه در می‌یابد که برای حل این مشکل باید عبارت‌های جست‌وجو را بزرگتر و پیچیده‌تر کند و انواع حالت‌های مختلف، با فضای خالی یا بدون فضای خالی، را در فرمول جست‌وجو بگنجاند، تا بتواند دست‌کم یکی از عوامل کاهش جامعیّت را حذف کند. در این حالت عبارت جست‌وجو به صورت زیر در می‌آید[۳]:

(منابع × طبیعی) + (منابع‌طبیعی)

با این شکل جست‌وجو به یکی از ویژگی‌های رایانه خدشه وارد می‌شود. ویژگی مهم رایانه سرعت آن است. با این روش تعداد جست‌وجوها افزایش خواهد یافت و برای هر جست‌وجوی اضافی وقت بیشتری صرف خواهد شد. البته باید اذعان کرد که مثال "منابع طبیعی" فقط دو حالت دارد و فرمول جست‌وجوی آن ساده است. ولی اگر بخواهیم جست‌وجویی انجام دهیم که ۵ واژه داشته باشد برای جلوگیری از کاهش جامعیّت به عبارت جست‌وجوی بزرگی نیاز داریم. مثلاً اگر اطلاعاتی دربارهٔ "منابع طبیعی ایران در قرن بیستم" مورد نیاز باشد و نوع فضای خالی میان واژه‌ها هم دقیقاً مشخص نباشد عبارت جست‌وجو به صورت زیر خواهد بود:

+ (بیستم × قرن × ایران × منابع‌طبیعی) + (بیستم × قرن × ایران × منابع × طبیعی)
... +(قرن بیستم × ایران× منابع× طبیعی) + (قرن‌بیستم × ایران × منابع‌طبیعی)

تصویر ۲

شاید عبارت جست‌وجوی بالا خیلی هوشمندانه نباشد ولی احتمالاً اولین شیوه‌ای است که به ذهن خطور می‌کند. عبارت جست‌وجویی مانند عبارت تصویر ۲ زمانی تحقق خواهد یافت که جست‌وجوگر به تمام احتمالات واقف باشد و تمام آنها را به رایانه بدهد. اگر جست‌وجوگر اعتقاد داشته باشد میان منابع طبیعی فاصلهٔ خالی نیست یا اگر شیوه‌های مختلف را فراموش کند در هر دو صورت جست‌وجو ناقص خواهد بود و جامعیّت مطلوب به دست نخواهد آمد.

رسم‌الخط و به‌ویژه نوع فاصله در آن مشکلات بسیاری را برای نظامهای رایانه‌ای به وجود آورده است که فقط به دو مشکل آن پرداخته شد. مشکل انتقال اطلاعات استاندارد و یکدست، مشکل استفاده‌کننده نهایی از شبکه و مانند آن مشکلاتی هستند که نیاز به بررسی دقیق‌تری دارد که جای آن در مقالهٔ دیگری است.

راه حل چیست؟

خط فارسی صدها سال است که توانسته ارتباط میان افراد را برقرار کند و هیچ مشکل غیرقابل حلی هم نداشته باشد. تمام شاعران و نویسندگان ایرانی با همین رسم‌الخط نوشته‌اند و افکار و آراء خود را به دیگران منتقل کرده‌اند[۴]. هم اکنون نیز اگر بحث در محدودهٔ خواندن و نوشتن باشد باز هم مشکلات پیچیده‌ای با این رسم‌الخط نخواهیم داشت. مشکل از جایی شروع می‌شود و حاد می‌گردد که پای تکنولوژی نوین به میان می‌آید. بنابراین می‌توان نتیجه گرفت که راه حل این مشکل را یا باید در رسم‌الخط و یا در تکنولوژی نوین، رایانه، یافت.

۱. یکی از منطقی‌ترین راه‌حل‌ها آن است که منتظر باشیم فرهنگستان زبان و ادب فارسی رسم‌الخط دقیق، منضبط، و مدون فارسی را ارائه کند[۵]. اگر چنین رسم‌الخطی تدوین گردد حتماً مشکل فاصلهٔ میان واژه‌ها و عبارت‌ها نیز حل خواهد شد. در آن زمان احتمالاً همه یکسان عمل خواهند کرد و مشکلات رسم‌الخط فارسی، از جمله نوع فاصله نیز، کاهش خواهد یافت. ویژگی عمدهٔ این راه‌حل آن است که هم نظام رایانه‌ای و هم کاربر از یک رسم‌الخط تبعیت می‌کنند و میزان ناهماهنگی میان نظام‌های رایانه‌ای و کاربران، از نظر رسم‌الخط و نوع فاصله، در پایین‌ترین حد خواهد بود.

این راه حل یک مشکل اساسی دارد یعنی معلوم نیست چه زمانی رسم‌الخط یکسان شدهٔ نهایی تدوین خواهد شد. در مورد پیوسته‌نویسی یا جدانویسی ده‌ها سال است که بحث و گفت‌وگو است و هنوز به نتیجهٔ قطعی نرسیده است. به جز آن به فرض که در زمانی خاص نتیجهٔ نهایی به دست آمد و رسم‌الخط دقیق فارسی مشخص شد تا همه آن را اجرا کنند باز هم به سالیان بسیاری وقت نیاز است. خلاصهٔ کلام اینکه تا چه وقت باید منتظر این رسم‌الخط بود؟

۲. برخی نرم‌افزارهایی که براساس نرم‌افزار سی.دی.اس. نوشته شده‌اند مدعی هستند که مشکل رسم‌الخط فارسی و به‌ویژه مشکل فضای خالی میان واژه‌ها را حل کرده‌اند. از نظر نرم‌افزاری امکان این مهم هست. یعنی می‌توان نرم‌افزار را به گونه‌ای طراحی کرد که هر گاه پانچ‌کننده به بود یا نبود فضای خالی میان واژه‌ها و عبارت‌ها شک کرد با زدن علامتی میان آن واژه‌ها و عبارت‌ها پانچ را ادامه دهد. مثلاً اگر پانچ‌کننده شک کرد که میان "منابع طبیعی" فضای خالی درج کند یا نکند می‌تواند از علامتی قراردادی، مانند □، استفاده کند. در این صورت ضمن پانچِ اطلاعات عبارت "منابع □ طبیعی" پانچ خواهد شد. کامپیوتر ضمن ایندکس‌سازی و با برخورد با علامت □ واژه‌های طبیعی، منابع، و منابع‌طبیعی را ایندکس می‌کند. ضمنِ جست‌وجو نیز اگر "منابع‌طبیعی" یا "منابع × طبیعی" تایپ شود نتیجهٔ جست‌وجو یکسان خواهد بود، زیرا از حالت‌های مختلف ایندکس تهیه شده است. ضمن چاپ اطلاعات نیز می‌توان دستور داد که هرگاه به علامت □ برخورد شد میان عبارت یا واژه فاصله بگذار یا فاصله نگذار.

همان‌طور که مشاهده شد با این شیوه، در صورت شک، انواع حالت‌ها، با فضای خالی و بدون فضای خالی، در ایندکس جای می‌گیرد. در مورد ترکیب‌های پیچیدهٔ این شیوه منبع موثقی در دست نیست ولی تا همین جا هم تکنیک جالب و کارایی است. ولی نقطه ضعف آن استوار بودن آن بر شک است. حال اگر شکی وجود نداشت چه باید کرد؟ مثلاً پانچ‌کننده یا کتابداری در کتابخانه‌ای یقین داشت که میان منابع طبیعی باید فضای خالی گذاشت و کتابدار و پانچ‌کننده در کتابخانهٔ دیگری یقین داشت که میان منابع طبیعی نباید فضای خالی درج کرد. در این صورت با دو قطب کاملاً متضاد سروکار داریم که هر دو هم یقین دارند. بنابراین میان نظام‌های رایانه‌ای هماهنگی به وجود نمی‌آید. یا مثلاً پانچ کننده‌ای امروز شک داشت ولی فردا به یقین رسید، در آن زمان چه باید کرد؟ یا بر عکس، امروز یقین داشت و فردا شک کرد. یا یقین این افراد از این قطب به آن قطب سیر کند، بدون هیچ شک و شبهه‌ای. آیا با توجه به شرایط گفته شده برای ورود هر بار اطلاعات باید به اطلاعات گذشته مراجعه کرد تا بود یا نبود فضای خالی را یافت و بعد به کار پانچ ادامه داد؟ یا اینکه فایلی به نام فایل شکیّات درست کرد؟

این سؤال‌ها، سؤال مهم‌تری را مطرح می‌کند، در مورد شک و یقین چه کسانی باید اجتهاد کنند؟ کتابدار یا پانچ‌کننده؟ باید تذکر داد که سالیان سال است در مورد مرکب یا بسیط بودن برخی واژه‌ها و عبارت‌ها، به ویژه فعل‌ها، میان استادان طراز اول زبان و ادب فارسی بحث و گفت‌وگو است. با این حساب چگونه می‌توان اجتهاد میان بود یا نبود فاصله را به کتابدار یا پانچ‌کننده واگذاشت؟

از اینها گذشته، برخی کتابخانه‌ها از نرم‌افزارهایی استفاده می‌کنند که امکان استفاده از علامت □ را ندارند، آنها چه باید بکنند؟ آیا اطلاعات آنها همچنان باید غیراستاندارد باشد؟ یعنی هماهنگی میان کتابخانه‌ها و اطلاعات رایانه‌ای آنها در سراسر کشور موکول به یک تکنیک است؟ این سؤال‌ها و ده‌ها سؤال مانند اینها را باید پاسخ داد تا به توانایی و امکان استفاده از این تکنیک اعتماد کرد.

سالیان بسیاری است که رایانه وارد کتابخانه‌ها شده است. در طی این سال‌ها کتابداران به انتظار نشسته‌اند تا مشکل رسم‌الخط فارسی حل شود ولی همان‌طور که مشاهده شد تا رسیدن به راه حل‌نهایی راهی طولانی در پیش است. با توجه به این شرایط آیا کتابداران نمی‌توانند راه‌حل‌هایی ابداع کنند که در کوتاه مدت، و تا رسیدن به رسم‌الخط واحد، مشکل را حل کند؟

اطلاعات کتابشناختی

در ایران و در حوزهٔ کتابداری عمدهٔ اطلاعاتی که ذخیره و بازیابی می‌شود اطلاعات کتابشناختی است و هنوز ذخیره و بازیابی چکیده و متن کامل منابع مرسوم نشده است. بنابراین مشکلات این نوع اطلاعات مورد توجه است. پیش از اینکه ویژگی‌های رکوردهای کتابشناختی مورد بررسی قرار گیرد بهتر است یک رکورد کتابشناختی از نزدیک مورد مطالعه قرار گیرد.

سعدی، مصلح بن‌عبدالله، ـ ۶۹۱؟ق.
[گلستان]
گلستان سعدی/ [به تصحیح محمد علی فروغی]؛ تصحیح متن و شرح لغات از سامیرا فردی تهرانی.ــ تهران: موسسه مطالعات و تحقیقات فرهنگی، ۱۳۵۶.
۲۰۸ ص.: مصور.ــ (گنجینهٔ ادب فارسی؛ ۸).
بهاء: ۱۲۰۰ ریال.
۱. نشر فارسی ـ قرن ۷ق. الف. فروغی، محمد علی، ۱۲۵۶ـ۱۳۲۱، مصصح. ب. فردی تهرانی، سامیرا. ج. عنوان.

تصویر ۳. نمونهٔ یک رکورد کتابشناختی

رکورد تصویر ۳، رکوردی فرضی است و با کمک رایانه می‌توان تمام اطلاعات آن را مورد جست‌وجو قرار داد. اما عملاً تعداد خاصی از اقلام اطلاعاتی مورد سؤال قرار می‌گیرد. اقلام قابل جست‌وجو در فهرست‌برگهٔ تصویر ۳ عبارت است از:

سرشناسه: سعدی؛ مصلح بن‌عبدالله، ـ ۶۹۱؟ق.
عنوان قراردادی: گلستان
عنوان: گلستان سعدی
محل نشر: تهران
ناشر: موسسه مطالعات و تحقیقات فرهنگی
سال نشر: ۱۳۵۶
فروست: گنجینهٔ ادب فارسی؛ ۸
موضوع: نثر فارسی ـ قرن ۷ق.
شناسه‌های افزوده: فروغی، محمد علی و فردی تهرانی، سامیرا

دو ویژگی مهم اطلاعات کتابشناختی عبارت است از:

۱. احتمال اینکه در رکوردهای کتابشناختی با ترکیب‌های پیچیدهٔ زبان فارسی مواجه شویم بسیار ناچیز است. مثلاً احتمال اینکه در فهرستبرگه فعل مرکب موجود باشد بسیار اندک است.

۲. اغلب اطلاعات کتابشناختی را کتابداران مستند می‌کنند. نام نویسندگان و مترجمان و ...، مؤسسات و سازمان‌ها، موضوع‌ها و... از جملهٔ این موارد هستند. و باید توجه کرد که اگر رسم‌الخط فارسی در ذخیره و بازیابی اطلاعات مشکل دارد این مشکل عمدتاً مربوط به ذخیره و بازیابی منابع به زبان طبیعی است ـ زبانی که ضمن گفت‌وگو و نگارش مستند نمی‌شوند و هر کس به هر شکل که دوست داشت در چارچوب‌های زبانی رفتار می‌کند ـ نه زبان اطلاعات کتابشناختی که مستند می‌شوند و کنترل شده هستند.

یکی از ویژگی‌های مهم کتابداری تبعیت کتابداران از منابع مستند است. کتابداران برای هماهنگی فهرستنویسی کتابخانه‌ها نیازمند تبعیت از این منابع هستند.

با توجه به این سه ویژگی و با عنایت به اینکه هنوز وضعیت رسم‌الخط فارسی مشخص نیست به نظر می‌رسد منابع مستند و مراکزی که این منابع را تدوین می‌کنند می‌توانند نوع فاصلهٔ میان واژه‌ها و عبارت‌ها را نیز مستند کنند.

برخی پیشنهادها

۱. در فاصله‌گذاری میان واژه‌ها و عبارت‌ها به صورت‌های فاصله، نیم‌فاصله، یا بی‌فاصله باید یکسان عمل کرد. متاسفانه در اکثر کتاب‌های مستند این مهم رعایت نشده است.

۲. شناخت انواع فاصله‌ها با توجه به نوع حروف کار بسیار مشکلی است، حتی افرادی که تجربهٔ کافی در امر چاپ و حروفچینی دارند در تشخیص نوع فاصله اشتباه می‌کنند. بنابراین فاصله‌های گوناگون میان واژه‌ها و عبارت‌ها در کتاب‌های مستند به گونه‌ای باشد که کتابداران و پانچ‌کنندگان اطلاعات بتواند آن را به سادگی بشناسند.

۳. منابع مرجع و مستند را از نظر فاصلهٔ خالی میان واژه‌ها و عبارت‌ها نیز باید ویرایش کرد. یعنی ویراستاری باید از این دید به منابع مستند نگاه و اشکالات را برطرف کند.

۴. تمام فاصله‌ها نه به دلخواه بلکه با توجه به زبان و دستور زبان فارسی تعیین و قانونمندی آن تدوین شود.

۵. در عبارت‌های اسمی، که در کتابداری بسیار کاربرد دارد، اکثراً کسره نشان دهندهٔ فاصلهٔ خالی میان واژه‌هاست.

۶. لازم نیست میان مقلوب عبارت‌های اسمی فاصله‌ای درج کرد. مانند زردکوه، سفیدرود.

۷. انواع پیشوندها و پسوندها یا پیوسته یا بدون فاصله نوشته می‌شود. بنابراین بیشتر یا بیش‌تر و بی‌نهایت یا بینهایت درست است.

۸. برخی عبارت‌های اسمی با اینکه گاه باکسره و گاه بی‌کسره خوانده می‌شود، ولی با فاصله نوشته می‌شوند، مانند: علوم اجتماعی، علوم تربیتی و ... در این مورد تکنیک نرم‌افزارهای سی.دی.اس کارساز است.

۹. قبل و بعد از حرف ربط نیاز به فاصلهٔ خالی است. مانند: مواد دیداری و شنیداری.

۱۰. بهتر است میان عبارت‌ها و واژه‌های لاتین که دقیقتاً منعکس‌کنندهٔ لفظ خارجی است فاصله‌ای درج نکرد. بنابراین لی‌فارمینک، سوپرساپ، و کوکلوکس‌کلان درست است نه لی فارمینک، سوپر ساپ، و کوکلوکس کلان(۶).

۱۱. در مورد اجزای نام، اصطلاح، یا عبارتی که در فارسی پذیرفته شده است درج فاصله ضروری است. مانند: استارت موتور دیزل

۱۲. بهتر است میان نام افراد مطابق فاصله‌گذاری همان زبان رفتار کرد. مانند آدام اسمیت.■

یادداشت‌ها

۱. علامت ضربدر برابر با ضرب منطقی، AND، در جبر بول فرض شده است.

۲. سعید اکبری‌نژاد. "نقش قابل مقلوب و عبارتهای جستجو در بازیابی اطلاعات". *پیام کتابخانه*. بهار ۷۱، ص. ۶۳-۷۲.

۳. علامت جمع برابر با جمع منطقی، OR، در جبر بول فرض شده است.

۴. سعید عریان. "فراگیری زبان یا فراموشی میراث مکتوب". *اطلاعات*. سه‌شنبه ۲۱ اسفند ۱۳۶۹. ص. ۷.

۵. عباس حری. "کامپیوتر و رسم‌الخط فارسی". *فصلنامهٔ پیام کتابخانه*. سال سوم، شماره اول، بهار ۷۲، ص. ۶-۱۱.

۶. *سرعنوان‌های موضوعی فارسی*. ویراستار پوری سلطانی و کامران فانی. ویرایش ۲. تهران: کتابخانه ملی جمهوری اسلامی ایران، ۱۳۷۴.