پژوهشگاه اطلاعات و مدارک علمی ایران

فصلنامه علوم و فناوری اطلاعات

دورهٔ ۲۱شماره‌ٔ ۲ (زمستان ۱۳۸۴) صفحه:۴۷-۶۹

روشی برای رفع چالش‌های محتواکاوی در وب‌های فارسی زبان

۱) سید مجتبی شهیدی۲) محسن صدیقی۳)کامران زمانی‌فر

چکیده: زبان فارسی از دو منظر برای ما ایرانیان دارای اهمیت است. اول آن‌که این زبان با تاریخ و فرهنگ و تمدن ما پیوندی دیرینه و ناگسستنی دارد و دوم آن‌که زبان فارسی زبان رسمی کشور و ابزار مبادلهٔ اندیشه‌ها وایده‌ها در عرصهٔ علمی و فرهنگی این مرزوبوم محسوب می‌گردد. رشد علمی و فنی و فرهنگی ما در گرو برقراری ارتباط زبانی و کلامی با دنیای الکترونیکی عرضهٔ دانش و فرهنگ است که وب نام دارد و این میسر نیست جز با تقویت کیفی زبان فارسی مورد استفاده در این دهکدهٔ جهانی. اما زبان فارسی، در تلاقی با جهان الکترونیکی، بخصوص از بَعد رسم‌الخط، دارای دشواری‌هایی است که کاوش در محتویات آن را دچار کم‌کیفیتی می‌نماید. این مقوله مستلزم تمهیداتی چند است تا زبان فارسی را به زبانی مناسب برای پهنهٔ الکترونیکی دادوستد دانش ـ وب ـ تبدیل نماید. مقالهٔ حاضر تلاشی است درجهت مرتفع‌سازی چالش‌های کاوش در وب‌های فارسی‌زبان که از دیدگاه رسم‌الخطی، با استفاده از نمایه‌سازی فارسی و دیدگاه مفهومی، با استفاده از انتولوژی قابل بحث هستند.

کلیدواژه‌ها: انتولوژی، نمایه‌سازی فارسی، کاوش وب‌های فارسی، وب‌کاوی، نرم‌افزار خزنده، محتواکاوی، رسم‌الخط فارسی، نمایه‌سازی.

۱. مقدمه

اهمیتی که وب فارسی به عنوان رسانه‌ای مستقل و مؤثر در دنیای ارتباطات ایرانیان پیدا کرده، انکارناپذیر است. به نظر می‌رسد که اکنون روآوردن برخی از روزنامه‌نگاران، پژوهشگران، دانشجویان، ... به وب فارسی و استفادهٔ خبری، علمی، ... از مطالب آن‌ها نیز موجب تقویت نقش رسانه‌ای وب فارسی شده است.

ولی با توجه به ماهیت خاص رسم‌الخط فارسی که آن را برای سیستم‌های رایانه‌ای نامناسب نموده، امروزه مشکلات بسیاری بر سر راه دانش‌پژوهان و بطورکلی استفاده‌کنندگان از وب‌های فارسی‌زبان وجود دارد. نبود حروف صدادار در فارسی به‌صورت یک موجودیت مجزا از یک طرف، و وجود صداهای یکسان با نمادهای حرفی گوناگون از طرف دیگر، باعث بروز چالش‌های جدی در امر نمایه‌سازی این زبان شده است. به‌نظر می‌رسد تلاش‌هایی لازم است تا زبان زیبای فارسی را برای حضور در عرصهٔ الکترونیکی دانش، آماده‌تر نماید.

۲. پیشینهٔ تحقیق و تعاریف ابتدایی

محتواکاوی وب[۱] یکی از سه شاخهٔ وب‌کاوی است که در واقع، کشف اطلاعات مفید از مستندات و داده‌های ساخت‌یافته و نیمه‌ساخت‌یافته و غیرساخت‌یافتهٔ وب می‌باشد. یک شاخهٔ دیگر این مقوله، ساختارکاوی وب[۲] است که به کشف مدل‌های پس‌زمینه‌ای حاکم بر ساختار فراپیوندهای وب می‌پردازد و هدف آن، ایجاد اطلاعاتی در رابطه با تشابه یا ارتباط بین سایت‌های مختلف وب یا مواردی از این دست می‌باشد. شاخهٔ دیگر آن، کاربردکاوی وب می‌باشد که سعی می‌کند از تعاملات کاربر با وب، اطلاعاتی کسب کند و از آن‌ها به‌صورت سابقه‌ای در مراجعات بعدی کاربر، سود ببرد.

در زمینهٔ محتواکاوی وب، نرم‌افزارهای خزنده[۳] به گشت‌وگذار در اقیانوس وب می‌پردازند و اقدام به نمایه‌سازی واژگان در پایگاه داده‌ای خود می‌نمایند و این نمایه در زمان جستجوهای کاربر، مورد استفادهٔ موتورهای کاوش قرار می‌گیرد. نمونهٔ بارز این روش، موتور کاوشگر «گوگل» است(Chakrabarti,1999).

در همین راستا ابزارهایی همچون«فاستوس»[۴]، در خلال این ماموریت، با هدف کشف گروه‌های مختلف واژگان (مانند اسامی، افعال، ترکیبات وصفی و اضافی،...) به تجزیه و تحلیل متون می‌پردازند و به این ترتیب به کشف دانش از محتویات وب کمک می‌کنند. این روش هم‌اکنون برای زبان‌های انگلیسی و ژاپنی پیاده‌سازی شده و به‌صورت بالقوه برای دیگر زبان‌ها قابل استفاده است (Feiyu,2001).

از طرف دیگر استفاده از انتولوژی[۵] در وب برای بهینه‌سازی کاوش در وب پیشنهاد می‌گردد. انتولوژی، یک فرهنگ واژگان مشترک بر اساس موضوع سایت تعریف می‌کند

تا برای استانداردسازی ارائهٔ مفاهیم آن به منظور قابل‌تفسیر شدن توسط ماشین، مورد استفاده قرار گیرد. انتولوژی یک جزء کلیدی در وب مفهومی[۶] است (Heflin,2000). شخصی‌کردن وب[۷] از دیگر روش‌هایی است که برای کاوش در وب، سودمند است. نمونهٔ این روش را در My Yahoo می‌توان مشاهده کرد.

یکی دیگر از راه‌های کاوش در حجم زیاد و غیرساخت‌یافته از اطلاعات وب، استفاده از پایگاه داده‌ای چندلایه[۸] (MLDB) است. هر لایه از این پایگاه داده، عمومیت بیش‌تری از لایهٔ قبلی دارد. همهٔ لایه‌ها بجز پایین‌ترین لایه (که خود وب است)، قابل کاوش توسط یک زبان پرس‌وجو (مثل «اس‌کیوال») می‌باشد (Osmar,2002). در پیاده‌سازی روش‌های ساختارکاوی وب، از تئوری «گراف» وب بهره‌مند خواهیم شد که در ایجاد دید ارزشمند در الگوریتم های جستجو، کشف ارتباطات، ... مؤثر است. در خصوص روش‌های کاربردکاوی وب، ناوبری کاربر در وب توسط مدل‌های ریاضی مارکف، و براساس میزان تجربهٔ کاربر و داشتن یا نداشتن راهنمای سایت، تجزیه و تحلیل می‌گردد (Velasquez,2003).

۳. خصوصیات وب‌های فارسی از نظر زبان

نبود استاندارد و شناور بودن ویژگی‌های رسم‌الخط و مفاهیم در زبان فارسی موجب گردیده که تقریباً به تعداد صفحات وب فارسی، سبک و سیاق نگارشی برای این زبان به‌کار رفته باشد. اما خصوصیات مشترک اکثر وب‌های فارسی‌زبان را می‌توان چنین ارزیابی نمود:

الف) نگارش برخی از وب‌های فارسی، زبان غیررسمی یا محاوره‌ای است.

ب) در وب‌های فارسی، بخصوص در متون علمی، اغلب از واژه‌های بیگانه، به‌کرّات استفاده می‌شود و بعضی از آن‌ها بارسم‌الخط زبان اصلی نوشته می‌شوند.

ج) رسم‌الخط وب‌های فارسی اصولاً غیراستاندارد و متغیر است و اغلب در معرض نوآوری است.

ه) نوشته‌های وب‌های فارسی حاوی غلط‌های تایپی و نگارشی نسبتاً زیادی است، هرچند که اغلب وب‌های فارسی مهم و پرخواننده، نگارش قابل‌قبولی دارند.

و) رسم‌الخط وب‌های فارسی، تابع محدودیت‌های محیط الکترونیکی و عدم تطبیق آن با الزامات خط فارسی است (اشرف زاده،۱۳۸۳).

۴. ابزارهای جستجو در وب‌های فارسی

در حال حاضر ابزارهای کاوش مختلفی در ایران ظهور کرده‌اند. اما ابزارهای جستجویی که امکان جستجوی اطلاعات به زبان فارسی را در اختیار قرار می‌دهند، محدودند. از طرف دیگر، امکانات و قابلیت‌های آن‌ها برای بازیابی مؤثر و مناسب اطلاعات، متغیر هستند. برخی از ابزارهای کاوش با امکانات جستجوی فارسی عبارت‌اند از «ان‌پی‌ایران»[۹]، «ایرانهو»[۱۰]، «ایران‌مهر»[۱۱]، «پارسیک»[۱۲]، «گوگل»[۱۳]. بجز سایت «ان‌پی‌ایران»، دیگر سایت‌ها دارای واسط جستجوی فارسی هستند و بجز «پارسیک»، هیچیک از ابزارهای موجود کاوش فارسی، چالش‌های زبان فارسی را با هدف بهینه‌سازی کاوش فارسی، فراروی خود قرار نداده‌اند و «پارسیک» نیز تنها مشکل کاراکترهای فارسی با یونیکدهای مختلف را حل نموده است.

در بین ابزارهای کاوش فوق، تنها موتور کاوش «گوگل» دارای برنامهٔ روبات به منظور شناسایی و نمایه‌سازی صفحات یا سایت‌های وب به زبان فارسی و نمایه‌سازی خودکار می‌باشد و قادر است صفحات فارسی را در قالب یونیکد، شناسایی و در پایگاه خود نمایه کند. سایت «پارسیک» نیز از پایگاه «گوگل» برای جستجو و بازیابی اطلاعات استفاده می‌کند. به تعبیر دیگر، ۴ ابزار کاوش دیگر، راهنمای موضوعی به شمار می‌آیند و انسان، فرایند شناسایی، بررسی و نمایه‌سازی سایت‌ها یا صفحات وب را بر عهده دارد (کوشا،۱۳۸۱).

به‌نظر می‌رسد که جای یک ابزار کاوش قوی ملی که با نظارت سازمان‌های انفورماتیکی و انجمن‌های زبان‌شناسی فارسی تهیه گردیده و منطبق با نیازهای اطلاعاتی کاربران اینترنت در ایران باشد و در آن، چالش‌های رسم‌الخطی و مفهومی فارسی و مشکلات ناشی از آن‌ها در نظر گرفته شده باشد، خالی است.

۵. مشکلات ومحدودیت‌های وب‌کاوی در سایت‌های فارسی‌زبان

در دهه‌های اخیر بیش‌ترین اختلاف نظر در باب شیوهٔ املای کلمات فارسی، بر سر جدانویسی یا پیوسته‌نویسی کلمات مرکب بوده است. فرهنگستان زبان و ادب فارسی در این باب راه میانه را برگزیده و کوشیده است تا فقط مواردی را که جدانوشتن یا پیوسته‌نوشتن آن‌ها ضروری است، تحت قاعده و ضابطه درآورد و شیوهٔ نگارش بقیهٔ کلمات مرکب را به ذوق و سلیقهٔ نویسندگان واگذار کند [فرهنگستان،۱۳۸۲].

بعضی چالش‌های زبان فارسی در رایانه و بخصوص در اینترنت که باعث تفاوت در نتیجهٔ جستجو در وب یا وب‌کاوی می‌شود از قرار زیر است:

الف) تنوع در نحوهٔ استفاده از «می» چسبان و غیر چسبان، مثل کلمات «می‌تواند» و «میتواند».

ب) تنوع در نحوهٔ به‌کاربردن«ها»ی چسبان و غیر چسبان، مثل «آن‌ها» و «آنها».

ج) به‌کاربردن بعضی پیشوندها و پسوندها، مثل «همین‌که» و «همینکه» یا «هیچ‌یک» و «هیچیک» یا «راه‌گشا» و «راهگشا».

د) به‌کاربردن همزه به‌صورت‌های مختلف، مثل «مسؤول» و «مسئول»، یا «مسأله» و «مسئله».

ه) استفاده کردن یا نکردن از «ه»، برای کلمات مختوم به «ها»ی بیان حرکت در حالت مضاف، مثل «خانهٔ مسکونی» و «خانه مسکونی».

و) تنوع در استفاده از «ی» در کلمات عربی مختوم به «ا»، مثل «موسی» و «موسا».

ز) تنوع املایی بعضی کلمات که همگی درست هستند، مثل «اتاق» و «اطاق».

ح) استفاده از کلمات اروپایی به‌صورت زبان اصلی یا ترجمهٔ فارسی، بخصوص در متون علمی، مثل «update» و «به‌روزآوری».

ط) استفاده کردن یا نکردن از جمع مکسر برای بعضی کلمات.

ی) تبدیل کلمات اروپایی به رسم‌الخط فارسی با همان تلفظ اصلی، مثل «سورس» به جای «source».

ک) استفاده از «ا» و «آ» به‌جای هم، مثل «فرایند» و «فرآیند».

ل) استفاده کردن یا نکردن از اِعراب برای کلمات.

به‌عبارت دیگر، یک کاربر ممکن است در جستجوی خود در وب کلمهٔ کلیدی خاصی را به‌کار برد، ولی در صفحات وب چنین کلمه‌ای به‌کار نرفته باشد و به جای آن (با توجه به مواردی که در مورد تنوع کاربری کلمات، بحث شد)، کلمهٔ مشابهی ثبت شده باشد؛ در نتیجه بسیاری از صفحات وب مطلوب کاربر، در مجموعهٔ بازیابی‌شده حضور نخواهند داشت.

۶. روش‌هایی برای بهبود کاوش در وب‌های فارسی

۶ـ۱. انتخاب مناسب سرعنوان‌های موضوعی در وب‌های فارسی

پیدا کردن اصول و معیارهای موضوع‌سازی ذهنی و فرایندی که در ذهن کاوشگران اطلاعات در هنگام بیان موضوعات، برای پاسخ‌یابی ماشینی، روی می‌دهد یک فرایند پیچیده، مهم و اثرگذار در جریان تهیهٔ سرعنوان‌های موضوعی است. از طرفی ترکیب‌بندی عبارات کاوش با یک زبان مشترک بین انسان و ماشین، از جمله مسائلی است که همیشه متخصصان بانک‌های اطلاعاتی و کاوشگران اطلاعات را دچار مشکل می‌سازد. به‌همین دلیل و با توجه به ساختار بانک‌های اطلاعاتی، حوزهٔ موضوعی کاوش، میزان آگاهی‌های عمومی‌کاوشگر، زبان تخصصی رایج در میان ورزیدگان یک رشتهٔ خاص موضوعی، مسائل و مشکلات زبانی، ساختار اصطلاحنامهٔ به‌کار گرفته‌شده در بانک اطلاعاتی، و ... است که راهبردهای کاوش، طراحی و اجرا می‌شوند. در این مسیر، سرعنوان‌های موضوعی، نقش عمده‌ای دارند. حل این مسائل می‌تواند به پیدا کردن راه‌حل‌های مؤثر برای سرعنوان‌های موضوعی بینجامد (بیگلو،۱۳۸۲).

۶ـ۲. استمداد از علم اصطلاح‌شناسی در نمایه‌سازی ماشینی

توجه به ضرورت روزآمد بودن واژگان علمی و تخصصی و لزوم کنترل ورود اصطلاحات بیگانه، امری است که ما را به استمداد از علم اصطلاح‌شناسی وامی‌دارد. در این خصوص «حسینی» پژوهشی ارائه کرده است که به‌جهت اشاره به تمهیدات وی در خصوص تشکیل یا بهینه‌سازی اصطلاحنامه‌ای مناسب برای نمایه‌سازی ماشینی، شمه‌ای از آن در اینجا ذکر می‌شود:

الف) کنترل مترادف‌ها و شبه‌مترادف‌ها به‌صورت ارجاع مترادف‌های غیرمرجح به اصطلاح مرجح؛

ب) هدایت کاوشگر از مفاهیم و اصطلاحات اخص به اعم (یعنی نزدیک‌ترین اصطلاح)؛

ج) حصول جامعیت با ارائهٔ روابط ساختاری مفاهیم اعم از سلسله‌مراتبی یا غیر سلسله‌مراتبی، و توسعهٔ کاوش با ارائهٔ طبقه‌های دارای ارتباط بسیار نزدیک تا از این طریق، مانعیت نیز با پیشنهاد اصطلاحات اخص، بهبود یابد؛

د) نظارت بر شکل دستوری، املایی، جمع و مفرد، اختصارات، و شکل مرکب اصطلاح؛

ه) گزینش از بین دو یا چند مترادف موجود برای بیان یک مفهوم؛

و) تصمیم‌گیری در خصوص پذیرش و نحوهٔ برخورد با انواع خاصی از اصطلاحات نظیر «واژه‌های قرضی»[۱۴]، «واژه‌های عامیانه»[۱۵]، اسامی تجاری، و اسامی خاص؛

ز) محدود کردن معانی یک اصطلاح که در یک فرهنگ ممکن است با توضیحات گوناگون همراه باشد.

توصیه‌های اضافی در خصوص تشکیل اصطلاحنامه به شرح زیر است:

الف) واژه‌های قرضی: واژه‌هایی که از زبان‌های دیگر قرض گرفته شده و در زبان قرض‌گیرنده تثبیت شده‌اند. چنانکه ترجمهٔ این اصطلاحات وجود داشته باشد ولی به‌طور رایج مورد استفاده قرار نگیرد، با اصطلاح ترجمه‌شده باید به‌صورت اصطلاح نامرجح رفتار کرد.

ب) «نوواژه‌ها»[۱۶]، اصطلاحات عامیانه و زبان حرفه‌ای: چنانچه جایگزینی که به‌طور گسترده توسط کاربران مورد استفاده قرار گیرد وجود نداشته باشد، نوواژه، اصطلاح عامیانه یا حرفه‌ای، به‌عنوان توصیفگر پذیرفته می‌شود.

ج) اسامی عامیانه و اسامی تجاری: توصیه می‌شود در جایی که اسم عامیانهٔ معادلی وجود دارد، از آن به‌جای اسم تجاری استفاده گردد.

د) اسامی مشهور و اسامی علمی: انتخاب بین این دو بر اساس احتمال بیش‌تر برای استفاده توسط کاربران می‌باشد.

هـ) اسامی مکان‌ها: در جایی که برای یک کشور یا منطقهٔ جغرافیایی درون یک جامعهٔ تک‌زبانی، بیش از یک اسم انتخاب می‌گردد، باید اسمی را به‌عنوان اصطلاح مرجح تعیین کرد که نزد کاربران، آشناتر است.

و) اسامی خاص مؤسسات، افراد، و ...: میزان نیاز به دستیابی به اسامی خاص بر اساس حوزهٔ عملکرد اصطلاحنامه، گنجاندن اسامی را در اصطلاحنامهٔ اصلی تعیین می‌کند.

ز) همنام‌ها و هم‌آواها: منظور کلماتی هستند که دارای املای یکسان و معانی متفاوت، یا دارای تلفظ یکسان و معانی متفاوت می‌باشند. روش معمول ابهام‌زدایی در چنین مواردی، اضافه کردن توضیحگری است که داخل پرانتز قرار می‌گیرد.

ح) مترادف‌ها: انتخاب مترادف‌ها باید بر اساس نیازهای کاربران باشد و این کار از نقطه نظر رواج و تخصص، صورت می‌گیرد.

ط) شبه‌مترادف‌ها: پذیرش شبه‌مترادف‌ها، از حوزهٔ موضوعی زیرپوشش اصطلاحنامه تأثیر می‌پذیرد، مثل «افراد بااستعداد» و «تیزهوشان». شبه‌مترادف‌ها ممکن است شامل متضادها هم باشند، مثل «سوادآموزی» و «بیسوادی» (حسینی،۱۳۸۳).

۶ـ۳. تعریف یک استاندارد برای مفاهیم و رسم‌الخط فارسی در وب:

همان‌طور که گفته شد، یک تفاوت زبان فارسی با زبان انگلیسی (و زبان‌های هم‌ارز)، تنوع املایی یا رسم‌الخطی کلمات آن است. به‌عبارت دیگر، در زبان انگلیسی نیز تنوع در مفهوم کلمات وجود دارد؛ یعنی برای بعضی مفاهیم، ممکن است از کلمات متنوعی استفاده شود، مثل کلمات Hello و Hi که دارای مفهوم یکسانی هستند. اما در فارسی، علاوه بر وجود کلمات متنوع برای مفاهیم یکسان (مثل «کامپیوتر» و «رایانه»)، تنوع در رسم‌الخط یک کلمه نیز فراوان به‌چشم می‌خورد. به‌عبارت دیگر، در حالی که شما به‌دنبال صفحات محتوی کلمهٔ «امپراتور» می‌گردید، ممکن است کلیهٔ صفحات محتوی کلمهٔ «امپراطور» را از دست بدهید.

به‌نظر می‌رسد که در تشکیل صفحات وب فارسی، جای یک استاندارد حاکم بر عملکرد تألیف نویسندگان وب، خالی است، استانداردی که در انتخاب بعضی کلمات دارای چندین رسم‌الخط و حتی انتخاب بعضی کلمات که بر مفاهیم متنوعی دلالت دارند، تکلیف را روشن کند و مؤلفان را از طرفی ترغیب به انتخاب گونهٔ زبانی مناسب برای تضمین کیفیت ارتباط و انتقال مؤثر پیام، و از طرف دیگر مؤظف به حفظ سلامت زبان و رعایت استانداردهای آن به‌عنوان یک وظیفهٔ رسانه‌ای نماید.

ایجاد و گسترش چنین استانداردی، با هماهنگی انجمن‌ها و شوراهای علمی یا صنفی انفورماتیک در ایران، به‌عهدهٔ «فرهنگستان زبان و ادب فارسی» است. تعویق در تنظیم این استاندارد، با توجه به رشد روزافزون وب‌های فارسی‌زبان، هزینه‌های جبران‌ناپذیری دربرخواهد داشت.

۶ـ۴. استفاده از مفرد و جمع در نمایه‌سازی

استفاده از اسامی جنس، نحوهٔ جمع بستن کلمات به‌صورت باقاعده یا بی‌قاعده معضلی است که در نمایه‌سازی واژگان فارسی رخ می‌نماید. در این خصوص «سمایی» در مقالهٔ خود قواعدی را برای نمایه‌سازی واژه‌های مفرد و جمع ارائه داده است که ذکر آن‌ها خالی از لطف نیست:

الف) از آنجا که کلیدواژه‌ها در زبان تخصصی به‌کار می‌روند و در بین اهل فن رایج و جاری‌اند، گاه اتفاق می‌افتد که صورت جمع، مرسوم باشد. در این حالت بهتر است که از صورت جمع استفاده شود، نظیر ترکیب «آثار باستانی». نکته‌ای که در این باره ذکرکردنی است، شیوهٔ جمع بستن اسامی در این موارد است؛ بدین معنا که گاهی نوع پسوند جمع یا شیوهٔ جمع بستن باعث می‌شود که اصطلاح به‌دست‌آمده، با سنت رایج در حوزهٔ تخصصی منطبق نباشد. مثلاً چنانچه لفظ «اثر» با «ها» جمع بسته و ترکیب «اثرهای باستانی» ساخته شود، با اصطلاح رایج اهل فن متفاوت خواهد بود. بعلاوه این که در برخی موارد، شیوهٔ جمع بستن باعث تفاوت در معنا می‌شود. «اثرها» در برخی بافت‌ها معادل «آثار» نیست. «آثار» در ترکیب با «باستانی» شامل خرابه‌ها و بناها و اشیای به‌جامانده از زمان قدیم می‌شود، در حالی که از لفظ «اثرها» بیش‌تر معنای ردّ و نشان تداعی می‌شود.

ب) گاهی صورت مفرد کلمه، معنایی متفاوت از معنای جمع دارد. این مسئله اغلب در کلمات عربی مصطلح در فارسی وجود دارد. ترکیب «مصالح راه‌سازی» از این دست است. «مصالح» به‌معنای مواد لازم برای ساختن بنا است، در حالی که معنای صورت مفرد آن یعنی «مصلحت» ــ به نقل از فرهنگ معین ــ از این قرار است: «آنچه که صلاح و سود شخص یا گروهی در آن باشد».

ج) در مواردی، با این که صورت مفرد و جمع کلمه، معنایی مشترک دارند استعمال صورت مفرد در زبان رایج نیست. به همین علت ترکیب‌هایی نظیر «منسوجات نظامی» و «الیاف کربنی» را نمی‌توان به‌شکل مفرد آورد و به جای منسوجات لفظ «منسوج» و به جای الیاف لفظ «لیف» را قرار داد.

د) در برخی واژه‌ها، صورت جمع توسّع معنا پیدا کرده و از این طریق، ارتباط صورت جمع و مفرد ضعیف شده است. در واژه‌ای نظیر «مهمّات» این اتفاق رخ داده و ارتباط «مهمات» با «مهم» از این دست است.

ه) بعضی ترکیب‌ها نظیر «ماشین‌آلات» وجود دارند که نه‌تنها نمی‌توان قسمت جمع آن‌ها را به شکل مفرد آورد، بلکه در مجموع، یک واحد نحوی ایجاد می‌کنند که از لحاظ معنایی تجزیه‌ناپذیر است.

و) گاهی هم اتفاق می‌افتد که جمع اسم با قاعدهٔ فارسی، در زبان مصطلح نیست و جمع عربی آن رایج است. بدیهی است که در این حالت چنانچه اسم مذکور، کلیدواژه شود یا در ترکیبی به کار رود و نتوان از صورت مفرد آن استفاده کرد، باید شکل جمع عربی آن را به کار برد. در ترکیب «اجزای پل» نمی‌توان به جای «اجزا» لفظ «جزءها» را به کار برد.

ز) برای جمع بستن اسامی لاتین استفاده از پسوند «ها» مرجح است: «کربامات‌ها» [سمایی،۱۳۸۲].

۶ـ۵. استفاده از یک واسط کاوش فارسی برای رفع چالش‌های رسم‌الخطی و مفهومی

استانداردسازی رسم‌الخط فارسی ممکن است در ابتدای تولید اولین صفحات وب فارسی بسیار مفیدتر به‌نظر رسد، ولی در حال حاضر، با وجود تعداد بسیار زیاد صفحات وب فارسی که در هر حال در نبود یک استاندارد ناظر، تولید شده‌اند، چندان سودمند واقع نمی‌شود؛ هرچند که ایجاد آن برای تولید صفحات وب فارسی آتی، لازم است. به‌عبارت دیگر، برای انجام عملیات وب‌کاوی در صفحات وب فارسی کنونی، باید روشی ابداع کرد تا با توجه به چالش‌های بحث‌شده، نتایج مطلوبی از وب‌کاوی در آن‌ها به‌دست آید.

با توجه به بحث‌های قبل، می‌توان دریافت که در کاوش وب، پارامترهایی که نتایج جستجو را برای کاربر مطلوب جلوه می‌دهند، از این قرارند:

الف) جامعیت نتایج[۱۷]: منظور از جامعیت نتایج این است که کلیهٔ صفحات وبی که بر اساس کلمهٔ کلیدی، مطلوب کاربر محسوب می‌گردند، نمایش داده شوند و هیچ صفحهٔ مطلوبی از قلم نیفتد.

ب) مانعیت نتایج: منظور از مانعیت نتایج این است که صفحات وبی اضافه بر نتایج جستجوی مطلوب کاربر ارائه نشود که به‌علت حجم زیاد نتایج، باعث سردرگمی‌کاربر گردد.

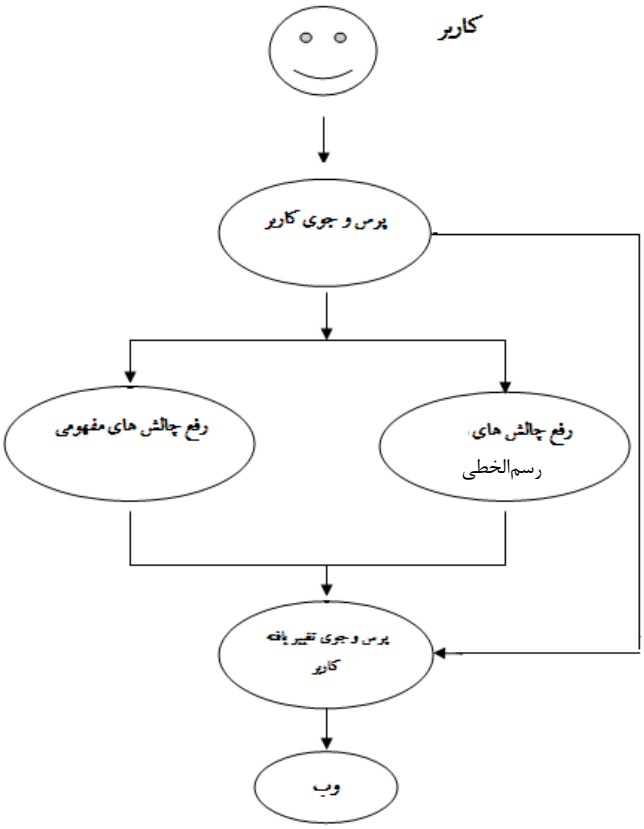
ج) تناسب نتایج[۱۸]: میزان مطلوب بودن نتایج، نسبت به مورد جستجو است که باید حداکثر باشد.

د)سرعت بازیابی: نکتهٔ مهم دیگر در کاوش، زمان صرف‌شده برای جستجو است که باید در حداقل باشد. این پارامتر به میزان ترافیک شبکه، سرعت خدمت‌دهنده‌ها، سرعت پایگاه دادهٔ نمایه‌شده، و موارد سخت‌افزاری وابسته است.

ولی جامعیت، مانعیت و تناسب نتایج، می‌توانند تحت تأثیر زبان استفاده‌شده در نگارش محتوای صفحات، تغییر نمایند، بخصوص در مواقعی که زبان مورد استفاده، زبانی همچون فارسی با چالش‌های رسم‌الخطی فراوان در گسترهٔ امور رایانه‌ای است و بشدت مستعد نتایج بی‌اعتبار و نامناسب می‌باشد. به‌عنوان مثال، کاربر جویندهٔ اطلاعات در بارهٔ «امپراتوری‌های قدیم»، از دیدن صفحات وب حاوی کلمهٔ «امپراطور» در نتایج جستجوی خود، محروم است.

از این رو در خصوص ارتقای کیفیت نتایج کاوش در وب‌های فارسی‌زبان، جای راهکارهایی که پارامترهای مذکور را تقویت نماید خالی است. از این رو، بر آن شدیم تا با ایجاد یک عامل هوشمند، نتایج جستجوها را بهینه کنیم. این‌کار را با اضافه کردن یک واسط هوشمند به موتورهای کاوش یا خزنده‌ها انجام دادیم. این واسط در واقع نقش یک پردازشگر پرس‌وجو[۱۹] را ایفا می‌کند.

این عامل از دو قسمت تشکیل شده است: یک قسمت به مرتفع‌سازی معضلات رسم‌الخط و بهبود بَعد جامعیت نتایج کاوش، و قسمت دیگر به رفع مشکلات مفهومی و بهینه‌سازی تناسب و مانعیت نتایج کاوش می‌پردازد.

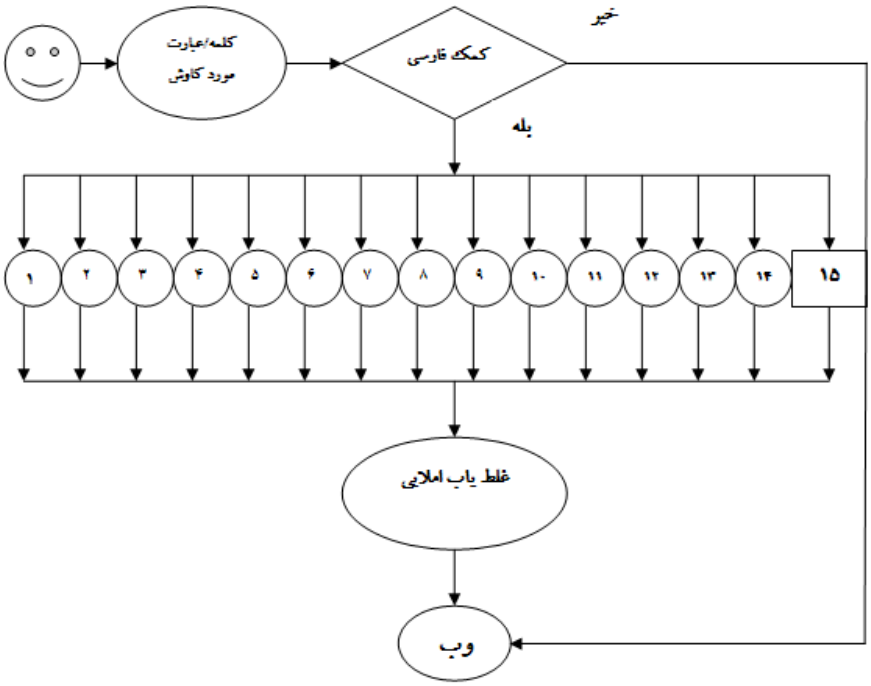


شکل ۱ شمای کلی واسط هوشمند فارسی

۷. واسط فارسی برای بهبود جامعیت کاوش

این قسمت از عامل، از یک پایگاه داده تشکیل شده است که حاوی چندین کلمهٔ معادل برای بعضی کلمات خاص که در چالش‌ها ذکر گردید، می‌باشد. این تناظر می‌تواند مربوط به معادل‌های رسم‌الخطی، معادل‌های مفهومی یا معادل‌هایی به زبان‌های غیرفارسی باشد. بدین‌صورت با عبوردادن کلمات مورد کاوش از این واسط یا با رجوع به این پایگاه داده، عملاً یک کاوش بر اساس یک کلمهٔ کلیدی خاص، منجر به

چند کاوش برای کلمات معادل آن کلمهٔ کلیدی خاص می‌گردد. با این ترفند، صفحات حاوی کلمات معادل از دست نمی‌روند و پارامتر جامعیت، تقویت می‌گردد.

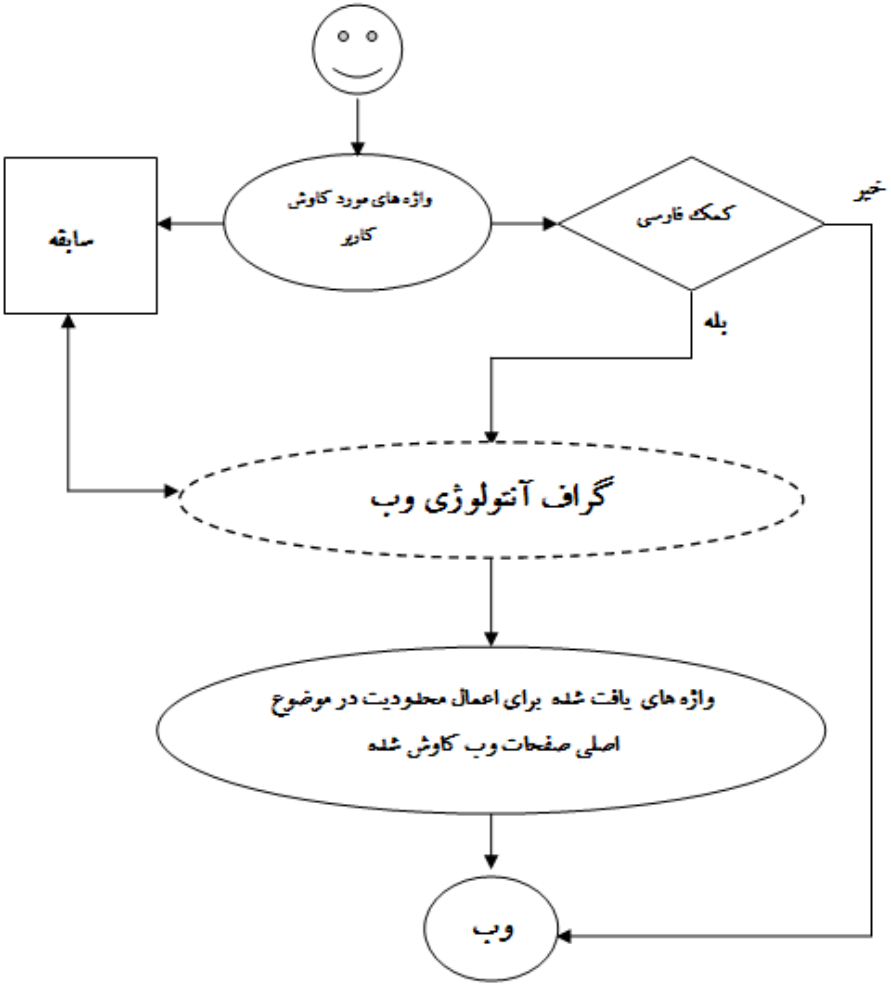


شکل ۲ ساختار واسط کمک فارسی برای بهبود جامعیت

۸. واسط فارسی برای بهبود مانعیت وتناسب کاوش

همان‌طور که در فصل قبل تشریح گردید، انتولوژی می‌تواند با ایجاد ساختار مفهومی مناسب، بر حسب موضوع اصلی سایت وب (مثل وب خدماتی، تولیدی، علمی، ...) کاوش ماشینی را تسهیل کند و با روشن بودن نوع وب از نظر موضوع و ساختار پیوندی آن (شکل گراف وب) با توجه به انتولوژی استانداردی که برای کاوشگر نیز شناخته شده، به‌طور بهینه‌ای می‌توان عملیات کاوش را به ثمر رسانید. به‌عبارت دیگر اگر کاربری در مورد «پایان‌نامه‌های تحصیلی» جستجو کند، موتور کاوشگر، با اطلاع از نوع سایت‌های

دانشگاهی، دقیقاً با مراجعه به پیوندهای مربوطه (مثل «پژوهش‌ها»، «پایان‌نامه‌ها»، ...) در سایت‌های مربوطه، نتایج بهینه‌ای را در زمانی اندک فراهم می‌نماید.



شکل ۳ ساختار واسط کمک فارسی برای بهبود مانعیت

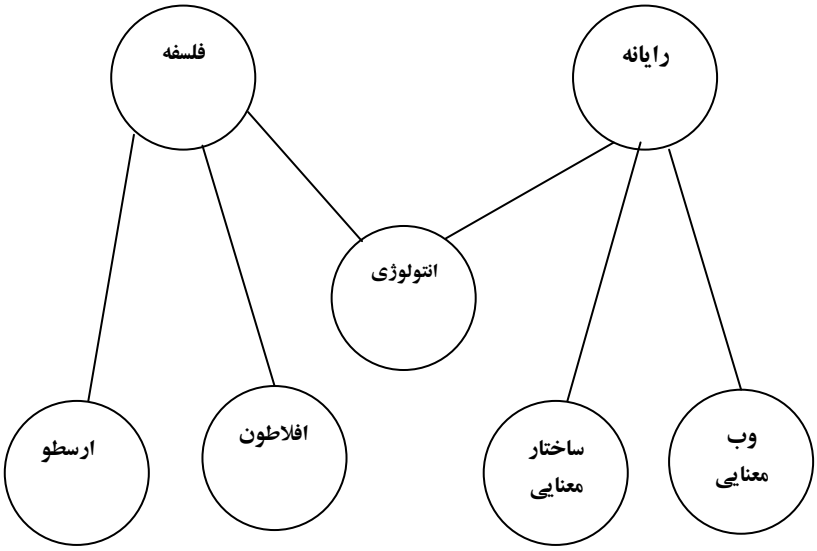
با تعریف یک گراف بدون جهت مفهومی (گراف انتولوژی) محتوی واژه‌های فارسی و ارتباط بین آن‌ها می‌توان گامی در جهت بهبود نتایج کاوش از نظر مانعیت برداشت. گراف عمومی ما شامل اجزای زیر است:

الف) گره‌ها که واژهٔ متناظر با آن‌ها، دلالت بر یک موجودیت فیزیکی یا مفهومی دارد؛ مثل ایران، میز، کشور، حیوان، درخت، اسب،... هریک از این گره‌ها، با یک کد مشخصهٔ گره، یک واژهٔ گویای موجودیت آن و یک عدد نشان‌دهندهٔ فاصلهٔ آن گره تا ریشه است.

ب) پیوند بین گره‌ها که شامل پیوندهای بین آن‌ها است و هر یک، معرف رابطهٔ مفهومی بین واژه‌ها می‌باشد. هر پیوند از نظر ساختار، حاوی یک کد مشخصهٔ پیوند، کد مشخصهٔ گره شروع و کد مشخصهٔ گره پایان می‌باشد.

زمانی که کاربر، کاوش خود را در مورد کلمهٔ خاصی آغاز می‌کند، معمولاً در ابتدا نتایج بسیار زیادی دریافت می‌کند که به‌علت عدم اعمال محدودیت روی کلمهٔ مورد کاوش وی می‌باشد. واسط هوشمند، در مراحل بعدی کاوش، با محدود کردن دامنهٔ جستجو با توجه به کلمات مورد نظر کاربر و بر اساس انتولوژی موجود، سعی در فیلتر کردن نتایج حاصل از کاوش دارد. واسط مفهومی ما با در نظر گرفتن کلمات استفاده شده در کاوش‌های قبلی کاربر، اقدام به یافتن مسیری در گراف عمومی مذکور، بین واژه‌های به‌کاررفته می‌نماید. اگر در گراف مذکور، مسیری با طول قابل قبول یافت شد، واسط، گره‌های موجود در بین این مسیر را شناسایی می‌کند، گره دارای کم‌ترین فاصله تا ریشه را، مبنایی برای فیلتر کردن سایت‌های وب کاوش‌شده قرار می‌دهد و از ارائهٔ سایت‌های دیگر به کاربر، که موضوع آن‌ها با واژهٔ گره اصلی همخوانی ندارد، اجتناب می‌ورزد و بدین طریق مانعیت و تناسب نتایج حاصل از کاوش را بهبود می‌بخشد.

مثلاً اگر کاربر کاوش خود را با واژهٔ «انتولوژی» شروع کند و در مراحل بعدی جستجوی خود، واژه‌های «ساختار معنایی»، «وب معنایی» و ... را مورد کاوش قرار دهد، واسط فارسی اقدام به یافتن مسیری از واژه «انتولوژی» تا واژه «وب معنایی» می‌نماید و در فاصلهٔ این مسیر، به گره اصلی «رایانه» می‌رسد. از این پس واسط، نتایج حاصل از کاوش کاربر را به صفحاتی با موضوع اصلی «رایانه» محدود می‌کند و از ارائهٔ صفحاتی با موضوع اصلی «فلسفه» خودداری می‌ورزد.



شکل ۴ قسمتی از گراف معنایی وب(انتولوژی)

جدول ۱ ساختار و محتویات نمونهٔ جدول گره‌های گراف معنایی در پایگاه داده

شناسه	واژهٔ گره در گراف	فاصله از ریشه
۱	انتولوژی	۱
۲	وب معنایی	۱
۳	رایانه	۰
۴	فلسفه	۰

جدول ۲ ساختار و محتویات نمونهٔ جدول پیوندهای گراف معنایی در پایگاه داده

شناسه	گره شروع	گره پایان
۱	وب معنایی	رایانه
۲	انتولوژی	رایانه
۳	فلسفه	انتولوژی

۸. معماری

این عامل هوشمند، در خصوص هر یک از چالش‌های رسم‌الخطی زبان فارسی رایانه‌ای، رفتار متفاوتی از خود نشان می‌دهد. این رفتارها به‌قرار زیر است:

الف) تنوع نحوهٔ استفاده از «می»، «ها»، پیشوندها و پسوندها: همان‌طور که قبلاً توضیح داده شد، موارد فوق به‌شکل چسبیده یا جدا از کلمه به‌کار برده می‌شوند. از این رو برای رفع چنین مشکلی، می‌توان در واسط هوشمند، با حذف کلیهٔ فواصل خالی موجود در عبارت مورد کاوش، اقدام به جستجو بر اساس دنباله‌ای از حروف همان عبارت، بدون هیچگونه فاصلهٔ خالی نمود.

ب) به‌کاربردن همزه به‌صورت‌های مختلف: برای حل مشکل فوق، در عمل هوشمند مورد بحث، فرآیندی ایجاد می‌گردد که در طی آن، اگر عبارت مورد کاوش حاوی صور مختلف همزه باشد، عملاً کاوش به چندین جستجو برای کلمات مشابه، با حالت‌های مختلف همزه تبدیل می‌شود. به‌عبارت دیگر کاوش کلمهٔ «مسئله» به کاوش برای کلمات «مسئله» و «مسأله» منجر می‌شود. می‌توان با جایگزینی «ی» به‌جای «ء» نیز دامنهٔ کاوش را وسیع‌تر نمود، مثل «رئیس» و «رییس».

ج) استفاده کردن یا نکردن از «ء» در ترکیب‌های اضافی یا وصفی: برای رفع این مشکل، در صورت استفادهٔ کاربر از «ء» در عبارت مورد کاوش خود، واسط هوشمند اقدام به جستجو برای عبارتی فاقد «ء» می‌نماید. در این صورت نتایج جستجو صفحاتی را که در محتوای متن آن‌ها از «ء» استفاده نشده است نیز شامل می‌گردد.

د) استفاده از «ا» و «آ»: در این مورد، واسط به‌محض برخورد با کلمهٔ مورد کاوش که در آن «ا» به‌صورت چسبان یا غیرچسبان به‌کار رفته یا شامل «آ» می‌باشد، جستجو را به کاوش برای کلمات جدیدی که با جایگزینی «ا» با «آ» یا «آ» با «ا» ساخته شده‌اند بسط می‌دهد. در نتیجهٔ کاوش برای کلمهٔ «فرایند» صفحات حاوی کلمهٔ «فرآیند»، از دست نمی‌رود.

هـ) استفاده از اصطلاحنامه برای حل مشکل تنوع املایی کلمات: این معضل شامل تنوع استفاده از «ی» در کلمات عربی مختوم به «ا»، تنوع املایی بعضی کلمات که همه درست هستند، استفاده از کلمات اروپایی به‌صورت ترجمهٔ فارسی، و استفاده کردن یا نکردن از جمع مکسر برای بعضی کلمات می‌باشد و حل مشکل کلیهٔ این موارد، با ایجاد

یک پایگاه داده در سمت خدمت‌گذار انجام می‌گیرد. این پایگاه داده شامل نمایه‌ای از این کلمات و کلمات مترادف می‌باشد. مثلاً کلمه «موسی» به «موسا» و کلمه «کامپیوتر» به «رایانه» متناظر شده است. عامل هوشمند با مراجعه به این پایگاه داده، برای عبارت مورد کاوش کاربر عبارات مشابهی را استخراج می‌کند و کاوش را به جستجو برای این عبارات (علاوه بر عبارت اصلی) بسط می‌دهد.

ایجاد چنین پایگاه داده‌ای با مشاورهٔ انجمن‌ها، بزرگان، و فرهنگستان ادب فارسی انجام می‌پذیرد و به‌روزآوری آن نیز به‌صورت دوره‌ای و با دخالت صاحب‌نظران مذکور صورت می‌گیرد. نمونه‌ای از محتویات این پایگاه داده در جدول زیر آمده است:

جدول ۳ نمونه‌ای از محتویات پایگاه داده‌ای مترادف‌ها

شناسه	واژهٔ اصلی	واژهٔ مترادف
۱	موسی	موسا
۲	امپراتور	امپراطور
۳	Ontology	انتولوژی
۳	انتولوژی	انتالوژی
۳	انتولوژی	هستی‌شناسی
۴	کامپیوتر	رایانه
۴	Computer	کامپیوتر
۵	Source	منبع
۵	Source	سورس

و) تبدیل کلمات اروپایی به رسم‌الخط فارسی با همان تلفظ اصلی[۲۰]: کاربری که به‌دنبال اطلاعاتی در خصوص برنامه‌های «Open Source» در اینترنت می‌باشد، شاید برای همیشه از دسترسی به صفحاتی که در آن‌ها کلمهٔ «سورس باز» به‌کار رفته محروم بماند، یا این که حداقل محکوم به اتلاف زمان زیادی تا رسیدن به چنین کلمه‌ای و به تبع آن، رسیدن به نتایج مطلوب باشد. بنابراین در صورتی که جستجو برای لغت «سورس»، به‌نحوی همزمان با کاوش برای کلمه «Source» ـ حتی بدون اطلاع کاربرـ

انجام پذیرد، می‌توان گفت که هم در سرعت و هم در جامعیت اطلاعات به‌دست آمده، ارتقایی صورت گرفته است.

وظیفهٔ واسط ما در این خصوص این است که با مراجعه به پایگاه داده، کاوش را به کلمهٔ ساخته‌شده بر اساس تلفظ انگلیسی متناظر نیز گسترش دهد. برای انجام فرآیند حل این مشکل به‌صورت خودکار و ضمناً استفاده از پایگاه دادهٔ معتبرتر و روزآمدتر به‌عنوان معیار عملکرد این واسط، می‌توان روشی پیشنهاد نمود که کلمهٔ متناظر تلفظ انگلیسی لغات که با رسم‌الخط فارسی تهیه می‌گردد، با مراجعه به پایگاه‌های دادهٔ بین‌المللی حاوی معادل‌های نمادین، تلفظ کلمات انگلیسی (که در لغتنامه‌های انگلیسی به انگلیسی آمده است)، کلمهٔ مذکور را تهیه کند و سپس کاوش را برای آن انجام دهد.

جدول ۴ اجزا و پردازش‌های مربوط به رفع اشکالات رسم‌الخط

نام جزء	پردازش مربوط	نام جزء	پردازش مربوط
C1	حذف «ء» از عبارت	C7	تبدیل «آ» به «ا»
C2	تبدیل «ؤ» به «ئ» و برعکس	C8	تبدیل «ا» به «آ» در ابتدای کلمات
C3	تبدیل «ئ» به «أ» و برعکس	C9	اضافه کردن «ء» به «ه» در ترکیبات
C4	تبدیل «ؤ» به «أ» و برعکس	C10	حذف اعراب‌ها
C5	تبدیل «یی» به «ئی» و برعکس	C11	تبدیل «ۀ» به «ه ء» یا به «ه» و برعکس
C6	تبدیل "ی" به کاراکتری با یونیکد مشابه	C12	مراجعه به پایگاه دادهٔ واژه‌های مترادف

۹. پیشنهاد

مطالعهٔ حاضر با هدف بهینه‌سازی امکانات جستجو و بازیابی اطلاعات در ابزارهای کاوش با واسط فارسی صورت گرفته است. به‌عنوان پژوهشی دیگر می‌توان تمهیداتی برای کاوش هر چه دقیق‌تر وب‌های فارسی‌زبان، با هدف به‌حداقل‌رساندن تأثیرهای سوء

چالش‌های رسم‌الخط فارسی، اندیشید و از این راه حل‌ها به صورت تلفیقی (سری و موازی) نیز استفاده نمود.

می‌توان نرم‌افزار واسط کمک فارسی مذکور را به‌صورت یک نوار ابزار، برروی مرورگرها نصب و از آن استفاده نمود. از طرف دیگر می‌توان به‌صورت یک نرم‌افزار که بر روی مرورگر نصب شده، به‌صورت پس‌زمینه‌ای، کلمات مورد کاوش را گرفت، بر روی آن‌ها اعمال نظر کرد، و کاوش جدیدی را ترتیب داد.

پژوهشی دیگر می‌تواند در صورت امکان روشی را جستجو کند که گراف معنایی مورد بحث را به‌صورت ماشینی ایجاد و گسترش می‌دهد.

۱۰. منابع

1. Velasquez, J., Hiroshi, Y., & Terumasa, A. (2003). Combinig the Web content and usage mining to understand the visitor behavior in a Web site. In X. Wu & A. Tuzhilin (Eds.), *Third IEEE International Conference on Data Mining (ICDM 2003): proceedings*. Los Alamitos, Calif.: IEEE Computer Society.
2. Chakrabarti, S., van den Berg, M., & Domc, B. (1999). Focused crawling: a new approach to topic-specific Web resource discovery. *Computer networks: Proceedings of the Eighth International World Wide Web Conference*, Canada, 31, 1623-1640.
3. Hobbs, J. R., Appelt, D., Bear, J., Israel, D., Kameyama, M., Stickel, M., & Tyson, M. (1997). FASTUS: A Cascaded Finite-State Transducer for Extracting Information from Natural-Language Text. In E. Roche & Y. Schabes (Eds.), *Finite State Devices for Natural Language Processing* (pp. 383-406). Massachusetts: MIT Press. Retrieved May 16, 2004, from www.isi.edu/~hobbs/fastus-schabes-jul95.pdf
4. Heflin, J., Hendler, J. (2000). Dynamic ontologies on the Web. *Proceedings of 17th National Conference on Artificial Intelligence (AAAI-2000)*, 443-449. Retrieved December 1, 2005, from www.cs.umd.edu/projects/plus/SHOE/pubs/aaai2000.pdf

5. Zaïane, O. R. (2002). *Principles of Knowledge Discovery in Data* , Chapter 9. Retrieved December 31, 2005, from www.cs.ualberta.ca/~joerg/courses/cmput695/fall2003

۶. اشرف‌زاده، بهرام. (۱۳۸۳). *زبان فارسی در وبلاگ‌های فارسی*. دسترسی در اول بهمن، ۱۳۸۴، از سایت زبان فارسی در دنیای ارتباطات:
http://www.persianfarsi.com/articles/zabaneweblog.htm

۷. جعفربیگلو، موسی. (۱۳۸۲). مقایسه فرایند موضوع‌سازی ذهنی جویندگان اطلاعات با ساختار سرعنوان‌های موضوعی فارسی. *فصلنامه علوم اطلاع‌رسانی*، ۱۴(۳و ۴)، ۲۱-۱۲.

۸. حسینی بهشتی، ملوک السادات. (۱۳۸۲). کاربرد اصطلاح‌شناسی و واژه‌گزینی در نمایه‌سازی ماشینی و بازیابی اطلاعات. *فصلنامه علوم اطلاع‌رسانی*، ۱۸(۳و ۴)، ۴۴-۳۱.

۹. خوانساری، جیران. (۱۳۸۲). تکامل وب و مقایسه ابزارهای جستجو در اینترنت. *فصلنامه علوم اطلاع‌رسانی*، ۱۶(۳و۴)، ۷۷-۶۹.

۱۰. رضازاده ملک، رحیم. (۱۳۸۰). *تبیین و تدوین قواعد املای فارسی*. تهران: نشر گلاب.

۱۱. سمایی، سید مهدی. (۱۳۷۹). مفرد و جمع در نمایه‌سازی. *فصلنامه علوم اطلاع‌رسانی*، ۱۶(۱و۲)، ۳۱-۲۷.

۱۲. فرهنگستان زبان و ادب فارسی. (۱۳۸۳). *دستور خط فارسی*. تهران: فرهنگستان زبان و ادب فارسی.

۱۳. کارنیرو، آلبرتو. (۱۳۸۳). نقش منابع هوشمند در مدیریت دانش (علیرضا گنجی، مترجم). *فصلنامه علوم اطلاع‌رسانی*، ۱۹(۳و۴)، ۷۷-۶۷.

۱۴. کمیجانی، احمد. (۱۳۸۲). ساختار نمایه‌سازی در موتورهای کاوش وب. *فصلنامه علوم اطلاع‌رسانی*، ۱۷(۳و۴)، ۴۸-۴۴.

۱۵. کوشا، کیوان. (۱۳۸۱). معیارهای ارزیابی ابزارهای کاوش اینترنت: مطالعهٔ مقایسه‌ای بر روی ابزارهای کاوش وب با واسط جستجوی فارسی. *مجله الکترونیکی کتابدار*، ۱(۱). دسترسی در اول بهمن، ۱۳۸۴، از سایت نشر کتابدار:
http://www.ketabdar.org/magazine/detailarticle.asp?number=25

۱۶. محققزاده، محمد صادق، و زارعیان، کاظم. (۱۳۸۳). ارائهٔ راه حل برای برخی مسائل اتوماسیون و نگارش فارسی. *فصلنامه علوم اطلاع‌رسانی*، ۱۹(۳ و۴)، ۸-۱.

۱۷. مرتضایی، لیلا. (۱۳۸۰). مسایل زبان و خط فارسی در ذخیره‌سازی و بازیابی اطلاعات. *فصلنامه علوم اطلاع‌رسانی*، ۱۷(۱و۲)، ۲۹-۲۴.

۱۸. هسی-یی، اینگرید. (۱۳۸۲). اینترنت: سازماندهی و جستجو (قاسم آزادی، مترجم). فصلنامه علوم اطلاع‌رسانی، ۱۸(۳و۴)، ۱۰۳-۹۴.

۱۹. محسنی، یاسمن. (۱۳۸۰). *ایجاد و نمایش آنتالوژی برای شبکهٔ مفاهیم مرتبط با حوزهٔ مخابرات نوری*. تهران: مرکز تحقیقات مخابرات ایران.

۲۰. عبداللهی، بهناز. (۱۳۸۰). *بررسی روش‌های طراحی و ایجاد آنتولوژی*. تهران: مرکز تحقیقات مخابرات ایران.

۲۱. جباری‌فر، معصومه. (۱۳۸۰). *بررسی پارامترهای ارزیابی و لیست دسته‌بندی شدهٔ جویشگرها*. تهران: مرکز تحقیقات مخابرات ایران.

۲۲. صالحی، مازیار، زارع بیدکی، علی محمد. (۱۳۸۰). *ارائه RFP برای یک جویشگر دوزبانهٔ فارسی/انگلیسی*. تهران: مرکز تحقیقات مخابرات ایران.

۲۳. میریان، مریم السادات. (۱۳۸۰). *ارائهٔ چارچوب کلی برای زیرسیستم‌های Information Retrieval and Query Processing*. تهران: مرکز تحقیقات مخابرات ایران.

پی‌نوشت‌ها

1. Web Content Mining
2. Web Structure Mining
3. Crawler
4. FASTUS:Finite-State Automaton Text Understanding System
5. Ontology
6. Semantic Web
7. Personalization
8. MLDB
9. NPiran
10. Iranhoo. www.iranhoo.com

11. IranMehre. www.iranmehr.com
12. Parseek. www.parseek.com
13. Google. www. Google.com/webhp?hl=fa
14. Loan Words
15. Slang Words
16. Neologisms
17. Recall
18. Precision
19. Query Processing
20. Cross language Retrieval

۱) کارشناس ارشد کامپیوتر- نرم‌افزار
پست‌الکترونیکی: mjshahidy@yahoo.com
۲) عضو هیئت علمی دانشگاه صنعتی اصفهان
۳)عضو هیئت علمی دانشگاه اصفهان