مركز اطلاعات و مدارك علمي ايران
مجله الکترونیکی
**شماره دوم دوره چهارم**

# روشي براي رفع چالش هاي محتواكاوي وب هاي فارسي زبان

**دكتر محسن صديقي**- عضو هيئت علمي دانشگاه صنعتي اصفهان
**دكتر كامران زماني فر**- عضو هيئت علمي دانشگاه اصفهان
**سيد مجتبي شهيدي**- دانشجوي كارشناسي ارشد دانشگاه آزاد نجف آباد

**چكيده**
زبان فارسي از دو منظر براي ما ايرانيان حائز اهميت است. اول آنكه اين زبان ، با تاريخ و فرهنگ و تمدن ما پيوندي ديرينه و ناگسستني دارد و دوم آنكه زبان فارسي ، زبان رسمي كشور و ابزار مبادله انديشه ها وايده ها در عرصه علمي و فرهنگي اين مرزوبوم محسوب مي گردد.رشد علمي و فني و فرهنگي ما در گرو برقراري ارتباط زباني و كلامي با دنياي الكترونيكي عرضه دانش و فرهنگ است كه وب نام دارد و اين ميسر نيست جز با تقويت كيفي زبان فارسي مورد استفاده در اين دهكده جهاني ، وب.ليكن زبان فارسي ، در تلاقي با جهان الكترونيكي ، بخصوص از بعد رسم الخط ، داراي مصائبي است كه كاوش در محتويات آن را دچار كم كيفيتي مي نمايد. اين مقوله مستلزم تمهيداتي چند است تا زبان فارسي را از يك زبان شعر وعرفان به زباني مناسب با پهنه الكترونيكي دادوستد دانش ، وب ، تبديل نمايند. مقاله حاضر ، تلاشي است درجهت مرتفع سازي چالش هاي كاوش در وب هاي فارسي زبان كه از دو ديدگاه رسم الخط ، با استفاده از نمايه سازي فارسي و مفهومي ، با استفاده از آنتولوژي قابل بحث هستند.
**واژه هاي كليدي** : آنتولوژي ، نمايه سازي فارسي ، كاوش وب هاي فارسي ، وب كاوي

## مقدمه

اهميتي كه پديده وب فارسي بعنوان رسانه‌اي مستقل و مؤثر در دنياي ارتباطات ايرانيان پيدا كرده است ، غيرقابل‌انكار است. به‌نظر مي‌رسد كه اكنون روآوردن برخي از روزنامه‌نگاران ، پژوهشگران ، دانشجويان ،... به وب فارسي و استفاده منابع خبري ، علمي ،... از مطالب آنها نيز، موجب تقويت نقش رسانه‌اي وب فارسي شده‌است.
ليكن با توجه به ماهيت خاص رسم الخط فارسي كه آن را براي سيستم هاي رايانه اي نامناسب نموده است ، امروزه مشكلات بسياري براي دانش پژوهشان و بطور كلي استفاده كنندگان از وب هاي فارسي زبان نموده است. عدم وجود حروف صدادار در فارسي بصورت يك موجوديت مجزا از يك طرف و وجود حروف يكسان با اشكال متعدد از طرف ديگر ، باعث بروز چالش هاي جدي در امر نمايه سازي اين زبان شده است. بنظر مي رسد تلاشهايي لازم است تا زبان زيباي فارسي را با وجود ماهيت عرفاني و شاعرانه آن ، جهت حضور در عرصه الكترونيكي دانش ، آماده نمايد.

## پيشينه تحقيق و تعارف ابتدايي

محتواكاوي وب(Web Content Mining) ، يكي از سه شاخه وب كاوي است كه در واقع ، كشف اطلاعات مفيد از مستندات و داده هاي ساختيافته و نيمه ساختيافته و غير ساختيافته وب مي باشد. يك شاخه ديگر اين مقوله ، ساختاركاوي وب(Web Structure Mining) است كه به كشف مدل پشت زمينه حاكم بر ساختار فرا پيوند هاي وب مي پردازد و هدف آن ، ايجاد اطلاعاتي همچون تشابه يا ارتباط بين سايت هاي مختلف وب است. شاخه ديگر آن كاربرد كاوي وب مي باشد كه سعي مي كند از تعاملات كاربربا وب ، اطلاعاتي كسب كند و از آن ها بصورت سابقه اي در مراجعات بعدي كاربر سود ببرد.
در زمينه محتواكاوي وب نرم افزارهاي خزنده(Crawler) ، به گشت و گذار در اقيانوس وب پرداخته ، اقدام به نمايه سازي واژگان در پايگاه داده خود مي نمايند كه مورد استفاده موتورهاي كاوش ، در زمان جستجوهاي كاربر قرار مي گيرد. نمونه بارز اين روش ، موتور كاوشگر Google است. [Chakrabarti,1999]
در همين راستا ابزارهايي همچون FASTUS:Finite-State Automaton Text Understanding System ، در خلال اين ماموريت به تجزيه و تحليل متون ، با هدف كشف گروه هاي مختلف واژگان مانند اسامي ، افعال ، تركيبات وصفي و اضافي ،... مي پردازند كه به كشف دانش از محتويات وب كمك مي كند. اين روش هم اكنون براي زبان هاي انگليسي و ژاپني پياده سازي شده است وبصورت بالقوه براي ديگر زبان ها قابل استفاده است. [Feiyu,2001]
از طرف ديگر استفاده از آنتولوژي(Ontology) در وب در بهينه سازي كاوش در وب پيشنهاد مي گردد. آنتولوژي ، يك

فرهنگ واژگان مشترك بر اساس موضوع سايت براي استاندارد سازي ارائه مفاهيم آن جهت قابل تفسير شدن توسط ماشين ، تعريف مي كند. آنتولوژي ، يك جزء كليدي وب مفهومي(Semantic Web) است. [Heflin,2000]
شخصي كردن وب(Personalization) ، از ديگر روش هاست كه در امر كاوش وب مثمر ثمر است. نمونه اين روش در My Yahoo قابل مشاهده است.
يكي ديگر از راه هاي كاوش در مقدار زياد و غير ساختيافته اطلاعات وب ، استفاده از پايگاه داده چند لايه اي (MLDB) است. هر لايه از اين پايگاه داده ، تعميم بيشتري از لايه قبلي است. همه لايه ها بجز پايين ترين لايه (كه خود وب است) ، قابل كاوش توسط يك زبان پرس وجو مثل SQL است. [Osmar,2002]
در پياده سازي روش هاي ساختاركاوي وب ، از تئوري گراف وب بهره مند خواهيم شد كه به ايجاد ديد ارزشمند در الگوريتم هاي جستجو ، كشف ارتباطات ،... موثر است.
در خصوص روش هاي كاربرد كاوي وب ، ناوبري كاربر در وب توسط مدل هاي رياضي ماركو(Markov) ، براساس ميزان تجربه كاربر و دارا بودن يا عدم داشتن راهنماي سايت ، تجزيه و تحليل مي گردد. [Velasquez,2003]

**خصوصيات وب هاي فارسي از نظر زبان**

عدم وجود يك استاندارد و شناور بودن ويژگيهاي رسم الخط و مفاهيم در زبان فارسي ، موجب گرديده است تا تقريبا بتعداد صفحات وب فارسي ، سبك و سياق نگارش اين زبان بكار رفته باشد. ليكن خصوصيات مشترك اكثر وب‌هاي فارسي زبان را مي توان چنين ارزيابي نمود :
الف) نگارش برخي از وب هاي فارسي ، زبان غيررسمي يا محاوره‌اي است.
ب) در وب‌هاي فارسي ، بخصوص در متون علمي ، اغلب واژه‌هاي بيگانه ، بكرات استفاده مي‌شود كه بعضي از آنها بارسم الخط زبان اصلي نوشته مي‌شوند.
ج) رسم‌الخط وب هاي فارسي ، اصولا غيراستاندارد و متغير است و اغلب در معرض نوآوري است.
ه) نوشته‌هاي وب‌هاي فارسي ، حاوي غلطهاي تايپي و نگارشي نسبتاً زيادي است، هرچند كه اغلب وب‌هاي فارسي مهم و پرخواننده، نگارش قابل‌قبولي دارند.
و) رسم‌الخط وب‌هاي فارسي، تابع محدوديت‌هاي محيط الكترونيكي و عدم تطبيق آن با الزامات خط فارسي است.] اشرف زاده،١٣٨٣[ابزارهاي جستجو در وب هاي فارسي

در حال حاضر ابزارهاي كاوش مختلفي در ايران ظهور پيدا كرده اند.ليكن ابزارهاي جستجويي كه امكان جستجوي اطلاعات به زبان فارسي را در اختيار قرار مي دهند ، محدودند. از طرف ديگر ، امكانات و قابليتهاي آن ها براي بازيابي موثر و مناسب اطلاعات متغير هستند. برخي از ابزارهاي كاوش با امكانات جستجوي فارسي عبارتند از ، Iranhoo ، NPiran Google ، Parseek ، IranMehre و Parseek .بجز سايت NPIran ، ديگر سايتها داراي واسط جستجوي فارسي هستند و بجز Parseek ، هيچيك از ابزارهاي موجود كاوش فارسي ، چالش هاي زبان فارسي را با هدف بهينه سازي كاوش فارسي ، فراروي خود قرار نداده اند و Parseek نيز تنها مشكل كاراكترهاي فارسي با يونيكدهاي مختلف را حل نموده است.

| رديف | نام ابزار كاوش با واسط جستجوي فارسي | نشاني دسترسي |
|---|---|---|
| ١ | Google | www.google.com/webhp?hl=fa |
| ٢ | IranClick | www.iranclick.com |
| ٣ | Iranhoo | www.iranhoo.com |
| ۴ | IranMehr | www.iranmehr.com |
| ۵ | Open Directory | http://dmoz.org/World/Farsi |
| ۶ | Parseek | www.parseek.com |

جدول (١) شش ابزار كاوش برگزيده با واسط جستجوي فارسي همراه با نشاني آنها.

در بين ابزارهاي كاوش فوق ، تنها موتور كاوش Google داراي برنامه روبات به منظور شناسايي و نمايه سازي صفحات يا سايتهاي وب به زبان فارسي و نمايه سازي خودكار مي باشد و قادر است صفحات فارسي را در قالب Unicode شناسايي و در پايگاه خود نمايه كند و سايت Parseek نيز از پايگاه Google براي جستجو و بازيابي اطلاعات استفاده مي كند. به تعبير ديگر، ٤ ابزار كاوش ديگر توسط نمايه سازي انساني اداره مي شوند و از اين لحاظ راهنماي موضوعي تلقي مي شوند. ابزارهاي كاوش ديگر ، راهنماي موضوعي به شمار مي آيند و انسان ، فرايند شناسايي، بررسي و نمايه سازي سايتها يا صفحات وب را بر عهده دارد.]كوشا،١٣٨١[

بنظر مي رسد جاي يك ابزار كاوش قوي ملي ، تحت نظارت سازمان هاي انفورماتيكي و انجن هاي زبان شناسي فارسي ، منطبق با نيازهاي اطلاعاتي كاربران اينترنت در ايران وبا در نظر گرفتن چالش هاي رسم الخط و مفهومي فارسي و مرتفع سازي مشكلات ناشي از آن ها خالي است.

**مشكلات ومحدوديت هاي وب كاوي در سايت هاي فارسي زبان**

در دهه هاي اخير ، بيشترين اختلاف نظر در باب شيوه املاي كلمات فارسي بر سر موضوع جدانويسي يا پيوسته نويسي كلمات مركب بوده است.فرهنگستان زبان و ادب فارسي ، در اين باب راه ميانه را برگزيده و كوشيده است تا فقط مواردي را كه جدانوشتن و يا پيوسته نوشتن آنها الزامي است ، تحت قاعده و ضابطه درآورد و شيوه نگارش بقيه كلمات مركب را به ذوق و سليقه نويسندگان واگذار كند.]فرهنگستان،١٣٨٢[

بعضي چالش هاي زبان فارسي در رايانه و بخصوص در اينترنت كه باعث تفاوت در نتيجه جستجو در وب يا وب كاوي مي شود از قرار زير است :

الف) تنوع نحوه استفاده از "مي" چسبان و غير چسبان ، مثل كلمات "مي تواند" و "ميتواند".

ب) تنوع نحوه بكاربردن چسبان و غير چسبان "ها" ، مثل "آن ها" و "آنها".

ج) بكار بردن بعضي پيشوند ها و پسوند ها ، مثل "همين كه" و "همينكه" ويا "هيچ يك" و "هيچيك" و يا "راه گشا" و "راهگشا".

د) بكاربردن "حمزه" بصورت هاي مختلف ، مثل "مسؤول" و "مسئول" يا "مسأله" و "مسئله".

ه) استفاده يا عدم استفاده از "ء" ، براي كلمات مختوم به هاي بيان حركت ، در حالت مضاف ، مثل "خانة مسكوني" و "خانه مسكوني".

و) تنوع استفاده از "ي" در كلمات عربي مختوم به "ا" ، مثل "موسي" و "موسا".

ز) تنوع املايي بعضي كلمات كه همه درست هستند ، مثل "اتاق" و "اطاق".

ح) استفاده از كلمات اروپايي بصورت زبان اصلي يا ترجمه فارسي بخصوص در متون علمي ، مثل "Update" و "بروزآوري".

ط) استفاده يا عدم استفاده از جمع مكسر براي بعضي كلمات.

ي) تبديل كلمات اروپايي به رسم الخط فارسي با همان تلفظ اصلي ، مثل "Source" و "سورس".

ك) استفاده از "ا" و "آ" بجاي هم ، مثل "فرايند" و "فرآيند".

ل) استفاده يا عدم استفاده از اعراب براي كلمات .

بعبارت ديگر ، يك كاربر ممكن است در جستجوي خود در وب ، كلمه كليدي خاصي را بكار برد ، ليكن در صفحات وب چنين كلمه اي بكار نرفته باشد و با توجه به مواردي كه در مورد تنوع كاربري كلمات ، بحث شد ،كلمه مشابهي ثبت شده باشد. بنابراين بسياري از صفحات وب مطلوب كاربر ، در مجموعه بازيابي شده ، وجود نداشته باشد.

**روش هايي براي بهبود كاوش وب هاي فارسي**

الف) انتخاب مناسب سرعنوان هاي موضوعي در وب هاي فارسي

پيدا كردن اصول و معيارهاي موضوع سازي ذهني و فرايندي كه در ذهن كاوشگران اطلاعات در هنگام بيان موضوعات ، براي پاسخ يابي ماشيني ، روي مي دهد يك فرايند پيچيده ، مهم و تاثيرگذار در جريان تهيه سرعنوان هاي موضوعي است. از طرفي تركيب بندي عبارات كاوش با يك زبان مشترك بين انسان و ماشين ، از جمله مسايلي است كه هميشه متخصصان بانك هاي اطلاعاتي و كاوشگران اطلاعات را دچار مشكل مي سازد. بهمين دليل و با توجه به ساختار بانك هاي اطلاعاتي ، حوزه موضوعي كاوش ، ميزان آگاهي هاي عمومي كاوشگر ، زبان رايج تخصصي ميان ورزيدگان يك رشته خاص موضوعي ، مسائل و مشكلات زباني ، ساختار اصطلاحنامه بكار گرفته شده در بانك اطلاعاتي و ... است كه راهبردهاي كاوش ، طراحي و اجرا مي شوند. در اين مسير ، سرعنوان هاي موضوعي ، نقش عمده اي را دارا هستند. حل اين مسائل مي تواند به پيدا كردن راه حل هاي موثري براي سرعنوان هاي موضوعي بيانجامد.]بيگلو،١٣٨٢[

ب) استمداد از علم اصطلاح شناسي(Terminology) در نمايه سازي ماشيني

توجه به اصطلاحات و اصطلاح سازي نيز با توجه به ضرورت روزآمد بودن واژگان علمي و تخصصي و لزوم كنترل ورود اصطلاحات بيگانه امري است كه ما را ناگزير به استمداد از علم اصطلاح شناسي وامي دارد.در اين خصوص "حسيني" پژوهشي ارائه كرده است كه بجهت اشاره به تمهيدات وي در خصوص تشكيل ويا بهينه سازي اصطلاحنامه اي مناسب براي نمايه سازي ماشيني ، شمه اي از آن در ادامه مشروح مي باشد :

الف) كنترل مترادف ها و شبه مترادف ها بصورت ارجاع مترادف هاي غير مرجح به اصطلاح مرجح.

ب) هدايت كاوشگر از مفاهيم و اصطلاحات اخص به اعم يعني نزديك ترين اصطلاح.

ج) با ارائه روابط ساختاري مفاهيم اعم از سلسله مراتبي يا غير سلسله مراتبي ، جامعيت حاصل مي گردد و كاوش را با ارائه طبقه هاي داراي ارتباط بسيار نزديك توسعه مي دهد. از اين طريق مانعيت نيز با پيشنهاد اصطلاحات اخص ، بهبود مي يابد.
د) نظارت بر شكل دستوري ، املايي ، جمع و مفرد و اختصارات و شكل مركب اصطلاح.
ه) گزينش بين دو يا چند مترادف موجود براي بيان يك مفهوم.
و) تصميم گيري در خصوص پذيرش و نحوه برخورد با انواع خاصي از اصطلاحات نظير "واژه هاي قرضي"(Loan Words) ، "واژه هاي عاميانه"(Slang Words) ، اسامي تجاري و اسامي خاص.
ز) محدود كردن معني يك اصطلاح كه در يك فرهنگ ممكن است با توضيحات گوناگون همراه باشد.
توصيه هاي اضافي در خصوص تشكيل اصطلاحنامه بشرح زير است :
الف) واژه هاي قرضي :
واژه هايي كه از زبانهاي ديگر قرض گرفته شده اند و در زبان قرض گيرنده تثبيت شده اند. چنانكه ترجمه اين اصطلاحات وجود داشته باشد ولي بطور رايج مورد استفاده قرار نگيرد با صاطلاح ترجمه شده بايد بصورت اصطلاح نامرجح برخورد كرد.
ب) نو واژه ها(Neologisms) ، اصطلاحات عاميانه و زبان حرفه اي :
چنانچه جايگزيني كه بطور گسترده توسط كاربران مورد استفاده قرار گيرد ، وجود نداشته باشد ، نو واژه ، اصطلاح عاميانه يا حرفه اي ، بعنوان توصيفگر پذيرفته مي شوند.
ج) اسامي عاميانه و اسامي تجاري :
توصيه مي شود ، در جايي كه اسم عاميانه معادلي وجود دارد ، بايد از آن بجاي اسم تجاري استفاده كرد.
د) اسامي مشهور و اسامي علمي :
انتخاب بين ايندو بر اساس احتمال بيشتر استفاده كاربران مي باشد.
ه) اسامي مكان ها :
در جايي كه براي يك كشور يا منطقه جغرافيايي درون يك جامعه تك زباني ، بيش از يك اسم ، انتخاب مي گردد ، بايد اسمي را بعنوان اصطلاح مرجح تعيين كرد كه نزد كاربران ، آشناتر است.
و) اسامي خاص موسسات ، افراد و...
ميزان نياز دستيابي به اسامي خاص بر اساس حوزه عملكرد اصطلاحنامه ، گنجاندن اسامي را در اصطلاحنامه اصلي تعيين مي كند.
ز) همنام ها و هم آوا ها :
منظور ، كلماتي هستند كه داراي املاء يكسان و معاني متفاوت يا داراي آواي يكسان و معاني متفاوت مي باشند. در چنين مواردي روش معمول ابهام زدايي ، اضافه كردن توضيحگر است كه داخل پرانتز قرار مي گيرد.
ح) مترادف ها :
انتخاب مترادف ها بايد بر اساس نياز هاي كاربران باشد كه از نقطه نظر رواج و تخصص ، صورت مي گيرد.
ط) شبه مترادف ها :
پذيرش شبه مترادف ها ، از حوزه موضوعي زير پوشش اصطلاحنامه ، متاثر است. براي مثال "افراد با استعداد" و "تيزهوشان". شبه مترادف ها ممكن است شامل متضاد ها هم باشند مثل "سوادآموزي" و "بيسوادي".]حسيني،١٣٨٣[

ج) تعريف يك استاندارد براي مفاهيم و رسم الخط فارسي در وب
همانطور كه گفته شد ، يك تفاوت زبان فارسي با زبان انگليسي (و زبان هاي هم ارز) ، تنوع املايي يا رسم الخطي كلمات آن است. بعبارت ديگر ، در زبان انگليسي ، تنوع در مفهوم كلمات وجود دارد. يعني براي بعضي مفاهيم ، ممكن است كلمات متنوعي استفاده شود. براي مثال كلمات "Hello" و "Hi" كه داراي مفهوم يكساني هستند. ليكن در فارسي ، علاوه بر وجود كلمات متنوع براي مفاهيم يكسان ، مثل "كامپيوتر" و "رايانه" ، تنوع در رسم الخط يك كلمه نيز فراوان بچشم مي خورد. بعبارت ديگر ، در حالي كه شما بدنبال صفحات محتوي كلمه "امپراتور" مي گرديد ، كليه صفحات محتوي كلمه "امپراطور" را از دست مي دهيد.
بنظر مي رسد ، در تشكيل صفحات وب فارسي ، جاي يك استاندارد حاكم بر عملكرد تاليف نويسندگان وب ، خالي است. استانداردي كه انتخاب بعضي كلمات داراي چندين رسم الخط و حتي انتخاب بعضي كلمات كه بر مفاهيم متنوعي دلالت دارند را منحصر بفرد نمايد و مولفان را از طرفي ترغيب به انتخاب گونه زباني مناسب، براي تضمين كيفيت ارتباط و انتقال مؤثر پيام و از طرف ديگر موظف به حفظ سلامت زبان و رعايت استانداردهاي آن به‌عنوان يك وظيفه رسانه‌اي نمايد.
ايجاد و گسترش چنين استانداردي بعهده "فرهنگستان زبان و ادب فارسي" و با هماهنگي انجمن ها و شوراهاي علمي يا صنفي انفورماتيك در ايران است. تعويق در تنظيم اين استاندارد ، با توجه به رشد روز افزون وب هاي فارسي زبان ،

هزينه هاي جبران ناپذيري در بر خواهد داشت.

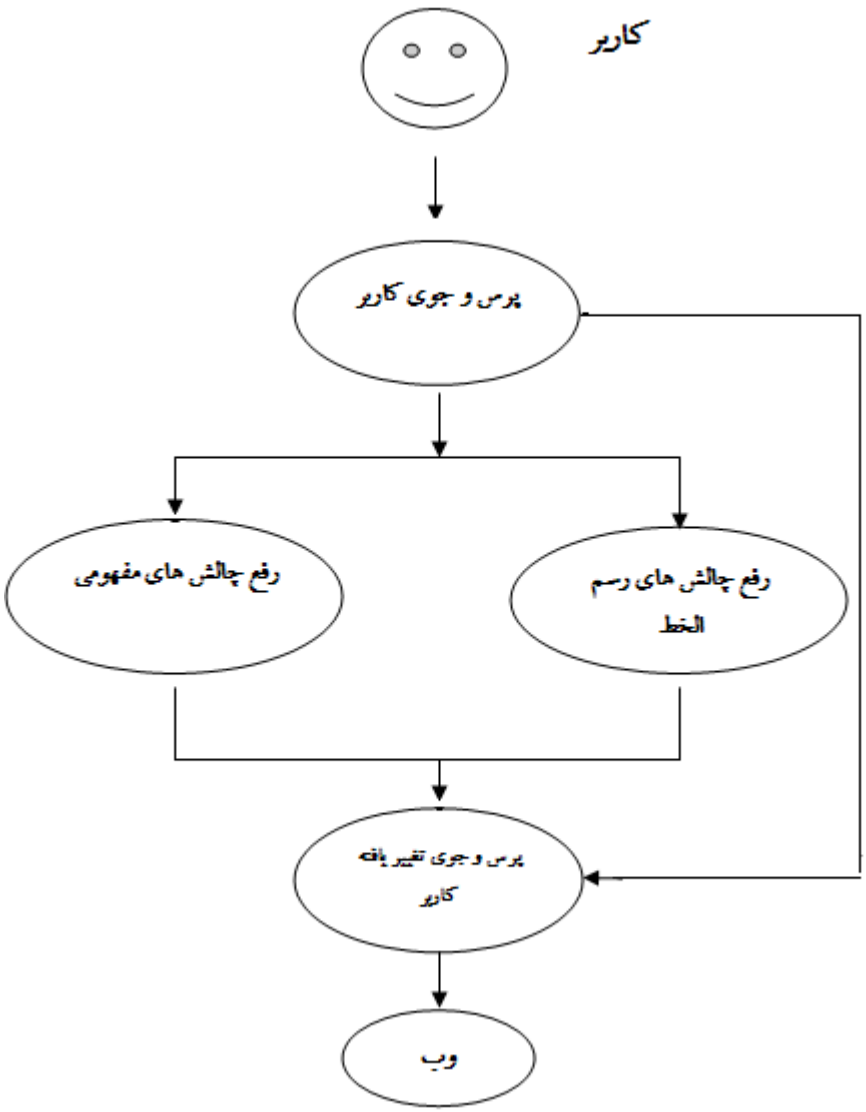
د) استفاده از مفرد و جمع در نمايه سازي

استفاده از اسامي جنس ، نحوه جمع بستن كلمات بصورت باقاعده يا بدون قاعده (جمع هاي مكسر) معضلي است كه در نمايه سازي واژگان فارسي معمولا بسياري از صفحات وب را شامل نمي شود. در اين خصوص "سمايي" در مقاله خود قواعدي را براي نمايه سازي واژه هاي مفرد و جمع ارائه داده است كه ذكر آن ها خالي از لطف نيست :
الف) از آنجا كه كليدواژه‌ها در زبان تخصصي بكار مي‌روند و در بين اهل فن رايج و جاري‌اند، گاه اتفاق مي‌افتد كه صورت جمع مرسوم باشد. در اين حالت بهتر است كه از صورت جمع استفاده شود. نظير تركيب "آثار باستاني". نكته‌اي كه در اين باره ذكركردني است، شيوة جمع بستن اسامي در اين موارد است. بدين معنا كه گاهي نوع پسوند جمع يا شيوة جمع بستن باعث مي‌شود كه اصطلاح به دست آمده، با سنت رايج در حوزه تخصصي منطبق نباشد. مثلاً چنانچه لفظ اثر با "ها" جمع بسته و تركيب "اثرهاي باستاني" ساخته شود، با اصطلاح رايج اهل فن متفاوت خواهد بود.
بعلاوه اين كه در برخي موارد، شيوة جمع بستن باعث تفاوت در معنا مي‌شود. "اثرها" در برخي بافت‌ها معادل "آثار" نيست. "آثار" در تركيب با "باستاني"» شامل خرابه‌ها و بناها و اشياي به جا مانده از زمان قديم مي‌شود، در حالي كه از لفظ "اثرها" بيش‌تر معناي ردّ و نشان تداعي مي‌شود.
ب) گاهي صورت مفرد كلمه، معنايي متفاوت از معناي جمع دارد. اين مسئله اغلب در كلمات عربي مصطلح در فارسي وجود دارد. تركيب "مصالح راه‌سازي" از اين دست است. "مصالح" بمعناي مواد لازم براي ساختن بنا است، در حالي كه معناي صورت مفرد آن يعني "مصلحت" ـ به نقل از فرهنگ معين ـ از اين قرار است: "آنچه كه صلاح و سود شخص يا گروهي در آن باشد".
ج) در مواردي ، با اين كه صورت مفرد و جمع كلمه، معنايي مشترك دارند استعمال صورت مفرد در زبان رايج نيست. به همين علت تركيب‌هايي نظير "منسوجات نظامي" "الياف كربني" را نمي‌توان بشكل مفرد آورد و به جاي منسوجات لفظ "منسوج" و به جاي الياف لفظ "ليف" را قرار داد.
د) در برخي واژه‌ها، صورت جمع توسّع معني پيدا كرده و از اين طريق، ارتباط صورت جمع و مفرد ضعيف شده است. در واژه‌اي نظير "مهمّات" اين اتفاق رخ داده و ارتباط "مهمات" با "مهم" از اين دست است.
ه) بعضي تركيب‌ها نظير "ماشين‌آلات" وجود دارند كه نه تنها نمي‌توان قسمت جمع آنها را به شكل مفرد آورد ، بلكه در مجموع، يك واحد نحوي ايجاد مي‌كنند كه از لحاظ معنايي تجزيه‌ناپذير است.
و) گاهي هم اتفاق مي‌افتد كه جمع اسم با قاعده فارسي، در زبان مصطلح نيست و جمع عربي آن رايج است. بديهي است كه در اين حالت چنانچه اسم مذكور، كليدواژه شود يا در تركيبي به كار رود و نتوان از صورت مفرد آن استفاده كرد، بايد شكل جمع عربي آن را به كار برد. در تركيب "اجزاي پل" نمي‌توان به جاي "اجزا" لفظ "جزءها" را به كار برد.
ز) براي جمع بستن اسامي لاتين استفاده از پسوند "ها" مرجح است: "كربامات‌ها".]سمايي،١٣٨٢[

ه) استفاده‌ازيك واسطكاوش فارسي براي رفع چالش‌هاي رسم‌الخط و مفهومي

استاندارد سازي رسم الخط فارسي در رايانه ها ، ممكن است در ابتداي توليد اولين صفحات وب فارسي بسيار مفيدتر بنظر رسد ، ليكن در حال حاضر ، با وجود تعداد بسيار زيادي صفحات وب فارسي كه در هر حال با عدم نظارت يك استاندارد ، توليد شده اند ، چندان مثمر ثمر واقع نمي شود. اگر چه ايجاد آن براي توليد صفحات وب فارسي آتي ، لازم است. بعبارت ديگر ، جهت انجام عمليات وب كاوي در صفحات وب فارسي كنوني ، بايستي روشي ابداع كرد تا با توجه به چالش هاي بحث شده ، نتايج مطلوبي از وب كاوي در آنها بدست آيد.
با توجه به بحث هاي قبل ، مي توان دريافت كه در كاوش هاي وب ، پارامترهايي كه نتايج جستجو را براي كاربر مطلوب جلوه مي دهد ، از قرار زير است :
الف) جامعيت نتايج(Recall) :
منظور از جامعيت نتايج ، اينست كه كليه صفحات وبي كه بر اساس كلمه كليدي ، مطلوب كاربر محسوب مي گردد ، نمايش داده شود. بعبارت هيچ صفحه مطلوبي از قلم نيافتد.
ب) مانعيت نتايج :منظور از مانعيت نتايج اينست كه صفحات وبي اضافه بر نتايج جستجوي مطلوب كاربر ارائه نشود ، كه بعلت حجم زياد نتايج ، باعث سردرگمي كاربر گردد.
ج) تناسب نتايج(Precision) :ميزان مطلوب بودن نتايج نسبت به مورد جستجو است كه بايد حداكثر باشد.
د)سرعت بازيابي:نكته مهم ديگر در كاوش ، زمان صرف شده براي جستجو است كه بايستي حداقل باشد. اين پارامتر ، به ميزان ترافيك شبكه ، سرعت خدمت دهنده ها ، سرعت پايگاه داده نمايه شده و موارد سخت افزاري وابسته است.

ليكن جامعيت ، مانعيت و تناسب نتايج ، مي توانند تحت تاثير زبان استفاده شده در نگارش محتواي صفحات ، تغيير نمايند. بخصوص موقعي كه زبان مورد استفاده ، زباني همچون فارسي با چالش هاي رسم الخط فراوان در گستره امور رايانه اي است و بشدت مستعد نتايج بي اعتبار و نامناسب مي باشد. بعنوان مثال كاربر جوينده اطلاعات در مورد "امپراتوري هاي قديم" از ديدن صفحات وب حاوي كلمه "امپراطور" در نتايج جستجوي خود ، محروم است.
لذا در خصوص ارتقاء كيفيت نتايج كاوش در وب هاي فارسي زبان ، جاي راهكارهايي كه پارامترهاي مذكور را تقويت نمايد خاليست. از اين رو ، بر آن شديم تا با ايجاد يك عامل هوشمند ، نتايج جستجوها را بهينه كنيم. اينكار را با اضافه كردن يك واسط هوشمند به موتورهاي كاوش يا خزنده ها انجام داديم. اين واسط در واقع نقش يك پردازشگر پرس و جو(Query Processing) را ايفاء مي كند.
اين عامل از دو قسمت تشكيل شده است. يك قسمت به مرتفع سازي معضلات رسم الخط و بهبود بعد جامعيت نتايج كاوش و قسمت ديگر به رفع مشكلات مفهومي و بهينه سازي تناسب و مانعيت نتايج كاوش مي پردازد.




شكل (١) شماي كلي واسط هوشمند فارسي.

**واسط فارسي براي بهبود جامعيت كاوش**

اين قسمت از عامل ، از يك پايگاه داده تشكيل شده است كه حاوي چندين كلمه معادل بعضي كلمات خاص كه در چالش ها ذكر گرديد ، مي باشد. اين تناظر ، مي تواند مربوط به معادل هاي رسم الخط ، معادل هاي مفهومي يا معادل هايي به زبان هاي غير فارسي باشد. بدين صورت كه با عبور دادن كلمات مورد كاوش از اين واسط و يا با رجوع به اين پايگاه داده ، عملا يك كاوش بر اساس يك كلمه كليدي خاص ، منجر به چند كاوش براي كلمات معادل آن كلمه كليدي خاص مي گردد. با اين ترفند ، صفحات حاوي كلمات معادل ، از دست نمي رود و پارامتر جامعيت را تقويت مي نمايد.

**واسط فارسي براي بهبود مانعيت وتناسب كاوش**

همانطور كه در فصل قبل تشريح گرديد ، آنتولوژي مي تواند ، با ايجاد ساختار مفهومي مناسب ، بر حسب موضوع اصلي سايت وب (مثل وب خدماتي ، توليدي ، علمي ،...) كاوش ماشيني را تسهيل كند. چرا كه با روشن بودن نوع وب از نظر موضوع وساختار پيوندي آن (شكل گراف وب) با توجه به آنتولوژي استانداردي كه براي كاوشگر نيز شناخته شده است ، بطور بهينه اي مي توان عمليات كاوش را بثمر رسانيد. بعبارت ديگر اگر كاربري در مورد "پايان نامه هاي تحصيلي" جستجو كند ، موتور كاوشگر ، با اطلاع از نوع سايت هاي دانشگاهي ، دقيقا با مراجعه به پيوند هاي مربوطه (مثل "پژوهش ها" ، "پايان نامه ها" ،...) در سايت هاي مربوطه ، نتايج بهينه اي را در اندك زماني فراهم نمايد.

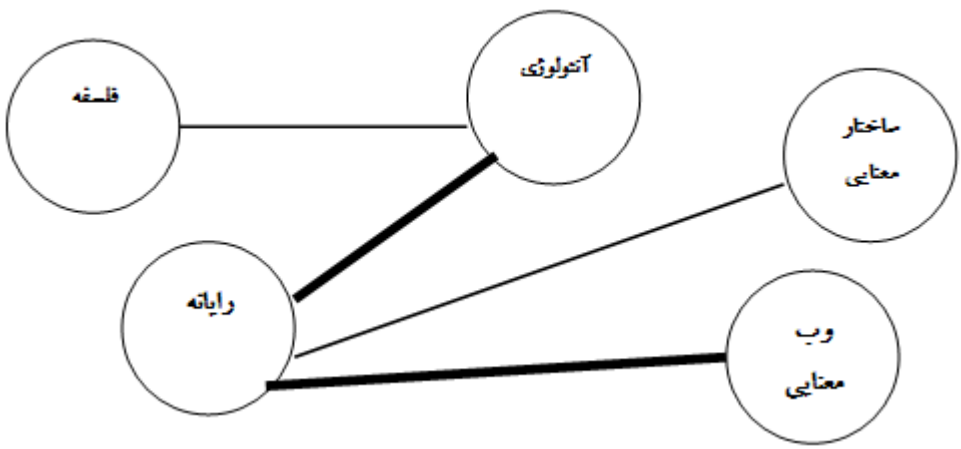
با تعريف يك گراف بدون جهت مفهومي (گراف آنتولوژي) محتوي واژه هاي فارسي و ارتباط بين آنها مي توان گامي در جهت بهبود نتايج كاوش از نظر مانعيت ، برداشت. گراف عمومي ما شامل اجزاء زير است :

الف) گره ها كه واژه متناظر با آن ها ، دلالت بر يك موجوديت فيزيكي يا مفهومي دارد. مثل ايران ، ميز ، كشور ، حيوان ، درخت ، اسب ،... هريك از اين گره ها ، با يك كد مشخصه گره ، يك واژه گوياي موجوديت آن و يك عدد نشان دهنده فاصله آن گره تا ريشه است.

ب ) پيوند بين گره ها : شامل پيوند هاي بين آنهاست كه هر يك معرف رابطه مفهومي بين واژه هاست. هر پيوند از نظر ساختار ، حاوي يك كد مشخصه پيوند ، كد مشخصه گره شروع و كد مشخصه گره پايان مي باشد.

زماني كه كاربر ، كاوش خود را در مورد كلمه خاصي آغاز مي كند ، معمولا درابتدا نتايج بسيار زيادي دريافت مي كند كه بعلت عدم اعمال محدوديت روي كلمه مورد كاوش وي مي باشد. واسط هوشمند ، در مراحل بعدي كاوش ، با محدود كردن دامنه جستجو با توجه به كلمات مورد نظر كاربر و بر اساس آنتولوژي موجود ، سعي در فيلتر كردن نتايج حاصل از كاوش دارد. واسط مفهومي ما با در نظر گرفتن كلمات استفاده شده در كاوش هاي قبلي كاربر ، اقدام به يافتن مسيري در گراف عمومي مذكور ، بين واژه هاي بكار رفته مي نمايد. اگر در گراف مذكور ، مسيري با طول قابل قبول يافت شد ، واسط ، گره هاي موجود در بين اين مسير را شناسايي نموده ، گره با كمترين فاصله تا ريشه را ، مبنايي براي فيلتر كردن سايت‌هاي وب كاوش شده قرار مي‌دهد واز ارائه سايت‌هاي ديگر به كاربر ،كه موضوع آنها با واژه گره اصلي همخواني ندارد ، اجتناب مي ورزد و بدين طريق به مانعيت و تناسب نتايج حاصل از كاوش بهبود مي بخشد.

براي مثال اگر كاربر كاوش خود را با واژه "آنتولوژي" شروع نموده و در مراحل بعدي جستجوي خود ، واژه هاي "ساختار معنايي" ، "وب معنايي" و... را مورد كاوش قرار دهد ، واسط فارسي اقدام به يافتن مسيري از واژه "آنتولوژي" تا واژه "وب معنايي" مي نمايد و در فاصله اين مسير به گره اصلي "رايانه" مي رسد. از اين پس ، واسط ، نتايج حاصل از كاوش كاربر را به صفحاتي با موضوع اصلي "رايانه" محدود مي كند و از ارائه صفحاتي با موضوع اصلي "فلسفه" خودداري مي ورزد.




شكل (٢) قسمتي از گراف معنايي وب(آنتولوژي)

| شناسه | واژه گره در گراف | فاصله از ريشه |
|---|---|---|
| ١ | آنتولوژي | ١ |
| ٢ | وب معنايي | ١ |
| ٣ | رايانه | ٠ |
| ۴ | فلسفه | ٠ |

جدول (٢) ساختار و محتويات نمونه جدول گره هاي گراف معنايي در پايگاه داده.

| شناسه | گره شروع | گره پايان |
|---|---|---|
| ١ | وب معنايي | رايانه |
| | | |

| | | |
|---|---|---|
| ٢ | آنتولوژي | رايانه |
| ٣ | فلسفه | آنتولوژي |

جدول (٣) ساختار و محتويات نمونه جدول پيوندهاي گراف معنايي در پايگاه داده.

**معماري**

اين عامل هوشمند ، در خصوص هر يك از چالش هاي رسم الخط زبان فارسي رايانه اي ، رفتار متفاوتي از خود نشان مي دهد. اين رفتارها بقرار زير است :

الف) تنوع نحوه استفاده از "مي" ، "ها" ، پيشوند ها و پسوند ها :

همانطور كه قبلا توضيح داده شد ، موارد فوق بطور چسبيده يا جدا از كلمه بكار برده مي شود. لذا جهت رفع چنين مشكلي ، مي توان در واسط هوشمند ، با حذف كليه فواصل خالي(Blanks) موجود در عبارت مورد كاوش ، اقدام به جستجو بر اساس دنباله اي از حروف همان عبارت ، بدون هيچگونه فاصله خالي نمود.

ب) بكاربردن "حمزه" بصورت هاي مختلف :

جهت حل مشكل فوق ، در عمل هوشمند مورد بحث ، فرآيندي ايجاد مي گردد ، كه طي آن ، اگر عبارت مورد كاوش حاوي صور مختلف "حمزه" باشد ، عملا كاوش ، به چندين جستجو براي كلمات مشابه ، با حالت هاي مختلف "حمزه" تبديل مي شود. بعبارت ديگر كاوش كلمه "مسئله " به كاوش براي كلمات "مسئله" و "مسأله" منجر مي شود. مي توان با جايگزيني "ي" بجاي "ء" نيز دامنه كاوش را وسيع تر نمود ، مثل "رئيس" و "رييس".

ج) استفاده يا عدم استفاده از "ء" در تركيب هاي اضافي يا وصفي :

جهت رفع اين مشكل ، در صورت استفاده كاربر از "ء" در عبارت مورد كاوش خود ، واسط هوشمند اقدام به جستجو براي عبارتي فاقد "ء" مي نمايد. در اين صورت نتايج جستجو ، صفحاتي را كه در محتواي متن آنها از "ء" استفاده نشده است نيز شامل مي گردد.

د) استفاده از "ا" و "آ" :

در اين مورد ، واسط ، بمحض برخورد به كلمه مورد كاوش كه در آن "ا" بصورت چسبان يا غيرچسبان بكار رفته باشد يا شامل "آ" باشد ، جستجو را به كاوش براي كلمات جديدي كه با جايگزيني "ا" با "آ" ويا "آ" با "ا" ، ساخته شده اند ، بسط مي دهد. در نتيجه كاوش براي كلمه "فرايند" ، صفحات حاوي كلمه "فرآيند" ، از دست نمي رود.

ه) استفاده از اصطلاحنامه(Thesaurus) براي حل مشكل تنوع املايي كلمات :

اين معضل شامل تنوع استفاده از "ي" در كلمات عربي مختوم به "ا" ، تنوع املايي بعضي كلمات كه همه درست هستند ، استفاده از كلمات اروپايي بصورت ترجمه فارسي و استفاده يا عدم استفاده از جمع مكسر براي بعضي كلمات مي باشد كه حل مشكل كليه موارد ، در ايجاد يك پايگاه داده در سمت خدمت گذار ، مستتر است. اين پايگاه داده شامل نمايه اي از اين كلمات و كلمات مترادف مي باشد. براي مثال كلمه "موسي" ، به كلمه "موسا" و كلمه "كامپيوتر" به كلمه "رايانه" متناظر شده است. عامل هوشمند با مراجعه به اين پايگاه داده ، براي عبارت مورد كاوش كاربر ، عبارات مشابهي استخراج كرده ، كاوش را به جستجو براي اين عبارات ، علاوه بر عبارت اصلي ، بسط مي دهد.

ايجاد چنين پايگاه داده اي ، با مشاوره انجمن ها ، بزرگان و فرهنگستان ادب فارسي انجام مي پذيرد و بروزآوري آن نيز بصورت دوره اي و با دخالت صاحب نظران مذكور صورت مي گيرد.نمونه اي از محتويات اين پايگاه داده در جدول زير آمده است :

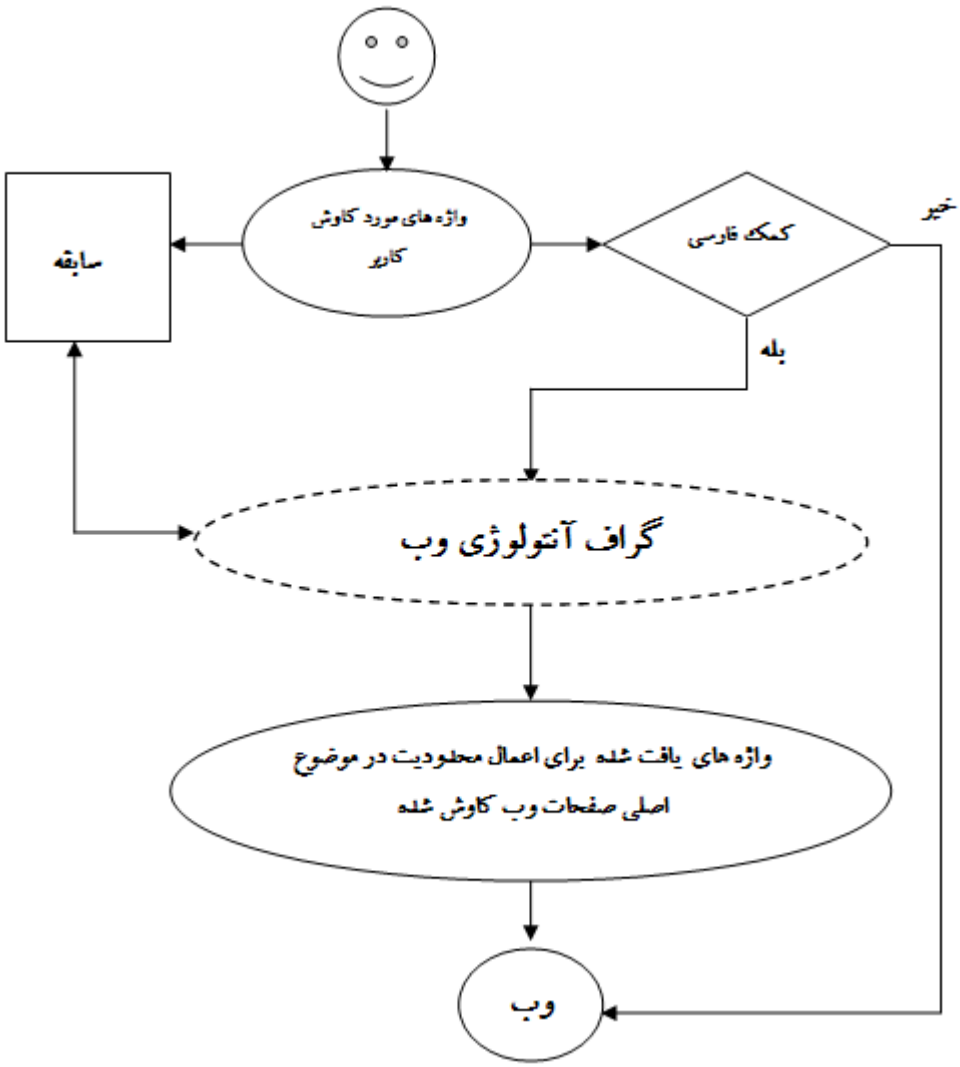
| شناسه | واژه اصلي | واژه مترادف |
|---|---|---|
| ١ | موسي | موسا |
| ٢ | اميراتور | اميراطور |
| ٣ | Ontology | آنتولوژي |
| ٣ | آنتولوژي | انتولوژي |
| ٣ | آنتولوژي | انتالوژي |
| ٣ | آنتولوژي | هستي شناسي |
| ۴ | كامپيوتر | رايانه |
| ۴ | Computer | كامپيوتر |
| ۵ | Source | منبع |
| ۵ | Source | سورس |

جدول (٤) نمونه اي از محتويات پايگاه داده مترادف ها.

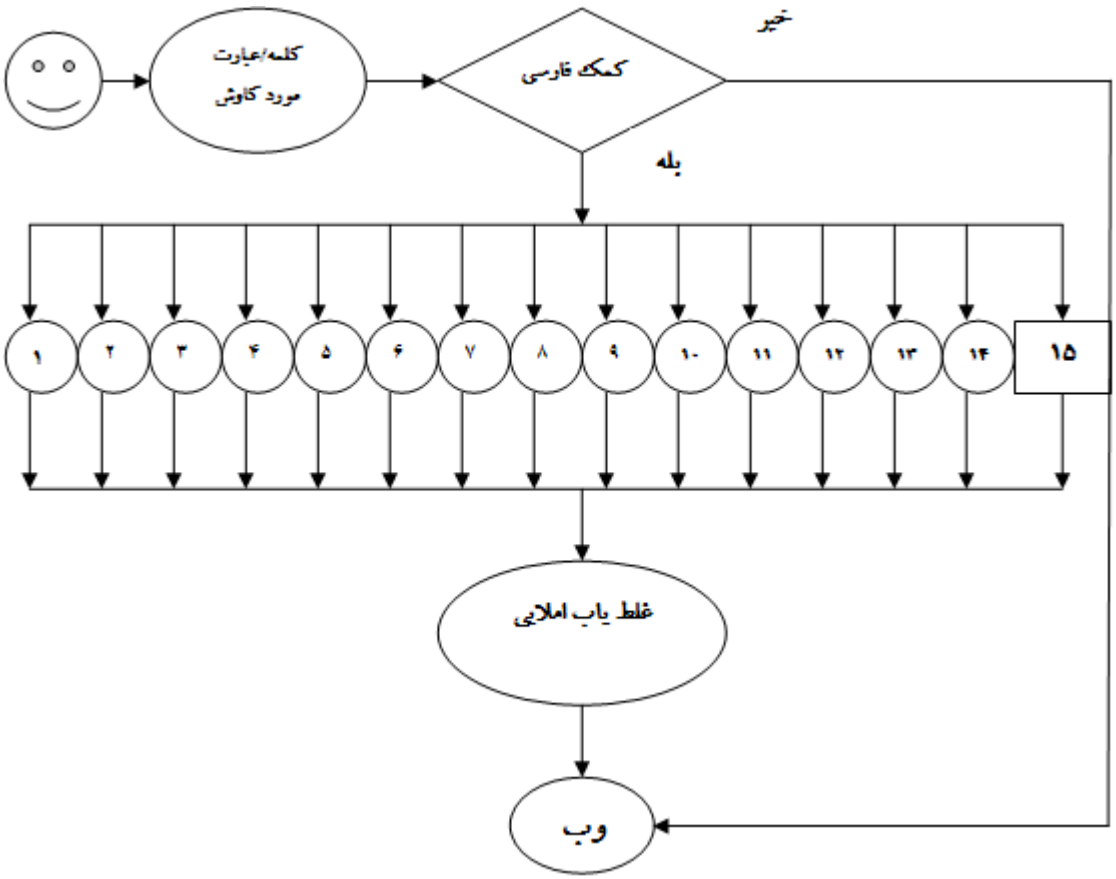
و) تبديل كلمات اروپايي به رسم الخط فارسي با همان تلفظ اصلي(Cross language Retrieval ) :

كـاربري كـه بـدنبال اطلاعـاتي در خصـوص برنامـه هـاي "Open Source" در اينترنـت مـي باشـد ، شـايد بـراي هميشـه از دسترسـي به صفحاتي كه در آنها كلمه "سورس باز" بكار رفته است ، محروم بماند يا حداقل محكوم به اتلاف زمان زيـادي تا رسيدن به چنين كلمه اي و به تبع ، نتـايج مطلـوب باشـد. لـذا در صـورتي كـه جسـتجو بـراي لغـت "سـورس" ، بنحـوي همزمان با كاوش براي كلمه "Source" ، حتي بدون اطلاع كاربر ، انجام پذيرد ، مي تـوان گفـت هـم در سـرعت و هـم در جامعيت اطلاعات بدست آمده ، ارتقايي صورت گرفته است.

وظيفه واسط ما در اين خصوص اينست كه با مراجعه بـه پايگـاه داده ، كـاوش را بـه كلمـه سـاخته شـده بـر اسـاس تلفـظ انگليسـي متناظر نيز گسترش دهد. براي انجام فرآيند حـل ايـن مشـكل بصـورت اتوماتيـك و ضـمنا اسـتفاده از پايگـاه داده معتبرتر و روزآمدتر بعنوان معيار عملكرد اين واسط ، مي توان روشـي پيشـنهاد نمود كه كلمه متناظر تلفظ انگليسـي لغـات كه با رسم الخط فارسي تهيه مي گردد ، با مراجعه به پايگاه هـاي داده بيـن الملـي حـاوي معـادل هـاي سـمبوليك تلفـظ كلمات انگليسـي (كه در كتاب هاي ديكشـنري انگليسـي به انگليسـي آمـده اسـت) ، كلمـه مـذبور را تهيـه نمـود و سـپس كاوش را براي آن انجام داد.




شكل (٣) ساختار واسط كمك فارسي براي بهبود مانعيت




شكل (٤) ساختار واسط كمك فارسي براي بهبود جامعيت

| نام جزء | پردازش مربوط | نام جزء | پردازش مربوط |
|---|---|---|---|
| C1 | حذف "ء" از عبارت | C7 | تبديل "آ" به "ا" |
| C2 | تبديل "ؤ" به "ئ" و بالعكس | C8 | تبديل "ا" به "آ" در ابتداي كلمات |
| C3 | تبديل "ئ" به "أ" و بالعكس | C9 | اضافه كردن "ء" به "ه" در تركيبات |
| C4 | تبديل "ؤ" به "أ" و بالعكس | C10 | حذف اعراب ها |
| C5 | تبديل "يي" به "ئي" و بالعكس | C11 | تبديل "ة" به "ه ء" يا به "ه" و بالعكس |
| C6 | تبديل "ي" به كاراكتر با يونيكد مشابه | C12 | مراجعه به پايگاه داده واژه هاي مترادف |

جدول (٥) اجزاء وپردازش هاي مربوط به رفع اشكالات رسم الخط

**پيشنهادات**

مطالعه حاضر با هدف بهينه سازي امكانات جستجو و بازيابي اطلاعات در ابزارهاي كاوش با واسط فارسي صورت گرفته است. بعنوان پژوهشي ديگر مي توان تمهيداتي جهت كاوش هر چه دقيقتر وب هاي فارسي زبان ، با هدف به حداقل رساندن تاثيرهاي سوء چالش هاي رسم الخط فارسي ، انديشيد و اين راه حل ها را بصورت تلفيقي (سري و موازي) نيز استفاده نمود.

مي توان نرم افزار واسط كمك فارسي مذكور را بصورت يك نوار ابزار ، برروي Browser ، نصب و استفاده نمود. از طرف ديگر مي توان بصورت يك نرم افزار كه بر روي Browser نصب شده بصورت پشت زمينه ، كلمات مورد كاوش را گرفته ، برروي آنها اعمال نظر كرده ، كاوش جديد خود را ترتيب دهد.

پژوهشي ديگر مي تواند در صورت امكان روشي را جستجو كند كه گراف معنايي مورد بحث را بصورت ماشيني ايجاد و گسترش دهد.

**فهرست منابع**

[1] . Velasquez Juan , Hiroshi Yasuda , Terumasa Aoki , 2003 , Combinig the Web content and usage mining

to understand the visitor behavior in a Web site , Third IEEE International Conference on Data Mining (ICDM'03)
www.cs.wisc.edu/~shavlik/ICDM_2003_Schedule.pdf
[2] . Chakrabarti S. , Martin van den Berg , Byron Domc , 1999 , Focused crawling: a new approach to topic-specific Web resource discovery , Computer Science and Engineering, Indian Institute of Technology, Bombay, 400076, India , FX Palo Alto Laboratory, 3400 Hillview Ave, Bldg 4, Palo Alto, CA 94304, USA ,c IBM Almaden Research Center, 650 Harry Rd, San Jose, CA 95120, USA
www.csd.uch.gr/~hy558/papers/chakrabarti99focused.pdf
[3] . Feiyu Xu , 2001 , Overview of FASTUS , DFKI LT-Lab
www.ics.mq.edu.au/~diego/publications/ALTSS03L2.pdf
[4] . Heflin J. , James Hendler , 2000 , Dynamic Ontologies on the Web ,Department of Computer Science University of Maryland College Park, MD 20742
www.cs.umd.edu/projects/plus/SHOE/pubs/aaai2000.pdf
[5] . Dr. Osmar , R. Zaïane , 2002 , Principles of Knowledge Discovery in Data , Chapter 9
www.cs.ualberta.ca/~joerg/courses/cmput695/fall2003
[6] اشرف زاده بهرام ، زبان فارسي در وبلاگ هاي فارسي ، ١٣٨٣ ،
http://www.persianfarsi.com/articles/zabaneweblog.htm

[٧]. جعفرقلي بيگلو موسي ، ١٣٨٢ ، مقايسه فرايند موضوع سازي ذهني جويندگان اطلاعات با ساختار سرعنوان هاي موضوعي فارسي ، علوم اطلاع رساني ، دوره ١٤ ، شماره ٣و ٤.
[٨]. حسيني بهشتي ملوك السادات ،١٣٨٣، كاربرد اصطلاح شناسي و واژه گزيني در نمايه سازي ماشيني و بازيابي اطلاعات ، عضو هيئت علمي مركز اطلاعات و مدارك علمي ايران ، علوم اطلاع رساني ، دوره ١٨ ، شماره ٣و ٤.
[٩]. خوانساري جيران ،١٣٨٢ ، تكامل وب و مقايسه ابزارهاي جستجو در اينترنت ، فصلنامه اطلاع رساني ، دوره ١٦ ، شماره ٣و٤ .
[١٠]. رضازاده ملك رحيم ، ١٣٨٠ ، تبيين و تدوين قواعد املاي فارسي ، گلرنگ يكتا ، چاپ اول.
[١١]. سمايي سيد مهدي ، عضو هيئت علمي مركز اطلاعات و مدارك علمي ايران١٣٨٢، مفرد و جمع در نمايه‌سازي ، فصلنامه اطلاع رساني. دوره ١٦، شماره ١و٢ ،
http://www.irandoc.ac.ir/ETELA-ART/16/16_1_2_3.htm
[١٢]فرهنگستان زبان و ادب فارسي ، ١٣٨٣ ، دستور خط فارسي.

[١٣]. كارنيرو آلبرتو ، ترجمه: عليرضا گنجي - دانشجوي كارشناسي ارشد كتابداري و ا طلاع‌رساني دانشگاه فردوسي مشهد ، ١٣٨٣ ، نقش منابع هوشمند در مديريت دانش ، فصلنامه اطلاع رساني. دوره ١٩، شماره٣ و ٤
http://www.irandoc.ac.ir/etela-art/19/19_3_4_9.htm
[١٤] كميجاني احمد ، ١٣٨٢، ساختار نمايه سازي در موتورهاي كاوش وب ، علوم اطلاع رساني ، دوره ١٧ شماره ٣و٤.

[١٥]. كوشا كيوان١٣٨١، معيارهاي ارزيابي ابزارهاي كاوش اينترنت: مطالعه مقايسه‌اي‌بر روي ابزارهاي كاوش وب با واسط جستجوي فارسي ، نشر كتابدار – مجله كتابدار.

[١٦]. محقق زاده محمد صادق - عضو هيات علمي دانشگاه علوم پزشكي شيراز ، زارعيان كاظم - كارشناس ارشد زبانشناسي ، ١٣٨٣ ، ارائه راه حل براي برخي مسائل اتوماسيون و نگارش فارسي ، فصلنامه اطلاع رساني. دوره ١٩، شماره ٣ و٤ ،
http://www.irandoc.ac.ir/etela-art/19/19_3_4_1.htm
[١٧] . مرتضايي ليلا ، ١٣٨٣ ، مسايل خط فارسي در ذخيره سازي و بازيابي اطلاعات ، فصلنامه اطلاع رساني . دوره ١٧ شماره ١و٢ .
[١٨]. هسي-يي اينگريد ، ترجمه: قاسم آزادي دانشجوي كارشناسي ارشد كتابداري و اطلاع‌رساني دانشگاه تهران ، ١٣٨٣ ، اينترنت : سازماندهي و جستجو ، فصلنامه اطلاع رساني. دوره ١٨، شماره ٣و ٤
http://www.irandoc.ac.ir/ETELA-ART/18/18_3_4_10.htm

[۱۹]. محسني ياسمن ، ايجاد و نمايش انتالوژي براي شبكه مفاهيم مرتبط با حوزه مخابرات نوري ، ۱۳۸۰ ، مركز تحقيقات مخابرات ايران.
[۲۰]. عبدالهي بهناز ، بررسي روشهاي طراحي و ايجاد انتولوژي ، ۱۳۸۰ ، مركز تحقيقات مخابرات ايران.
[۲۱]. جباري فر معصومه ، بررسي پارامترهاي ارزيابي و ليست دسته بندي شده جويشگرها ، ۱۳۸۰ ، مركز تحقيقات مخابرات ايران.
[۲۲]. صالحی مازیار ، زارع بیدکی علی محمد ، ارائه RFP برای یک جویشگر دوزبانه فارسی/ انگلیسی ، ۱۳۸۰ ، مرکز تحقیقات مخابرات ایران.
[۲۳]. ميريان مريم السادات ، ارائه چارچوب كلي براي زيرسيستمهاي Information Retrieval and Query Processing ، ۱۳۸۰ ، مركز تحقيقات مخابرات ايران.

---

- هرگونه استفاده تجاري از اين پژوهش ، بدون اجازه پژوهشگران ممنوع است.
- حق انتشار (الكترونيكي و چاپ)براي مركز اطلاعات و مدارك علمي ايران محفوظ است