نام مقاله: مهندسي خودكار هستي شناسي: امكان سنجي استخراج روابط معنايي از متون فارسي و تعيين ميزان پيدايي آنها

نام نشريه: فصلنامه كتابداري و اطلاع رساني (اين نشريه در www.isc.gov.ir نمايه مي شود)

شماره نشريه: 46 \_ شماره دوم، جلد 12

پديدآور: عاطفه شريف

مهندسي خودكار هستي شناسي: امكان سنجي استخراج روابط معنايي از متون فارسي و تعيين ميزان پيدايي آنها

عاطفه شريف

**چكيده**

**در اين مقاله ضمن تبيين فرايند مهندسي هستي شناسي‌ها، استخراج روابط معنايي با تكيه بر روشهاي مبتني بر الگو، مطالعه شده است. نمونه‌اي از متون تخصصي فارسي در حوزة ربط تحليل و روابط معنايي موجود در آن استخراج و دسته‌بندي گرديد. همچنين، تعيين ميزان پيدايي روابط معنايي در نمونة مورد تحليل، در پاسخ به دومين پرسش پژوهش انجام پذيرفت. امكان استخراج و تعيين روابط معنايي در نمونه مورد تحليل، تأييد شد. اين در حالي است كه ميزان پيدايي روابط، در سطح پايين ارزيابي مي‌شود.**

**كليدواژه‌ها: استخراج متن، استخراج دانش مفهومي، الگوهاي معنايي، پيدايي روابط معنايي، خودكارسازي، مهندسي هستي‌شناسي.**

**حرف اول: هستي شناسي وب**

به منظور كاهش مشكلاتي چون محدوديت دامنة معنايي و انعطاف پذيري اندك نظامهاي ذخيره و بازيابي، همچنين افزايش قابليتهاي استنتاج در نظامها، حركتهايي به سمت نظامهاي بازنمون دانش جديد در جريان است. اين حركتها كه بيشتر در حوزة هوش مصنوعي مشاهده مي‌شود، با ظهور كتابخانه‌هاي ديجيتال و انديشة وب معنايي، رنگ ديگري به خود گرفته است. تحقيقات در زمينة هستي‌شناسي‌ها به عنوان ابزارهاي جديد بازنمون دانش، از جمله اين حركتهاست (شريف، 1385، ص 67). هستي‌شناسي‌ها به منزلة ابزار بازنمون دانش در نظامهاي ذخيره و بازيابي، استفاده مي‌شوند (ونگ و ديگران، 2006) و آن را مجموعه‌اي از مفاهيم، خصيصه‌ها، و روابط ميان آن مفاهيم تعريف كرده‌اند (هس و اشليدر، 2006، ص545؛ ). اين تعريف در حوزة مدل‌سازي مفهومي، چندان جديد نيست. در مدلهاي موجوديت ـ رابطه كه از دهه 1970 در پايگاه‌هاي اطلاعاتي استفاده مي‌شود و در مدلهاي گسترش يافتة آن نيز چنين الگويي از مفاهيم، خصيصه‌ها و روابط قابل شناسايي است. اما دليل اين همه استقبال از هستي شناسي ها در اين نكته نهفته است كه هستي‌شناسي‌ها برخلاف مدلهاي مفهومي پيش گفته، استنتاج هوشمند را ممكن مي‌سازند (هس و اشليدر، 2006، ص545).

معماريهاي متنوعي در طراحي و اجراي هستي شناسي ها وجود دارد. اين معماريها در سه دسته قابل بررسي اند (ميكا ، 2006، ص 289):

**1. هستي‌شناسي واحد:** در اين صورت، تمام افراد سازمان به تفاهم وتوافق بر هستي شناسي واحدي مي رسند. اين معماري تنها در موارد بسيار ايده‌آل و هوشمند مؤثر خواهد بود. نقطه قوّت اين نوع معماري آن است كه در صورت تمايل به اشتراك، به هيچ‌گونه ترسيمي نياز نيست.

2**. هستي‌شناسي‌هاي محلي:** در كنار هستي شناسي واحد: در اين نوع معماري، گروه‌هايي از افراد سازمان فعال در حوزه‌اي خاص از هستي‌شناسي محلي استفاده مي‌كنند. در اين صورت چنانچه اشتراك دانش ميان گروه ها نياز باشد، استفاده از ترسيم هستي شناسانه ضروري است.

**3. هستي‌شناسي‌هاي فردي:** هنگامي كه هر يك از افراد در سازمان از هستي‌‌شناسي منحصر به فرد و مستقلي استفاده كند، اين معماري تحقق مي‌يابد. در اين صورت، استفاده از ترسيمها با وسعت بيشتري انجام مي‌پذيرد.

از ميان سه مدل معماري ذكر شده، دومين معماري، قابل دفاع ترين نوع معماري به شمار مي‌رود. «برنرزلي» در توضيح اين مطلب مي افزايد: معماري نخست بر كنترل مركزي بنا شده است، در حالي كه سومين معماري به مكاني آرماني نظر دارد! (نقل در ميكا ،2006، ص 289). بر مبناي معماري دوم، هستي‌شناسي واحد مشتركي در سطح بالا تعريف مي‌شود و همزمان، امكان افزودن و گسترش هستي‌شناسي به‌طور محلي براي هر يك از نظيرها [كارگروه‌هاي مشابه] فراهم مي‌آيد. اغلب اين نوع معماري به‌طور سلسله مراتبي به اجرا مي‌رسد؛ به گونه‌اي كه تمام زيرمجموعه‌ها ـ نظيرها ـ از هستي‌شناسي واحد مشتركي استفاده مي‌كنند و هر زيرمجموعه مي‌تواند گسترشهاي لازم را به هستي‌شناسي سطح بالا بيفزايد و هستي‌شناسي محلي خود را داشته باشد (ميكا ، 2006، ص 289).

امروزه با توجه به روند رو به رشد استفاده از هستي‌شناسي‌ها در نظامهاي اطلاعاتي، ساخت هستي‌شناسي‌ها، روش‌شناسي ساخت، ابزارهاي ساخت، ساخت خودكار و يادگيري هستي‌شناسي‌ها، از مباحث مطرح در ميان محققان است (شمس فرد، عبداله‌زاده بارفروش، 1381: ص 49). رويكردهاي متنوعي در ساخت، توسعه و روزآمدسازي هستي‌شناسي‌ها در جريان است. طيفي از روشهاي متنوع در استخراج و مدل‌سازي دانش در متون مختلف قابل شناسايي است. اغلب در ساخت هستي‌شناسي‌ها از روشهاي نيمه خودكار استفاده مي‌شود، اما مطالعاتي در زمينة خودكارسازي كامل اين فرايند نيز در جريان است (ونگ و ديگران، 2006). صرف نظر از نوع معماري انتخابي در ساخت و توسعه هستي‌شناسي‌ها، در مورد شيوه‌هاي استخراج مفاهيم و روابط نيز بايد مطالعه و تصميم گيري شود.

ساخت و توسعه هستي شناسي ها به سه شيوة كلي انجام پذير است:

ساخت دستي

·استفاده از ابزارهاي مهندسي هستي شناسي [كه انجام بخشي از كار را به صورت نيمه خودكار امكان پذير مي سازد] و

روشهاي خودكار.

ساخت دستي هستي‌شناسي‌ها، براي قلمروها و كاربردهاي مختلف، پرهزينه، وقت‌گير و مستعد خطاست و هستي‌شناسي‌هايي كه به صورت دستي ساخته مي‌شوند، معمولاً گران، متمايل به نظرهاي شخصي طراح، انعطاف‌ناپذير در مقابل تغييرات و دقيقاً خاص منظوري هستند كه براي آن تهيه شده‌اند (شمس فرد، عبداله زاده بارفروش، 1381: ص 49). در ديگر سوي، ابزارهاي مهندسي هستي شناسي مانند Protégé واسط كاربري فراهم مي آورند كه در آن امكان تعريف مفاهيم، نمونه‌ها، ويژگيها و محدوديتهاي مفاهيم و همچنين روابط، وجود دارد. در اين صورت، باز هم هزينة زيادي در ساخت هستي‌شناسي‌ها صرف مي‌شود و افراد متخصصي براي ساخت هستي‌شناسي در فعاليت شركت دارند. مزيت استفاده از ابزارها آن است كه در صورت وجود، مي‌توان از مفاهيم و روابط موجود در ساير هستي‌شناسي‌ها به منظور ساخت و توسعة هستي‌شناسي جديدي استفاده كرد؛ لذا بخشي از كار به صورت خودكار انجام پذير است.

اما در روشهاي خودكار ـ سومين شيوة مهندسي و توسعه هستي‌شناسي‌ها ـ دخالت عامل انساني كاهش و سرعت ساخت و توسعة هستي شناسي ها افزايش مي يابد. در اين صورت، بايد به اكتساب يا استخراج دانش پرداخت. از آنجا كه هستي‌شناسي را مجموعة مفاهيم و روابط آن مي‌دانيم، اين استخراج در دو بُعد انجام مي‌پذيرد: الف) استخراج مفاهيم و ب) استخراج روابط معنايي ميان اين مفاهيم.

**مروري اجمالي بر پيشينه**

مشكل اصلي انسان با ماشين، چگونگي آموختن مفاهيم به آن است. در يك نگاه كلي، روشهاي يادگيري ماشيني به دو دستة «روشهاي رياضي/ آماري» و «روشهاي زبان پايه» قابل تقسيم است. براي مثال، فنون مورد استفاده در نمايه‌سازي معناي پنهان صرفاَ بر محاسبات رياضي استوار است، بدون اينكه نيازي به درك مفهوم از سوي ماشين باشد (يو و ديگران، 2002). در اين روش، محدوديت زباني به حداقل مي‌رسد و نوع زبان به فراموشي سپرده مي‌شود. در واقع، در روش نمايه‌سازي معناي پنهان ويژگيهاي زباني ناديده گرفته مي‌شود. از سويي، بسياري از روشهاي يادگيري بر روشهاي زبان پايه استوار است (مانند ميلر و ديگران، 2000 و زلنكو و ديگران، 2003؛ نقل در زو و زانگ، 2007، ص970). كاربرد مؤلفه‌هاي زبان شناختي در دامنة وسيعي ـ دانش واژگاني، درخت تجزيه، درخت وابستگي و مؤلفه‌هاي معنايي ـ در برخي از متون مورد توجه قرار گرفته است. به‌طور مثال، «زانگ، زو، او» (2008) با بررسي مؤلفه‌هاي نحوي ساخت يافته، روابط معنايي را استخراج كرده‌اند.

پروژه‌هاي مختلف ساخت و به كارگيري هستي‌شناسي‌ها بويژه در محيطهاي اينترانت، از جمله مواردي است كه نمونه هاي فراواني از آنها در حوزه هاي موضوعي متنوع موجود است (از جمله راسكين و پن، 2005؛ كازانوا و ديگران 2005). «كازانوا و ديگران» (2005) مراحل و چگونگي ساخت هستي‌شناسي موضوعي حقوق را با استفاده از زبان طبيعي و بر مبناي پژوهش قوم نگارانه بيان مي دارند. گفته شد كه ساخت هستي‌شناسي‌ها اغلب نيمه خودكار صورت مي‌پذيرد. با وجود اين، پژوهشهايي نيز در زمينه خودكارسازي كامل فرايند مهندسي هستي‌شناسي‌ها در جريان است (مانند ونگ و ديگران، 2006؛ آقابك، آيدين، اوزمل، و آكسوي، 2006) در بسياري از پژوهشها بر روشهاي استفاده از ابزارهاي بازنمون سنتي دانش مانند پايگاه‌هاي اطلاعاتي موجود (كيوره، 2003) و يا اصطلاحنامه‌ها (سورگل و ديگران ، 2004) در ساخت و توسعه هستي‌شناسي‌ها تأكيد شده است.

اگر چه مطالعات زيادي در زمينة تحليل و استخراج دانش مفهومي در متوني به زبان انگليسي انجام پذيرفته است، تنها مطالعه‌اي كه در ايران و بر روي زبان فارسي انجام شده پايان‌نامة دكتراي شمس فرد (1381) مي‌باشد. «هستي» كه نظام يادگير هستي‌شناسي است، حاصل پژوهش شمس فرد است. در اين نظام، مفاهيم و روابط معنايي با استفاده از الگوهاي زباني و معنايي استخراج شده‌اند. «هستي» سيستمي براي استخراج دانش مفهومي از متون سادة زبان فارسي و ساخت هستي‌شناسي از روي آنهاست. «هستي»، از پايه، به ساخت خودكار هستي شناسي مي پردازد. منظور از «پايه»، نبود هستي‌شناسي مبنا (اعم از عمومي يا تخصصي) و همچنين نبود واژگان معنايي براي كمك به فرايند يادگيري است. در ابتداي كار نظام، واژگان تقريباً تهي و هستي شناسي فقط حاوي هستة اوليه يادگيري است كه به صورت دستي ساخته شده است.

هستة اوليه يادگيري شامل الگوهاي زباني و معنايي است. اين الگوها كه از طريق تحليل متن به زبان فارسي انجام پذيرفته است در نظام قرار مي‌گيرد و از آن پس، استخراج با توجه به الگوها انجام مي‌پذيرد. بخشي از الگوهاي معنايي مورد استفاده در «هستي» در (شمس فرد، عبداله زاده بارفروش، 1381) بيان شده است.

**طرح پرسش**

يكي از مهم‌ترين دغدغه‌هاي انسان در عصر فناوري اطلاعات، طراحي و گسترش ابزارها، امكانات و خدمات مربوط به گردآوري، ذخيره سازي و پردازش داده‌هاي زباني است (عاصي، رضاقلي فاميان، آقاجاني، 1385، ص 125). علاقه به استخراج خودكار، بويژه با توجه به افزايش روزافزون اطلاعات متني دسترس‌پذير در محيط وب و آرشيوهاي ديجيتال، رو به افزايش است (زو و زانگ، 2007،ص 969) استخراج مفاهيم و روابط از طريق تحليل داده ها صورت مي‌پذيرد. داده‌ها در متن مانند داده‌هاي موجود در يك صفحه وب، يا يك مقاله، و يا خارج از متن، به طور مثال در يك پايگاه كتابشناختي قرار دارند. به طور كلي، داده‌هاي اوليه كه براي تحليل مورد استفاده قرار مي‌گيرند، در سه دسته تقسيم‌بندي مي‌شوند:

داده‌هاي ساخت يافته مانند داده هاي موجود در يك پايگاه كتابشناختي [در فيلدها و فيلدهاي فرعي]

داده‌هاي نيمه ساخت يافته مانند مستنداتي در زبان XML يا HTML ؛ و سخت‌ترين حالت

تحليل و استخراج از متوني به زبان طبيعي.

ساخت خودكار هستي‌شناسي‌ها با استفاده از نظامهاي يادگير هستي‌شناسي انجام مي‌پذيرد. اين در حالي است كه استفاده از ابزارهايي چون Protégé  تنها به منزلة پشتيبان ساخت به كار مي‌روند. «شمس فرد و عبداله زاده بارفروش» (1381) در توضيح «هستي»، به نقطه شروع استخراج دانش مفهومي مي پردازند و نظامهاي يادگير هستي شناسي را به دو دسته تقسيم مي‌كنند:

دسته‌اي از نظامها كه از دانش زباني (دستور زبان، دانش لغوي، الگوها، و ...) به عنوان دانش پيش زمينه استفاده مي‌كنند و از منابع ورودي، دانشهاي جديد را مي‌آموزند.

دسته‌اي كه اين فرايند را با استفاده از دانش مفهومي (هستي‌شناسي مبنا) به انجام مي‌رسانند. (ص 50).

اما اغلب نظامهاي موجود يادگير هستي‌شناسي، دسترسي به هر دو را ترجيح مي‌دهند؛ بدين ترتيب كه از ابزارهايي چون WordNet ،كه نظامي از واژگان معنايي از پيش تعريف شده است، نيز استفاده مي‌كنند. اين نظام واژگاني در تشخيص و ايجاد تمايز ميان رابطه‌ها كاربرد دارد و از مشكل پراكندگي داده‌ها در استخراج روابط مي‌كاهد (زو و زانگ، 2007، ص 975) وردنت علاوه بر دانش زباني، حاوي بخشي از دانش مفهومي نيز مي‌باشد.

استخراج دانش در دامنه‌اي از روشهاي با دانش ضعيف (مانند تكنيكهاي آماري) تا روشهاي غني از دانش (مانند استدلال منطقي) گسترده اند (شمس فرد، عبداله زاده، 1381، ص 51). در نظامهايي كه به روشهاي آماري عمل مي كنند، بسامد تكرار و يا بسامد هم وقوعي و هم مكاني كلمات و عبارتها مورد توجه است و از تحليل آماري داده‌هاي هم وقوع براي يادگيري طبقات و روابط مفهومي استفاده مي‌شود. برخي از نظامهاي ديگر، روشهاي نماديني چون روشهاي منطقي مبتني بر الگو و زبان ـ پايه را براي استخراج دانش به كار مي‌گيرند. روشهاي زبان ـ پايه مانند تحليل كامل نحوي (ميلر و ديگران، 2000 نقل در زو و زانگ، 2007، ص 970)، تحليل ساختواژي ـ نحوي (اسدي، 1997، نقل در شمس فرد، عبداله‌زاده، 1381، ص51)، تجزية الگوهاي لغوي ـ نحوي (فينكستاين ـ لندو و مورين، 1999، نقل در شمس فرد، عبداله زاده، 1381، ص 51)، پردازش معنايي و درك متن عموماً وابسته به زبان هستند و براي استخراج دانش از منابع غيرساخت يافته (زبان طبيعي) به كار مي‌روند. در روشهاي مبتني بر الگو، ورودي (معمولاً متن) به دنبال الگو يا كلمات كليدي خاص كه نشانگر روابط مفهومي خاصي است، جستجو و اطلاعات مورد نظر از متن استخراج مي‌شود.

استخراج مفاهيم و روابط در ساخت هستي‌شناسي‌ها پيچيدگي‌هاي فراواني دارد و مسيرهاي متفاوتي نيز در ساخت و توسعة آنها طي شده است. در اين ميان، تحليلها بيشتر بر روي زبان انگليسي صورت گرفته و مطالعات اندكي بر روي زبان فارسي انجام پذيرفته است (پايان‌نامه شمس فرد، 1381). همچنين، نظام واژگاني چون WordNet نيز در زبان فارسي موجود نيست. اكنون با توجه به آنچه بيان شد، اين پرسش مطرح است كه «آيا مي‌توان با استفاده از تحليل متون علمي زبان فارسي، به منزلة يك پيكره زباني كه در مقايسه با ساير متون ابهام كمتري دارد، روابط معنايي ميان مفاهيم را استخراج كرد؟» منظور از روابط معنايي در اين پرسش، مجموعه روابط معنايي معمول در ابزارهاي بازنمون سنتي مانند اصطلاحنامه‌ها ـ سلسله مراتبي و هم بسته- به علاوة گروهي از روابط غني‌تر معنايي است كه در نظامهاي جديد بازنمون ـ هستي‌شناسي‌ها- مطرحند. چنانچه پاسخ اولين پرسش مثبت است، «پيدايي يا وضوح روابط معنايي در متون علمي زبان فارسي تا چه ميزان است؟» به بياني «آيا مي‌توان به كشف الگوهايي براي استخراج روابط معنايي در اين متون اميدوار بود؟» پژوهش حاضر در راستاي پاسخ به دو پرسش طرح شده، اجرا گرديد.

**هدف پژوهش**

گفته شد بخشي از خودكارسازي فرايند ساخت هستي‌شناسي‌ها با كمك ردگيري الگوهاي معنايي امكان‌پذير است. چنانكه پيشتر آمد، در روشهاي مبتني بر الگو، الگو يا كلمات كليدي خاص كه نشانگر روابط باشند، مورد جستجو هستند و از متن استخراج مي‌شوند. ردگيري الگوهاي معنايي با اين پيش فرض انجام مي‌پذيرد كه قابليت و امكان استخراج روابط به روش تحليل متن عملي است و اين روابط پيدايي لازم را براي استخراج دارند، زيرا تنها در صورتي مي‌توان به استخراج خودكار روابط اميدوار بود كه پيدايي اين روابط در متن در حد قابل قبولي باشد. لذا هدف از اين پژوهش، بيان الگوهاي معنايي نيست، بلكه پژوهشگر قصد دارد پيش فرض روشهاي مبتني بر الگو را بررسي كند.

**دامنة كار و روش پژوهش**

با توضيحي كه در هدف پژوهش مورد اشاره قرار گرفت و با توجه به اينكه:

هستي‌شناسي حاصل بحث و توافق نظر بر سر معناست و هدف از ايجاد آن، بيان واضح توافقي است كه در باب «چگونه بودن جهان هستي» از جنبه‌اي خاص [در حوزه‌اي مشخص] حاصل آمده است (مانسيني و شام، 2006، ص 1169) و لذا در حوزه‌اي كاربردي، طراحي و مورد استفاده قرار مي‌گيرد.

تفاوتهاي زباني در يك جامعة مشخص به واسطة تفاوت در تجربيات مختلف افراد در گروه‌هاي اجتماعي و حرفه‌اي خاص ايجاد مي‌شود و چندان دور از ذهن نيست كه ساختار دانش در يك جامعه، به گونه‌اي ويژه و گاه متفاوت از ساير جوامع شكل گيرد. به بياني، افراد در گروه‌هاي اجتماعي خاص (به عنوان مثال در جامعه‌اي حرفه‌‌اي) ساختار دانش ويژه‌اي دارند. اين افراد به هنگام بازنمايي دانش خويش در متون، از واژه‌ها و مفاهيم مشتركي استفاده مي كنند (شريف، 1387) كه مي‌تواند متفاوت از واژگان حوزه‌اي ديگر باشد و در ساخت جمله و انتقال معنا نقش دارند.

مقاله‌هاي علمي يك رشته، اغلب نمايش مفاهيم جديد، همراه با جزئيات آن، از جمله روابط ميان مفاهيم است و ابهام كمتري دارد. همچنين

استخراج دانش، به دو بخش «استخراج مفاهيم« و «روابط» تقسيم مي‌شود.

لذا، در اين پژوهش تمركز بر متون تخصصي در يك حوزة موضوعي ويژه ـ ربط ـ قرار گرفت. بدين منظور، مقاله‌هاي مجلة اطلاع‌شناسي (1383، سال دوم، شماره اول) كه ويژه‌نامة ربط است، براي تحليل و استخراج روابط معنايي انتخاب شد. اگر چه به واسطة تحليل انجام شده بر روي متن، مفاهيم نيز استخراج شده است، در اين پژوهش تأكيد بر روابط معنايي و ميزان پيدايي آنهاست.

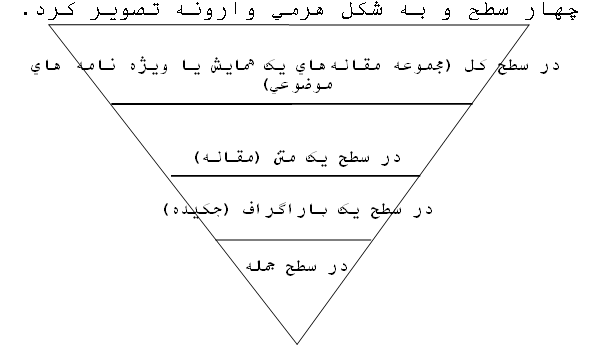
دانش قابل استخراج از متون به دو دستة «ضمني» و «عيني» تقسيم‌بندي مي‌شود. دانش عيني، دانشي است كه در متن و با استفاده از واژگان به بيان آمده است، اما دانش ضمني وابسته به پيش داشته‌هاي فرد انتزاع كننده آن است و به طور صريح در متن و در قالب واژگان نيامده است؛ بدين معنا كه اگر چه رابطه معنايي به طور صريح در متن نيامده است، خوانندة مطلب كه در اين جا فرد انتزاع كننده روابط معنايي (پژوهشگر) مي باشد، با توجه به دانش زمينه‌اي، قادر به انتزاع و استخراج رابطه‌هاست. در اين پژوهش، سعي شده هر دو نوع دانش استخراج شود. علت استخراج هر دو نوع دانش آن است كه ميزان پيدايي روابط معنايي را مي توان از مقايسه و تحليل اين دو يافته تعيين نمود. در عين حال، واضح است كه استخراج خودكار روابط معنايي بر پاية دانش عيني و الگوهايي است كه به طور عيني در متن آمده است و استخراج آنها را ممكن مي‌سازد.

در مقاله‌هاي مورد بررسي، دو بخش چكيده و بحث و نتيجه گيري براي استخراج روابط و الگوهاي آن انتخاب شده است. مبناي اين انتخاب بر اين استدلال استوار است كه در چكيده مقاله‌ها، اصلي‌ترين مطلب مقاله به صورت مختصر بيان مي‌شود و در بخش بحث و نتيجه‌گيري، حاصل سخن مقاله. همچنين عنوان و كليدواژه ها نيز به منظور استخراج بخشي از دانش ضمني مورد تحليل قرار گرفت.

بنابراين، نمونة مورد تحليل متشكل از9 مقاله به زبان فارسي است كه در ويژه‌نامة ربط مجله اطلاع شناسي به چاپ رسيده است. استخراج روابط با تجزيه متن به پاره‌هاي مورد تحليل، صورت گرفت. بدين منظور، علامت نقطه (.) در متن به منزلة مرز پاره‌هاي مورد تحليل، در نظر گرفته شد. تحليل بدون توجه به ساختار نحوي جمله‌ها و تنها با استناد به دانش موضوعي پژوهشگر، انجام پذيرفت.

**مراحل و روش تحليل**

مفاهيم در قالب زنجيره‌اي از واژگان به هم مرتبط، به بيان و بنان در مي‌آيند. در اين صورت است كه جمله متولد مي‌شود. جمله در ارتباط با ساير جمله‌هاي يك پاراگراف، در خدمت انتقال معنايي كلي‌تر است كه از عهدة يك جمله به تنهايي بر نمي‌آمد. در اين مورد، چكيدة مقاله علمي، تك پاراگراف منحصر به فردي است كه قصد دارد بازنموني مختصر و تا حدّ ممكن جامع از رويكرد، روشها و البته معناي مورد نظر نويسنده باشد. مجموعه پاراگرافها در خدمت نويسنده‌اند تا مقصود و معناي مورد نظر به تمامي منتقل شود. هر مقاله نيز با ساير مقاله‌هاي هم موضوع در ارتباط است و در كنار آنها زيست مي‌كند. ويژه‌نامه‌هاي تخصصي، مجالي هستند براي بروز اين زيستن.

****

**شكل 1. هرم تحليل استخراج مفاهيم و روابط**

**در نخستين سطح، سطح جمله، بخشي از مفاهيم و روابط شناسايي مي‌شود. در اين سطح از تحليل، دانش عيني آن جمله مورد نظر است و فراتر از همان جمله نمي توان حركت كرد؛ اما چنانچه سطح تحليل به پاراگراف گسترش يابد، مي‌توان مفاهيم را در سطحي انتزاعي‌تر و به بياني ضمني‌تر ديد. به طور مثال، اگر در جمله اول يكي از روشهاي درمان بيماري خاصي بيان شود، تنها همين يك روش درماني استخراج مي‌گردد؛ اما اگر پاراگراف در نظر گرفته شود و جمله اي ديگر از اين پاراگراف به همان روش با عبارتي ديگر بپردازد، رابطة ضمني هم ارزي اين دو مفهوم قابل شناسايي است، در حالي كه در متن نيامده است. بدين ترتيب، روابط ضمني‌تري قابل شناسايي و استخراج خواهد بود؛ روابطي كه شايد در سطح جمله امكان استخراج آنها وجود نداشت. بتدريج، با حركت در سطوح تحليل، مجموعه‌اي از روابط جديد ميان مفاهيم و حتي ميان رابطه ها آشكار و در نهايت هستي آن كل، به شكل مجموعه‌اي از مفاهيم و روابط متجلي مي‌شود.**

**تحليل و استخراج روابط معنايي مطابق با هرم تحليل صورت گرفت. عمليات تحليل و استخراج توسط پژوهشگر انجام شد؛ بدين ترتيب كه ابتدا تك تك خطوط متن مورد تحليل، تحليل شد و روابط معنايي استخراج گرديد. سپس چند خط به طور همزمان مورد نظر قرار گرفت و روابط انتزاعي تر استخراج شد. بديهي است، اين‌گونه تحليل، مانند ساير روشهاي كيفي، فارغ از سوگيريهاي پژوهشگر نيست. داده‌هاي حاصل از تحليل ـ مفاهيم و روابط معنايي ـ در نرم‌افزارMS Excel 2003  ثبت شد. در مواردي كه روابط چندين بار تكرار شده است نيز رخداد تكرار ثبت شد. بدين ترتيب، هر دو مفهوم به علاوة رابطه ميان آن دو، در يكي از رديفهاي نرم‌افزار قرار گرفت و رخداد تكرار آن ثبت شد.**

**به منظور تعيين ميزان پيدايي روابط معنايي سه سطح ـ انطباق كامل، انطباق نسبي، و عدم انطباق ـ تعيين شد. انطباق كامل زماني رخ مي‌دهد كه رابطه معنايي استخراج شده به طور كامل در جملة مورد تحليل آمده باشد؛ مثلاً در جمله «تكنيك فرامتن راهكار افزايش ربط است» رابطه «راهكار» ميان دو مفهوم «تكنيك فرامتن» و «افزايش ربط» نمونه يك رابطه با انطباق كامل است. در جمله‌اي مانند «ساختار نظامهاي فرامتن و تأثير آنها بر مسئله ربط در بازيابي و نيز راهبردهاي جستجو ...» رابطة «تأثير دارد بر»، داراي انطباق نسبي است. اما هنگامي كه از جمله‌اي مانند «پيوندهاي فرامتني راه ديگري را براي مسيريابي در ميان انبوه اطلاعات فراهم مي‌كند»، رابطه «كاربرد دارد در» استخراج مي‌شود، اين رابطه مصداق نبود انطباق ميان واژه‌هاي متن و رابطه، استخراج شده است. بدين ترتيب، مجموعه‌اي از روابط معنايي همراه با ميزان انطباق آنها با متن استخراج شد.**

**روابط معنايي مورد استخراج**

**مجموعه روابط معنايي قابل تعريف در ابزارهاي بازنمون سنتي از جمله در اصطلاحنامه‌ها عبارتند از: روابط سلسله مراتبي، هم‌ارز و هم‌بسته. رابطة هم ارز ميان دو اصطلاح پذيرفته شده و پذيرفته نشده. رابطة سلسله مراتبي اصطلاحات اعم و اخص را مشخص مي‌سازد و اصطلاحاتي كه نوعي تبادر به ذهن مي‌‌آورند، در دستة روابط هم‌بسته تعريف مي‌شوند. اما روابط معنايي حاكم در ابزارهاي بازنمون دانش جديد، از جمله هستي‌شناسي‌ها، محدود به اين روابط ثابت نيست و پويايي بيشتري دارد و عمق آن بسته به عمق هستي‌شناسي‌ها، متنوع است.**

**روابط معنايي در گستره‌هاي متفاوت، به گونه‌ها و با جزءنگري‌هاي متنوعي تقسيم‌بندي شده‌اند كه در اين مقاله به آنها نخواهيم پرداخت. رابطه‌هاي منعطفي چون «راه انتقال» براي يك بيماري خاص در هستي شناسي بيماريها و «نام تجاري» براي هستي‌شناسي داروها، نمونه‌اي از رابطه‌هاست. رابطه‌هاي استخراج شده در تحليل متن انتخابي نيز با همين رويكرد، استخراج شده است.**

**تجزيه و تحليل يافته‌ها**

**تجزيه و تحليل نهايي بر روي داده هاي ثبت شده، در نرم‌افزار اكسل انجام پذيرفت. روابط دسته‌بندي و در 7 مقولة كلي قرار گرفت. اين هفت مقوله كلي شامل 24 زير مقوله هستند كه در 122 مورد، رخداد داشته‌اند. جدول شماره 1، رابطه‌هاي استخراج شده را همراه با رخداد آنها نشان مي‌دهد.**

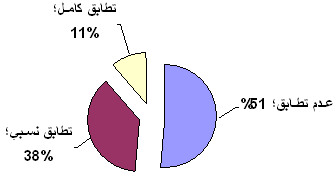
|  |  |  |  |
| --- | --- | --- | --- |
| **رخداد** | **مقولة فرعي رابطه استخراجي** | **مقولة كلي رابطة استخراجي** | **رديف** |
| 4 | معادل است با | هم ارزي | 1 |
| 2 | تفاوت دارد با | تفاوت (نا هم ارزي) | 2 |
| 2 | دسته بندي مي شود به | سلسه مراتبي | 3 |
| 10 | انواع دارد |
| 8 | در ارتباط است با | همبستگي | 4 |
| 1 | استفاده مي شود در | كاربردي | 5 |
| 2 | راهبرد دارد |
| 1 | راهكار |
| 1 | راهكار سنجش عملكرد |
| 8 | كاربرد دارد در |
| 2 | تأثير دارد بر | تأثيري (تأثيرگذاري/تأثيرپذيري) | 6 |
| 10 | نقش دارد در |
| 2 | تأثير دارد در |
| 3 | تأثير مي‌گذارد بر |
| 4 | تحت تأثير |
| 6 | براساس |
| 1 | برپايه |
| 2 | زيربناي |
| 1 | معيار است براي |
| 3 | معيار دارد |
| 3 | مقياس دارد | توصيفي | 7 |
| 37 | ويژگي دارد |
| 3 | شاخص دارد |
| 6 | چالش دارد |
| 122 | مجموع | | |

**جدول1. روابط معنايي استخراج شده همراه با بسامد رخدادهاي هر يك در متن مورد تحليل**

**همان‌طور كه در جدول 1 مشخص است، روابط معنايي استخراج شده در متن مورد تحليل، در 7 مقولة كلي قابل تقسيم است. اگر دقت كنيد، روابطي چون هم ارزي، سلسه مراتبي و همبستگي، همان‌گونه كه در اصطلاحنامه‌ها مطرح بود، استخراج شده است. علاوه بر آنها، روابطي كه نشانگر تأثيرگذاري/تأثيرپذيري، توصيف و كاربرد بودند نيز شناسايي شد. بدين ترتيب، در پاسخ به نخستين پرسش پژوهش بايد گفت، روابط معنايي را مي توان با استفاده از تحليل متون علمي به زبان فارسي استخراج نمود. استخراج روابط به اين روش، روابط ضمني را نيز شامل مي‌شود؛ روابطي كه در رويكرد دستي قابل شناسايي هستند.**

**اما در پاسخ به دومين پرسش و تعيين ميزان پيدايي روابط معنايي، ميزان تطابق واژه‌هاي نشانگر رابطه با رابطه استخراج شده، بررسي گرديد. در پاسخ به اين پرسش، نسبت روابط ضمني و عيني مشخص مي شود. نمودار 1 حاصل تحليل ميزان تطابق است. چنانكه در نمودار مشخص است، 51% روابط به صورت كاملاً تلويحي (ضمني) در متن موجود بوده است و 38% حاصل تطابق نسبي است، به طوري كه بخشي از رابطه در متن آمده است و تنها 11% تطابق كامل وجود داشته است. بدين ترتيب، ميزان پيدايي روابط معنايي در سطح پاييني ارزيابي مي‌شود.**

|  |
| --- |
|  |

****

**نمودار1. ميزان پيدايي روابط معنايي استخراج شده**

**جمع بندي و سخن پاياني**

**هم اكنون حجم وسيعي از اطلاعات به صورت ديجيتالي توليد و در چرخة توليد و بازتوليد قرار مي‌گيرد. پيكرة عظيم اطلاعات متني موجود، بويژه در بستر وب، فرصت مناسبي است براي گسترش مطالعات حوزة هوش مصنوعي. ساخت و توسعه ابزارهاي بازنمون دانش كه با هدف سازماندهي اطلاعات و دانش انجام مي‌پذيرد، تمايلي به سمت و سوي شيوه‌هاي خودكار دارد. ساخت و توسعة خودكار هستي‌شناسي‌ها كه مجموعه‌اي از مفاهيم و روابط معنايي‌اند، به استخراج ـ مفاهيم و روابط ـ وابسته است.**

**در اين پژوهش، پيش فرض روشهاي مبتني بر الگو در استخراج خودكار روابط معنايي در مورد متوني به زبان فارسي بررسي و به دو پرسش طرح شده پاسخ داده شد. نخستين پرسش كه بر امكان استخراج روابط معنايي متمركز بود، به پاسخي مثبت انجاميد. بدين منظور، استخراج روابط معنايي با رويكرد تحليل متن انجام شد و هفت رابطة كلي ـ هم ارزي، تفاوت، همبستگي، سلسه مراتبي، توصيفي، كاربردي، و تأثيري ـ و 24 رابطة جزئي‌تر شناسايي گرديد. اين روابط به دو صورت ضمني و عيني، در متن مورد تحليل آمده است.**

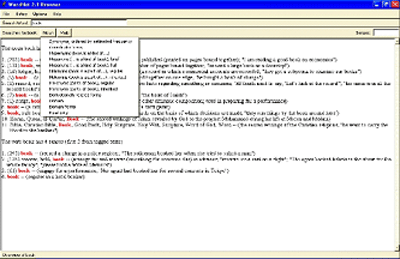
**اما در پاسخ به دومين پرسش پژوهش، ميزان پيدايي روابط معنايي و نسبت روابط عيني به ضمني طبق روشي كه توضيح داده شد، بررسي گرديد. با توجه به درصد يافته‌ها، مشخص است كه در متن مورد تحليل، نسبت روابط عيني به ضمني اندك است، به طوري كه تقريباً نيمي از روابط كاملاً ضمني‌اند و در متن نيامده است و از 49% باقي‌مانده تنها 11% روابط به طور كاملاً عيني در متن آمده است. اين يافته مشخص مي‌سازد در روش خودكار بخشي از دانش ـ كه در اين مورد (51%) قابل توجه نيز هست ـ از دست مي‌رود.**

**به نظر مي‌رسد ردگيري الگوها به تنهايي با توجه به ميزان پيدايي اندك روابط معنايي در متون، راهگشاي مسئله خودكارسازي نيست. استفاده از شبكه هاي واژگاني بخشي از مشكل را حل كند؛ زيرا شبكه‌هاي واژگاني، شبكه‌اي از مفاهيم، همراه با روابط معنايي ميان آنهاست كه به شيوه‌اي ساخت‌يافته طراحي شده است و در جايي كه تنها 49% روابط، با تطابق نسبي و كامل در متن آمده‌اند، راه گشاست. در اين صورت، لازم است شبكه واژگاني زبان فارسي با همكاري متخصصان زبان شناسي، زبان و ادبيات فارسي، رايانه و البته كتابداري و اطلاع‌رساني تدوين شود تا بتوان از قابليتهاي آن در اين حوزه‌ها بهره‌برداري كرد.**

**تحليل متن تحت تأثير عوامل ذهني و غير ذهني صورت مي‌گيرد. اگر عوامل غير ذهني را عوامل زباني بدانيم، در استخراج خودكار متن تنها بايد بر اين عوامل تكيه كرد. حال آن كه در اين صورت بخشي از تحليل دچار كاهش شده است. «اسپارك جونز» (1991) معتقد است ذهن انسان در موقعيتها و زمانهاي مختلف به صورت بسيار پيچيده و غيرقابل پيش‌بيني و با توجه به عوامل محسوس و نامحسوس قادر است به پردازش، تحليل و تفسير بپردازد و قضاوت كند (نقل در فتاحي، 1383، ص 20). هنوز تا خودكارسازياستخراج متن فاصله‌اي ناپيموده وجود دارد؛ فاصله‌اي كه شايد پيمودن آن تنها از عهدة ذهن پيچيده انساني برآيد.**

**پي‌افزود1. تصويري از Protégé، ابزار معندسي هستي شناسي**

**اين ابزار در سال 1998 توسط متخصصان انفورماتيك پزشكي** **دانشگاه استنفورد** **طراحي شد و توسعه يافت. مؤسساتي چون آژانس طرحهاي پژوهشي پيشرفته دفاعي****، مؤسسه ملي سرطان****، مؤسسه ملي استاندارد و فناوري****، كتابخانه ملي پزشكي****بنياد ملي علوم از جمله پيشتيبانان اين طرح نرم‌افزار بوده‌اند. در حال حاضر ويرايش (2007)3.3.1 اين نرم‌افزار در سايت رسمي آن** **موجود است. در اين سايت، امكان دانلود نرم‌افزار فراهم آمده است.**

****

**پي افزود2: تصويري از شبكه واژاگاني WordNet**

**اين شبكه واژگاني در آزمايشگاه علوم شناختي** **دانشگاه پرينستون** **طراحي شده است. امكان دانلود نرم‌افزار از سايت رسمي** **آن وجود دارد.**

**منابع**

**ـ شريف، عاطفه (1385). «شناختي از روابط معنايي در هستي‌شناسي وب». اطلاع شناسي، 4(1و2): 65-84**

**ـــــــــــــــــــ (1387، فروردين 27 و 28). مدل سازي مفهومي در طراحي محيط آموزش مجازي وبپايه. مقاله ارائه شده در همايش ملي فناوري آموزشي در عصر اطلاعات و ارتباطات، اهواز.**

**ـ شمس‌فرد، مهرنوش و احمد عبداله‌زاده بارفروش (1381). «استخراج دانش مفهومي از متن با استفاده از الگوهاي زباني و معنايي». تازه‌هاي علوم شناختي، 4(1): 48-66.**

**ـ عاصي، مصطفي؛ علي رضاقلي فاميان و داريوش آقاجاني (1385). «به سوي طراحي شبكه واژگاني صفات زبان فارسي». زبان و زبان‌شناسي، 2(4) : 125- 136.**

**ـ فتاحي، رحمت‌الله (1383). «تحليل عوامل مؤثر بر نسبي بودن ربط در نظامهاي بازيابي اطلاعات». اطلاع‌شناسي، 2(1): 7-22.**

**- Akgo¨bek ,O.; Aydin, Y.S.; O¨ ztemel, E.; Aksoy, M.S (2006). ”A new algorithm for automatic knowledge acquisition in inductive learning”. Knowledge-Based Systems, 19 : 388–395. Retrieved Jan ,20, 2008, From ScienceDirect Database**

**- Casanovas, P. et al (2005). " Iuriservice II Ontology Development". Retrieved Sep ,5, 2007, From** [**http://www.aifb.uni-karlsruhe.de/WBS/dvr/publications/ivrcasanovas.pdf**](http://www.aifb.uni-karlsruhe.de/WBS/dvr/publications/ivrcasanovas.pdf)

**- Cure, O (2003). "Mapping Databases to ontologies to design and maintain data in a semantic web environment". Retrieved Des ,25, 2006, From** [**http://www.iiisci.org/journal/cvs/sci/pdfs/p704935.pdf**](http://www.iiisci.org/journal/cvs/sci/pdfs/p704935.pdf)

**- Hess, C. ; Schlieder, C (2006). "Ontology-based verification of core model conformity in conceptual modeling" . Computers, Environment and Urban Systems, 30: 543- 561. Retrieved May ,5, 2007, , From Elsevier Database.**

**- Mancini, C. ; Shum, S. J. B (2006). Modeling discourse in contested domains: a semiotic and cognitive framework. Intelligent Journal of Human-Computer Studies. 64: 1154-1171. Retrieved Des ,20, 2006, From ScienceDirect Database**

**- Mika, P (2006). "A Methodology for Distributed Knowledge Management Using Ontologies and Peer-to-Peer . In Staab, S , Stuckenschmidt , H. (Eds) Semantic Web and Peer-to-Peer: Decentralized Management and Exchangeof Knowledge and Information. (pp.283-302). Koblenz: Springer.**

**- Raskin, R. G. ; Pan. M. J (2005). "Knowledge representation in the semantic web for Earth and environmental terminology (SWEET)". Computers & Geosciences, 31 : 1119–1125. Retrieved May ,5, 2007, From Elsevier Database**

**- Soergel, D. et all (2004). Reengineering Thesauri for new Applications: the AGROVOC Example" . Journal of Digital Information , 4(4). Retrieved Oct ,13, 2006, From** [**http://jodi.ecs.soton.ac.uk/Articles/v04/i04/Soergel/**](http://jodi.ecs.soton.ac.uk/Articles/v04/i04/Soergel/)

**- Weng, S. et al (2006). " Ontology construction for information classification". Expert Systems with Applications, 31 : 1–12. Retrieved May ,5, 2007, , From Elsevier Database.**

- Yu, C., et al (2002). “Patterns in Unstructured Data: Discovery, Aggregation, and Visualization”. Retrieved Oct ,13, 2004, From: <http://javelina.cet.middlebury.edu/lsa/out/cover_page.htm>

**- Zhang, M. ; Zhou, G. ; Aw, A (2008). “Exploring syntactic structured features over parse trees for relation extraction using kernel methods”. Information Processing and Management , 44 :687–701. Retrieved Jan ,20, 2008, From ScienceDirect Database**

**- Zhou, G. ; Zhang, M (2007). “Extracting relation information from text documents by exploring various types of knowledge”. Information Processing and Management, 43: 969-982. Retrieved Des ,20, 2006, From ScienceDirect Database.**

**1. دانشجوي دكتراي كتابداري واطلاع رساني دانشگاه فردوسي مشهد atefehsharif@gmail.com**

**.Ontology.**

**. Weng et al.**

**.Properties.**

**.Hess & schlieder.**

**.Entity-Relationships (ER).**

**.Extended Entity-Relationships (EER).**

**. Mika.**

**.Single ontology.**

**.Mapping.**

**local**

**1. Berners-Lee.**

**2. Top level.**

**3. Extension.**

**4. Peer.**

**1. توضيحي مختصر به همراه تصويري از اين نرم‌افزار در پي‌افزود 1 آمده است.**

**1. Latent Semantic Indexing (LIS).**

**2. Yu et al.**

**3. Miller et al.**

**4. Zelenko et al.**

**5. Zhou & Zhang.**

**6. lexical knowledge.**

**7. Syntactic parse trees.**

**8. Dependency trees.**

**9. Semantic features.**

**10. Zhang, Zhou, & Aw.**

**11. Syntactic structured features.**

**1. Raskin & Pan.**

**2. Casanovas et al.**

**3. Ethnographic.**

**4. Akgo¨bek, Aydin ,O¨ ztemel , Aksoy.**

**5. Cure.**

**6. Soergel et al.**

**1. Context.**

**1. تصويري از اين ابزار در پي افزود 2 آمده است.**

**2. Syntactic full pars.**

**3. Assadi.**

**4. Finkelstein-Landau, Morin.**

**1.بر پاية‌ گفتگويي كه با دكتر شمس فرد انجام پذيرفت، مشخص شد اين نظام در حال ساخت است اما هنوز به مرحلة استفاده نرسيده است. همچنين، شبكه واژگاني صفات زبان فارسي نيز در دست طراحي است. براي كسب اطلاعات بيشتر به مقالة عاصي، رضاقلي فاميان و آقاجاني با عنوان «به سوي طراحي شبكه واژگاني صفات زبان فارسي» كه اطلاعات كتابشناختي آن در بخش منابع آمده است، مراجعه كنيد.**

**2. چارچوب و محدودهاين تحليلدر بخشروش پژوهشبه روشنيخواهد آمد.**

**3. Hierarchical (Broader and Narrower Terms).**

**4. Equivalent (Used & Used for).**

**5. Related terms.**

**1. How the World is.**

**2. Domain.**

**3. Mancini & Shum.**

**4. Shared.**

**1. Medical Informatics.**

**2. Stanford University.**

**3. Defense Advance Research Projects Agency.**

**1. National Cancer Institute.**

**2. National Institute of Standards and Technology.**

**3. National Library of Medicine.**

**4. National Science Foundation.**

**5.** [**http://protege.stanford.edu**](http://protege.stanford.edu/)

**6. Cognitive Science Laboratory.**

**7. Princeton University.**

**8.** [**http://wordnet.princeton.edu/**](http://wordnet.princeton.edu/)