دانشگاه صنعتی امیرکبیر

دانشکده مهندسی کامپیوتر و فناوری اطلاعات

پایان‌نامه کارشناسی ارشد رشته مهندسی کامپیوتر گرایش نرم‌افزار

بازیابی معنایی اطلاعات با استفاده از بسط مفاهیم حاصل از جستجوی مبتنی بر کلید واژه

نگارش

وحید جلالی

استاد

دکتر محمدرضا مطش بروجردی

۱۳۸۷

بسمه تعالي

دانشگاه صنعتی امیرکبیر
(پلی تکنیک تهران)

فـــرم اطلاعـات
پـايـانـنـــامـه
كـارشنـــاسي- ارشد
و دكـتـــرا

شماره مدرک:

مشخصات دانشجو	نام خانوادگی : جلالی	نـــام : وحید	شماره دانشجویی: 85131049
	دانشکـــده : مهندسی کامپیوتر	رشته : مهندسی نرم‌افزار	گروه : نرم‌افزار
عنوان	بازيابي معنايي اطلاعات با استفاده از بسط مفاهيم حاصل از جستجوي مبتني بر كليد واژه		
Title	Semantic information retrieval using result concepts of a keyword based query		

استاد راهنمـا	نام خانوادگی : مطش بروجردی	درجه و رتبه	استاد راهنمـا	نام خانوادگی :	درجه و رتبه
	نـــام : محمد رضا	استـا ديـا ر		نـــام :	
استاد مشـاور	:نام خانوادگی	درجه و رتبه	استاد مشـاور	:نام خانوادگی	درجه و رتبه
	:نـــام			:نـــام	

دانشنـامه	○ ارشـد ●دکتـرا ○ کارشناسی	سال تحصیلی :87
نوع پـروژه	کاربردی● نظـری○ توسعـه ای●بنیـادی ○	
مشخصات ظاهری	تعداد صفحـات 107 / تصویر ● جدول● نمودار● نقشه○ واژه‌نامه● / تعداد مراجـع 44 / ضمائم ○ :تعداد صفحات	
زبان متن	فارسـی ● انگلیسی ○	چکیده: فارسـی ● انگلیسی ●
یاداشـت		
توصیفگـر		
کلید واژه فارسی	بازخورد ارتباطی مفهومی، بازیابی اطلاعات مبتنی بر هستان‌شناسی، گسترش جستار، معیارهای شباهت معنایی.	
Key word of English	*Concept-based Pseudo Relevance Feedback, Ontology-based Information Retrieval, Query Expansion, Semantic Similarity Measures.*	

تقدیم به

دوستانی که نشناختم.

با تشکر از زحمات و حمایت‌های بی‌دریغ دکتر بروجردی عزیز

چکیده

اگرچه بازیابی اطلاعات به طور سنتی مستندات و جستارها را از نقطه نظر کلیدواژه‌های به کار رفته در آنها مورد بررسی قرار می‌دهد، در نظر گرفتن روابط معنایی آنها در کنار شباهت‌های نحویشان می‌تواند به ایجاد سیستم‌های کاراتر در بازیابی اطلاعات منجر شود. به طور کلی می‌توان گفت وجود نداشتن واژگان مشابه میان دو متن دلیل بر نامربوط بودن آنها به یکدیگر نیست. حالت‌های متعددی ممکن است وجود داشته باشد که در آنها کاربر، نیاز اطلاعاتی خود را با لغات و واژگانی متفاوت از آنچه در مستندات سیستم وجود دارد بیان کند. و یا این که به علت کمبود اطلاعات قادر نباشد برخی کلید واژه‌های مناسب برای جستجوی خود را در اختیار سیستم قرار دهد. در تمامی این شرایط استفاده از روابط مفهومی میان مستندات و جستار کاربر، این امکان را به سیستم خواهد داد تا نتایج دقیق‌تر و کامل‌تری را به کاربران خود بازگرداند. هدف از این پایان نامه بررسی شیوه‌های مختلف به کارگیری دانش موجود در هستان‌شناسی‌ها در امر بازیابی اطلاعات و معرفی ایده‌های نوین در این رابطه می‌باشد. به طور خلاصه در این پایان نامه روش‌های جدیدی در زمینه کاربرد هستان‌شناسی در گسترش‌جستار، بازیابی معنایی مبتنی بر معیارهای شباهت مفهومی و بازیابی معنایی همراه با بازخورد ارتباطی خودکار مفهومی معرفی می‌شوند. دامنه مورد بررسی در این پایان نامه مستندات پزشکی می‌باشند. هستان‌شناسی به کار رفته در روش‌های معرفی شده MeSH یا Medical Subject Headings است که در برگیرنده در حدود ۲٤۰۰۰ مفهوم پرکاربرد در قلمرو پزشکی می‌باشد. همچنین برای مقایسه تاثیر به کاربردن هستان‌شناسی عمومی در مقابل هستان‌شناسی مختص دامنه در بخشی از تجربیات این پایان نامه از هستان‌شناسی Wordnet استفاده شده است. مجموعه آزمون‌های به کار رفته برای ارزیابی روش‌های معرفی شده نیز عبارتند از Medline و OHSUMED که از مجموعه آزمون‌های شناخته شده در ارزیابی سیستم‌های بازیابی اطلاعات می‌باشند.

کلمات کلیدی: بازخورد ارتباطی مفهومی، بازیابی اطلاعات مبتنی بر هستان‌شناسی، گسترش جستار، معیارهای شباهت معنایی.

فهرست

فهرست شکل ها

فهرست جدول ها

فصل اول

مقدمه و تعاریف

۱- مقدمه

در اثر افزایش حجم مستندات مربوط به زمینه‌های مختلف دانش بشری، نحوه ذخیره‌سازی و دسترسی به این مستندات تبدیل به مساله‌ای اساسی در دنیای امروز گشته‌است. چنانچه بتوان روش‌های نوینی ارائه نمود که نیاز روزافزون بشر به اطلاعات را با کیفیت بالاتری سیراب نمایند این روش‌ها از ارزش فوق العاده بالایی برخوردار خواهند بود. در این پایان نامه سعی شده است تا برای حوزه مستندات پزشکی، موتورهای جستجو با الگوریتم‌های معنایی نوینی معرفی شوند تا نیازهای اطلاعاتی کاربران را بهتر درک کرده و در نتیجه نتایج دقیق‌تری را برای جستجوهای کاربران به آنها بازگردانند.

۱-۱- تعاریف و مفاهیم اولیه

۱-۱-۱- مقدمه‌ای بر وب‌معنایی و هستان‌شناسی

از آنجا که بخش اعظم مستندات موجود در شبکه جهانی در حال حاضر در قالب HTML ذخیره شده اند این مستندات بیشتر مورد استفاده انسان قرار می گیرند تا عامل های نرم افزاری، به همین دلیل حیطه عملکرد و قدرت چنین عامل هایی در شبکه جهانی بسیار محدود و ناچیز است.

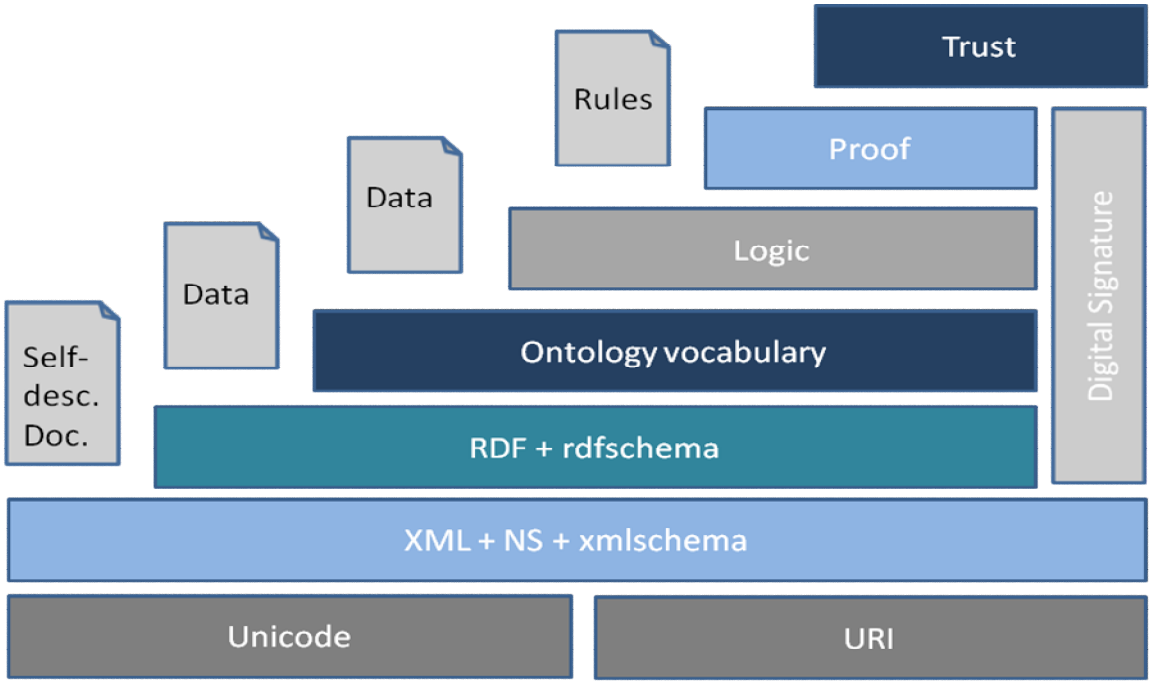
با گسترش روز افزون محتوی شبکه جهانی و تعریف کاربرد های جدید برای آن به نظر می رسد که منابع و مستندات موجود در آن باید به گونه ای قابل فهم و تفسیر توسط ماشین ارائه شوند. علاوه بر این، روش‌های بازیابی اطلاعات در حال حاضر به علت گستردگی موضوعات موجود و حجم بالای مستندات، کارایی نسبتا نامطلوبی دارند. وجود چنین مشکلاتی بود که موجب معرفی و ظهور وب‌معنایی در سال های پایانی هزاره دوم میلادی گردید.

وب‌معنایی بر پایه هستان شناسی بنا شده است و کارایی آن تا حد زیادی وابسته به جامعیت ، قدرت و ساختارمند بودن منابع هستان‌شناسی آن می باشد. از دیدگاه ساختاری وب‌معنایی را می توان به دو بخش موتور استنتاج و منابع هستان‌شناسی آن تقسیم کرد. آن چه در این جا بیشتر طرف توجه ما خواهد بود مسائل پیرامون هستان‌شناسی و موارد موجود در این راستاست.

۱-۱-۲- معرفی اجمالی ساختار وب‌معنایی

همان طور که در شکل ۱-۱ مشاهده می کنید [۴۲] وب‌معنایی بر پایه لایه های گوناگونی استوار است که یکی بر روی دیگری بنا شده است. هرچه از لایه های پایین این ساختار به سمت بالا حرکت کنیم حجم کار های انجام شده در آن زمینه کاهش یافته و مسائل حل نشده بیشتری باقی می ماند.

در سال‌های اخیر برای معرفی منابع موجود در وب استانداردهای فراوانی معرفی شده است. به عنوان نمونه می‌توان به استانداردهایی از قبیل RDF(S)، OIL، DAML و OWL اشاره کرد. استانداردی که در حال حاضر از سوی W۳C پیشنهاد می‌شود OWL است. OWL بر پایه RDF(S) بنا شده و از نوشتار XML سود می‌برد، ولی نسبت به RDF(S) امکانات بیشتری برای توصیف منابع دارد. OWL دارای سه نوع Lite، DL و Full می‌باشد. هر یک از این سه نوع دارای شرایط ویژه و دامنه استفاده مشخصی هستند. شایان ذکر است که هستان‌شناسی به کار رفته در این پژوهش به صورت استاندارد در قالب XML وجود دارد و چالش‌های مربوط به زبان‌های سطح بالاتر معرفی شده، مانند RDF(S) و یا OWL در مورد آن مطرح نمی‌شود، لیکن به دلیل معرفی حالت کلی وب‌معنایی در این بخش اشاره‌ای مختصر به ساختار کلی آن بی‌فایده نیست.



شکل ۱-۱- معماری وب‌معنایی

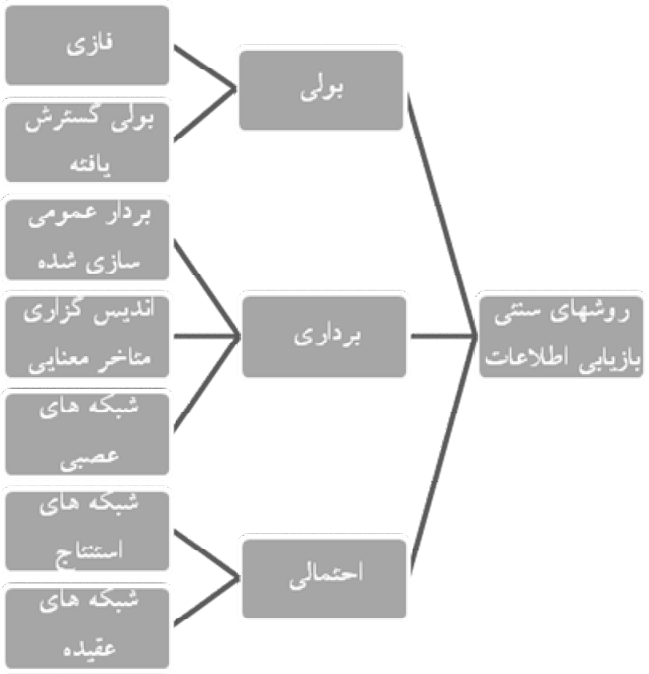
۱-۱-۳- مقدمه‌ای بر بازیابی اطلاعات

روش‌های سنتی بازیابی اطلاعات از طریق تطبیق کلیدواژه‌ها با یکدیگر مستندات را بازیابی می‌کنند. این نوع نگرش نسبت به بازیابی اطلاعات متضمن این مطلب است که معنای مستندات، توسط کلیدواژه‌های به کار رفته در آنها تعیین می‌شود. هرچند که چنین نگرشی به وضوح ساده سازی مساله

بازیابی اطلاعات است، کاربردهای بسیاری بر پایه همین دید در زمینه بازیابی اطلاعات شکل گرفته و طی سالیان متمادی مورد استفاده قرار گرفته‌اند. این ساده سازی در واقع به این علت رخ می‌دهد که جایگزینی یک مستند با تعدادی از کلیدواژه‌های به کار رفته در آن همواره باعث از دست دادن قسمتی از معنای مستند خواهد شد که توسط کلیدواژه‌های انتخاب شده قابل پوشش نمی‌باشد. در واقع به همین علت است که در اکثر موتورهای جستجوی سنتی همواره مقدار قابل توجهی از مستندات نامربوط در ازای جستار کاربر به او بازگردانده می‌شود و یا برخی از مستنداتی که به جستار کاربر مربوط هستند به طور کلی به عنوان مستندات مربوط، تشخیص داده نشده و به کاربر بازگردانده نمی‌شوند. در ادامه روش‌های اصلی بازیابی سنتی اطلاعات را مرور می‌کنیم.

سه روش اصلی بازیابی سنتی اطلاعات عبارتند از روش بولی، برداری و احتمالی. در روش بولی مستندات و جستارها به عنوان مجموعه‌هایی از کلیدواژه‌ها در نظر گرفته می‌شوند. بنابر این مدل بولی یک مدل از نوع تئوری مجموعه‌ها محسوب می‌شود. در روش برداری، مستندات و جستارها به عنوان بردارهایی در فضای t بعدی در نظر گرفته می‌شوند که بعدهای این بردارها را کلیدواژه‌ها تعیین می‌نمایند. به همین دلیل روش برداری را از جمله روش‌های جبری در نظر می‌گیرند. در روش احتمالی نیز چارچوب مدل کردن مستندات و جستارها، یک چارچوب مبتنی بر احتمال است و به همین دلیل همان طور که از نام آن بر می‌آید یک روش احتمالی محسوب می‌شود.

با گذشت زمان مدل‌های جایگزینی برای هریک از این سه روش سنتی معرفی شده است. بر طبق مدل‌های جایگزین مجموعه‌ای، مدل‌های فازی و گسترش یافته بولی معرفی شده‌اند. همچنین روش‌های برداری عمومی‌سازی شده ، اندیس‌گذاری متاخر معنایی و مدل‌های شبکه عصبی بر اساس مدل‌های جایگزین جبری ارائه شده‌اند. و در پایان بر اساس مدل‌های جایگزین احتمالی، مدل‌های شبکه‌های استنتاج و شبکه‌های اعتقاد معرفی می‌شوند. در شکل **۱-۲** می‌توانید یک دسته‌بندی از این مدل‌های بازیابی اطلاعات را مشاهده نمایید [۴۳].



شکل ۱-۲- دسته بندی روش‌های سنتی بازیابی اطلاعات و مدل‌های گسترش یافته آنها

۱-۱-٤- مشکلات روش‌های سنتی بازیابی اطلاعات

از آن‌جا که در روش‌های سنتی بازیابی اطلاعات، مستندات و جستارها با کلیدواژه‌های موجود در آن‌ها نمایش داده می‌شوند، حالت‌های متعددی به وجود می‌آید که جستار و مستند راجع به یک مفهوم خاص صحبت می‌کنند ولی به علت به کار بردن واژه‌های متفاوت برای اشاره به یک مفهوم، توسط موتورهای بازیابی سنتی اطلاعات، نامربوط تشخیص داده می‌شوند. همچنین موارد بسیاری وجود دارد که کاربر به علت عدم آگاهی از فرهنگ واژگان به کار رفته در یک حوزه، قادر نخواهد بود نیاز اطلاعاتی خود را به دقت بیان کند. برای حل چنین مشکلاتی بود که روش‌های بازیابی معنایی اطلاعات در مقابل روش‌های سنتی مطرح شدند.

جستجوی مفهومی به معنای جستجو بر اساس معنا به جای رشته‌ای از حروف، انگیزه حجم زیادی از تحقیقات و مطالعات در زمینه بازیابی اطلاعات بوده است. این دید نسبت به بازیابی اطلاعات در زمینه‌های شناخته شده‌ای مانند اندیس‌گزاری متاخر معنایی، مفهومی سازی زبانی و استفاده از فرهنگ‌های جامع و رده‌بندی‌ها دیده می‌شود.

۱-۲- بازیابی معنایی اطلاعات

بازیابی معنایی اطلاعات در برابر بازیابی سنتی اطلاعات مطرح می‌شود و هدف از آن، از میان برداشتن محدودیت‌ها و مشکلاتی است که در بازیابی سنتی اطلاعات وجود دارد. یکی از روش‌های به کار رفته در بازیابی اطلاعات استفاده از هستان‌شناسی برای بهبود دقت موتورهای جستجوست. استفاده از هستان‌شناسی در این راستا می‌تواند به دو صورت کلی انجام پذیرد. روش اول استفاده از هستان‌شناسی برای افزودن کلید واژه‌های مناسب به جستار اولیه کاربر است و روش دوم معرفی معیاری از شباهت معنایی میان مفاهیم برای استفاده در بازیابی اطلاعات در کنار شباهت‌های مبتنی بر کلیدواژه‌هاست.

۱-۳- معیارهای شباهت معنایی

برای تعیین شباهت‌های معنایی میان عبارات مختلف، روش‌های مختلفی وجود دارد که به چهار دسته زیر تقسیم می‌شوند:

- روش‌های شمارش یال: شباهت میان دو عبارت یا مفهوم را به عنوان تابعی از مسیر اتصال دهنده آن دو و موقعیت مفاهیم در یک رده بندی محاسبه می‌کند.
- روش‌های محتوای اطلاعات: تفاوت محتوای اطلاعات دو مفهوم یا عبارت را به عنوان تابعی از احتمال حضور آنها در دسته‌متن در نظر می‌گیرد. این احتمال حضور می‌تواند وابسته به دسته‌متن یا مستقل از آن باشد. به عنوان مثال چنان چه از نظر آماری تکرار یک عبارت در دسته‌متن مورد بررسی قرار بگیرد و احتمال بر پایه آن محاسبه شود، معیار حاصل وابسته به دسته‌متن خواهد بود و چنانچه این احتمال با توجه به موقعیت مفاهیم در یک سلسله مراتب که خود مستقل از دسته متن ایجاد شده است محاسبه گردد، معیار حاصل مستقل از دسته‌متن خواهد بود.
- روش‌های مبتنی بر ویژگی: شباهت دو عبارت را به عنوان تابعی از ویژگی‌های آنها (به عنوان مثال تعریف یا Glosses در Wordnet و یا scope noteها در MeSH) و یا مبتنی بر رابطه‌اشان با دیگر عبارات مشابه در رده‌بندی در نظر می‌گیرد. در این حالت تعداد بیشتر ویژگی‌های یکسان، شباهت معنایی را افزایش داده و برعکس تعداد ویژگی‌های غیریکسان، شباهت معنایی را کاهش می‌دهد.

- روش‌های ترکیبی: این روش‌ها ایده‌های موجود در روش‌های بالا را با یکدیگر ترکیب می‌کنند. به عنوان مثال شباهت عبارات را بر مبنای تطابق مترادف‌ها، عبارات همسایه و ویژگی‌های عبارات محاسبه می‌نمایند.

از منظری دیگر روش‌های تشخیص شباهت معنایی به دو دسته تقسیم می‌شوند:

- دسته اول روش‌هایی هستند که مفاهیم مورد نظر در یک هستان‌شناسی را با یکدیگر مقایسه می‌کنند.
- دسته دوم روش‌هایی هستند که عبارات مختلف از دو هستان‌شناسی متفاوت را با یکدیگر مقایسه می‌کنند.

یک نکته مهم در مورد معیارهای شباهت‌های معنایی مختلف این است که اکثر آن‌ها بالاترین شباهت را به مفاهیمی اختصاص می‌دهند که اولا در هستان‌شناسی مربوطه، به یکدیگر نزدیک هستند و ثانیا نسبت به مفاهیمی با فاصله مشابه، در قسمت عمیق‌تری از سلسله مراتب واقع شده باشند (عبارات خاص‌تری باشند) [۳۸].

روش‌های مبتنی بر شمارش یال و محتوای اطلاعات از طریق استفاده از ساختارهای اطلاعاتی مانند موقعیت قرارگیری عبارات عمل می‌کنند و روش‌های مبتنی بر محتوای اطلاعات بیشتر برای عباراتی از یک هستان‌شناسی یکسان مناسب هستند. علت این مطلب آن است که ساختار و محتوای اطلاعات هستان‌شناسی‌های مختلف به طور مستقیم قابل مقایسه نیستند. از طرف دیگر روش‌هایی که از دو هستان‌شناسی برای معرفی معیار شباهت معنایی سود می‌برند معمولا از روش‌های ترکیبی و یا مبتنی بر ویژگی استفاده می‌کنند.

شما می‌توانید در جدول ۱-۱ اطلاعاتی راجع به چند روش شناخته شده محاسبه شباهت معنایی را مشاهده کنید. در این جدول اطلاعاتی از قبیل تاثیر موقعیت مفاهیم در هستان‌شناسی، تاثیر ویژگی‌های یکسان بر افزایش یا کاهش شباهت، متقارن بودن معیار معرفی شده و نرمال بودن آن ذکر شده‌اند.

جدول ۱-۱- معیارهای شباهت معنایی و خصوصیات آنها

روش	نوع روش	افزایش با اشتراک	کاهش با تفاوت	تقارنی	نرمال شده در [۱و۰]	موقعیت در سلسله مراتب
Rada [۱]	شمارش یال	آری	آری	آری	آری	**خیر**
Wu [۲]	شمارش یال	آری	آری	آری	آری	**آری**
Li [۳]	شمارش یال	آری	آری	آری	آری	**آری**
Leacock [٤]	شمارش یال	خیر	آری	آری	خیر	**آری**
Richardson [٥]	شمارش یال	آری	آری	آری	آری	**آری**
Resnik [٦]	محتوای اطلاعات	آری	خیر	آری	خیر	**آری**
Lin [۷]	محتوای اطلاعات	آری	آری	آری	آری	**آری**
Lord [۸]	محتوالی اطلاعات	آری	خیر	آری	آری	**آری**
Jiang [۹]	محتوای اطلاعات	آری	آری	آری	خیر	**آری**
Tversky [۱۰]	ویژگی	آری	آری	خیر	آری	**خیر**
Rodriguez [۱۱]	ترکیبی	آری	آری	خیر	آری	**خیر**

در ادامه توضیحی مختصر راجع به هریک از این روش‌ها ارائه می‌شود.

- معیار Rada: یکی از قدیمی‌ترین معیارهای شباهت معنایی است که از روی فاصله بین دو مفهوم در یک سلسله مراتب میزان شباهت آنها به یکدیگر را محاسبه می‌کند.
- معیار Wu: Wu و Palmer معیار شباهت معنایی خود را برای دو مفهوم c۱ و c۲ به صورت زیر تعریف می‌کنند:

$$sim_{Wu\&Palmer}(c_1,c_2)=\frac{2\times N_3}{N_1+N_2+2\times N_3} \qquad (۱-۱)$$

که در آن N۱ تعداد گره‌های موجود در مسیر c۱ به خاص‌ترین مفهومی است که ابر مفهوم c۱ و c۲ است. N۲ نیز تعداد گره‌های موجود در مسیر c۲ به خاص‌ترین مفهومی است که ابر مفهوم c۱ و c۲ است و همچنین N۳ تعداد گره‌های موجود در مسیر خاص‌ترین ابر مفهوم c۱ و c۲ به ریشه هستان‌شناسی می‌باشد.

- معیار Li: این معیار از طریق رابطه‌ای که در آن فاصله دو مفهوم در هستان‌شناسی و عمق خاص‌ترین ابر مفهوم مشترکشان را در نظر می‌گیرد به صورت زیر محاسبه می‌شود:

$$sim_{Li}(c_1,c_2) = e^{-\alpha L} \times \frac{e^{\beta H} - e^{-\beta H}}{e^{\beta H} + e^{-\beta H}} \qquad (۲-۱)$$

در این رابطه α و β پارامترهای غیر منفی هستند که نحوه تعامل طول مسیر میان دو مفهوم و عمق گره مشترک در برگیرنده آنها را مشخص می‌کند.L نشان‌دهنده طول مسیر میان دو مفهوم c۱ و c۲ است و H نیز عمق خاص‌ترین ابر مفهوم مشترک c۱ و c۲ را نشان می‌دهد.

- معیار Leacock: در این معیار که در رابطه (۱-۳) معرفی شده از سلسله مراتب IS-A و کوتاه‌ترین مسیر میان نام‌ها در این سلسله مراتب به عنوان معیاری برای محاسبه شباهت معنایی آنها استفاده می‌شود.

$$sim_{Leacock}(c_1,c_2) = -\log(\frac{N_p(c_1,c_2)}{2D}) \qquad (۳-۱)$$

در این رابطه، NP(c۱,c۲) نشان دهنده کوتاه‌ترین مسیر بر حسب گره بین مجموعه ترادف‌های c۱ و c۲ است. و D بیشینه عمق رده‌بندی را نشان می‌دهد.

- معیار Richardson: در این معیار شباهت معنایی به صورت شباهت کلاسی بیان می‌شود. و به طور کلی شباهت بین دو کلاس از روی محتوای اطلاعات اولین (خاص‌ترین) کلاسی که هردوی آنها را در بر می‌گیرید تخمین زده می‌شود. همچنین محتوای اطلاعات یک کلاس نیز از روی احتمال حضور آن در یک دسته‌متن بزرگ تخمین زده می‌شود. در رابطه (۱-٤) شما می‌توانید نحوه محاسبه معیار Richardson را مشاهده کنید:

$$sim_{Richardson}(c_1, c_2) = \max_{ci}[\log \frac{1}{P(c_i)}] \quad (٤-١)$$

که در آن Ci مجموعه کلاس‌هایی است که هر دو کلاس c۱ و c۲ را در بر می‌گیرند و P(ci) کلاس احتمال کلاس ci است و همچنین log ۱/p(ci) محتوای اطلاعات کلاس ci را نشان می‌دهد.

- معیار Resnik: یک معیار شباهت معنایی برای سلسله مراتبی از نوع IS-A است که بر اساس محتوای اشتراکی اطلاعات معرفی می‌شود. در رابطه زیر شما می‌توانید نحوه محاسبه این معیار را مشاهده کنید:

$$sim_{Resnik}(c_1, c_2) = \max_{c \in S(c_1, c_2)} [-\log p(c)] \quad (٥-١)$$

که در آن S(c۱,c۲) مجموعه کلاس‌های در برگیرنده c۱ و c۲ هستند و p(c) نیز بیانگر احتمال برخورد با نمونه‌ای از مفهوم c است.

- معیار Lin: در این معیار شباهت بین دو نمونه از مفاهیم بر حسب احتمال حضور آن مفاهیم در کل دسته متن و از روی رابطه زیر محاسبه می‌شود:

$$sim_{Lin}(x_1, x_2) = \frac{2 \times \log P(c_0)}{\log P(c_1) + \log P(c_2)} \quad (٦-١)$$

در این رابطه p(c۰)، p(c۱) و p(c۲) از روی آمارهای مربوط به این مفاهیم در یک دسته‌متن برچسب‌گذاری شده بدست می‌آیند. به این صورت که تعداد دفعات تکرار این مفاهیم در دسته‌متن محاسبه شده و با توجه به کل مفاهیم موجود در متن نرمالایز می‌شوند. بدیهی است که در این رابطه c۱ و c۲ مفاهیم مورد نظر و c۰ خاص‌ترین مفهوم در برگیرنده c۱ و c۲ است.

- معیار Lord: این معیار از روی رابطه زیر محاسبه می‌شود:

$$sim_{Lord}(c_1, c_2) = -\ln Pms(c_1, c_2) \quad (٧-١)$$