آکادمی داده
dataacademy.ir

بکارگیری تکنیکی از داده‌کاوی برای بهبود کارآیی سیستم‌های استخراج اطلاعات از متن

احمد عبداله زاده[۱] مجتبی شکری[۲]

۱ تهران- خ سمیه - دانشگاه صنعتی امیرکبیر – دانشکده کامپیوتر و فناوری اطلاعات
۲ تهران – خ کارگر شمالی – ساختمان شماره ۲۴۰– طبقه پنجم – شرکت راهبر
ahmad@ce.aut.ac.ir

چکیده:

با رشد سریع حجم متون قابل دسترس به‌خصوص در شبکه‌ی جهانی اینترنت کار استخراج اطلاعات از متن اهمیت ویژه‌ای یافته است. خروجی سیستم استخراج اطلاعات، پایگاه داده‌ای از اطلاعات موجود در متن می‌باشد. از طرفی تکنیک‌های داده‌کاوی این امکان را به ما می‌دهند که دانشی را به شکل مجموعه‌ای از قواعد از محتویات پایگاه داده و ارتباط بین آن‌ها بدست آوریم. در این مقاله از تکنیک تولید قواعد تداعی‌گر در داده‌کاوی استفاده می‌کنیم. بدین منظور با استفاده از الگوریتم استاندارد C4.5RULES و اعمال آن بر روی خروجی سیستم استخراج اطلاعات، دانشی(مجموعه‌ی قواعد پیشگو) بدست می‌آوریم که از آن برای بالاتر بردن احتمال درستی اطلاعات استخراج شده، پیشگویی اطلاعات استخراج نشده و مقایسه بین قواعد استخراج اطلاعات استفاده می‌کنیم. در ادامه، طرحی برای معماری سیستم‌های استخراج اطلاعات پیشنهاد می‌کنیم تا بتوانند از این تکنیک برای بهبود کارآیی‌شان استفاده کنند. سپس نتایج پیاده‌سازی این تکنیک را بر روی سیستم WHISK (یکی از سیستم‌های باز استخراج اطلاعات موجود) مورد ارزیابی قرار می‌دهیم.

واژه‌های کلیدی: استخراج اطلاعات از متن - داده‌کاوی - وب‌کاوی - استخراج دانش از پایگاه داده - معماری سیستم‌های استخراج اطلاعات از متن

۱- مقدمه

زمانیکه بخواهیم اطلاعات خاصی را از انبوه متون در دسترس استخراج کنیم، اهمیت سیستم‌هایی که بتوانند متون را به صورت خودکار پردازش کنند، بیشتر مشخص خواهد شد. در وهله‌ی اول باید بتوانیم از انبوه مستندات موجود، مستنداتی را انتخاب کنیم که مربوط به زمینه‌ی مورد نظر ما باشد. این وظیفه به عهده‌ی سیستم‌های بازیابی اطلاعات است. در مرحله‌ی دوم باید در این مستندات جستجو کرده و اطلاعات مورد نیازمان را جمع‌آوری کنیم. اینجاست که سیستم‌های استخراج اطلاعات به کمک‌مان می‌آیند. استخراج اطلاعات از متن شامل ارائه‌ی یک قالب ساختارمند (مانند یک پایگاه داده) از اطلاعات دلخواه موجود در متن می‌باشد[۱]. دو معیار مهم برای ارزیابی کارآیی سیستم‌های استخراج اطلاعات وجود دارد. اول اینکه، چه درصدی از اطلاعات استخراجی صحیح هستند و دوم اینکه، چه درصدی از اطلاعات موجود در متن استخراج یافته‌اند[۲].

---

[۱] دانشیار، هیئت علمی دانشکده کامپیوتر دانشگاه صنعتی امیرکبیر

[۲] کارشناسی ارشد نرم افزار دانشگاه صنعتی امیرکبیر، هیئت علمی دانشگاه آزاد واحد دماوند

بکارگیری تکنیک‌های دیگری که در هوش مصنوعی وجود دارند، می‌تواند منجر به بهبود معیارهای کارآیی سیستم‌های استخراج اطلاعات شود[۳]. موضوع مقاله مزبور استفاده از تکنیک تولید قواعد تداعی‌گر در داده‌کاوی برای کاراتر کردن سیستم‌های استخراج اطلاعات است. در داده‌کاوی از روش‌های آماری یا یادگیری ماشین برای کشف روابط تازه در پایگاه‌های داده‌ی رابطه‌ای بهره برده می‌شود. در سیستم استخراج اطلاعات از مستندات زبان طبیعی، اطلاعات دلخواه استخراج می‌شود. به عبارت دیگر از متن بدون ساختار، اطلاعات ساختارمند مانند یک پایگاه داده استخراج می‌شود. پیوند داده‌کاوی و سیستم استخراج اطلاعات زمانی است که خروجی سیستم استخراج اطلاعات را به عنوان ورودی برای داده‌کاوی در نظر بگیریم و با استفاده از تکنیک‌های موجود در داده‌کاوی، روابط تازه‌ای در اطلاعات استخراجی کشف کنیم. حاصل ادغام استخراج اطلاعات از متن و داده‌کاوی، استخراج دانش از متن خواهد بود. حال دانشی در اختیار ماست که هم خود یک خروجی باارزش است و هم می‌توان از آن برای منظورهای دیگر مانند بالا بردن کارآیی سیستم‌های استخراج اطلاعات استفاده کرد. ما از تکنیک‌های موجود در داده‌کاوی از روش C4.5RULES [۴] برای تولید قواعد تداعی‌گر استفاده خواهیم کرد که یک روش استقرای قاعده با استفاده از ساخت درخت تصمیم از روی محتویات پایگاه‌های داده می‌باشد. سیستم WHISK [۱] که یکی از سیستم باز استخراج اطلاعات از متن می‌باشد را برای اضافه کردن تکنیک مورد نظر انتخاب کرده‌ایم و نتایج حاصله را مورد ارزیابی و تجزیه وتحلیل قرار داده‌ایم.

۲- پیش‌زمینه

در این بخش در ابتدا درباره‌ی سیستم‌های استخراج اطلاعات بطور کلی و سیستم WHISK بطور خاص صحبت می‌کنیم و سپس درباره‌ی داده‌کاوی و روش استقرای قاعده C4.5RULES صحبت خواهیم کرد.

۲-۱ سیستم‌های استخراج اطلاعات

وظیفه سیستم استخراج اطلاعات، ارائه اطلاعات دلخواه، بصورت ساختارمند از اطلاعات موجود در متن است. نوع متن ورودی به سیستم‌های استخراج اطلاعات در سه دسته‌ی ساختارمند، شبه ساختارمند و بدون ساختار تقسیم‌بندی می شوند[۵]. متن ساختارمند از یک قالب خاص پیروی می‌کند و از یک واژگان محدود استفاده می‌کند مانند گزارشات آزمایشگاهی. متن شبه ساختارمند دارای قالب کاملا مشخصی نمی‌باشد ولی مانند متن بدون ساختار نیز نیست و اغلب بصورت اطلاعات غیر گرامری و تلگرافی می‌باشد. متن بدون ساختار شامل جملات زبان طبیعی مانند داستان‌ها می‌باشد. سیستم استخراج اطلاعات برای استخراج اطلاعات از متن احتیاج به دانشی دارد که اغلب به صورت مجموعه قواعد استخراج اطلاعات می‌باشد که الگویی را در متن جستجو می‌کند و خروجی را بر اساس آن بدست می‌آورد[۱و۶]. در سال‌های اخیر تمرکز تحقیقات بر روی استفاده از روشهای آماری و یادگیری ماشین برای بدست آوردن مجموعه قواعد استخراج اطلاعات به صورت خودکار بوده‌است. سیستم‌های استخراج اطلاعات را می‌توان به عنوان پس‌پردازش بر روی خروجی بدست آمده از سیستم‌های بازیابی اطلاعات (مانند موتورهای جستجو) استفاده کرد و اطلاعات دقیق‌تری بدست آورد[۸].

زمانی یک سیستم استخراج اطلاعات را موفق می‌نامیم که به این اهداف رسیده باشد[۷]:

۱- جملاتی از متن را بیابد که حاوی اطلاعات مرتبط باشد.

۲- اطلاعات مرتبط با زمینه‌ی مورد علاقه را استخراج کند.

۳- اطلاعات را به هم پیوند دهد و در یک قالب از پیش تعیین شده در خروجی ارائه کند.

۴- از اطلاعات غیر مرتبط صرف نظر کند (بسیار مهم).

۲-۲ سیستم WHISK

در سال‌های اخیر چندین سیستم استخراج اطلاعات ساخته شده است که اغلب از روش‌های یادگیری ماشین بهره برده‌اند. از جمله مهمترین آنها می‌توان به سیستم‌های AutoSlog در سال ۹۳، CRYSTAL در سال ۹۵، WIEN در سال ۹۵،

RAPIER در سال ۹۷، SRV در سال ۹۸، SoftMealy در سال ۹۸، WHISK در سال ۹۸ و STALKER در سال ۹۹ اشاره کرد[۸].

در سیستم WHISK برای بدست آوردن مجموعه قواعد استخراج اطلاعات از یادگیری با نظارت استفاده شده است. قواعد استخراج اطلاعات در WHISK بر پایه‌ی عبارات باقاعده می‌باشد که در آن الگویی ارائه می‌شود که باید در متن جستجو شده و در صورت تطابق آن با یک قسمت از متن، خروجی لازم بدست می‌آید. برای نمونه در شکل ۱ قاعده‌ای برای استخراج اطلاعات از آگهی‌های اجاره خانه آمده است. Pattern، الگویی را نشان می‌دهد که باید در متن جستجو شود و Output، خروجی مربوطه را نشان می‌دهد[۱].

Pattern:: * (Nghbr) * (Digit) 'Bdrm' * '$' (Number)

Output:: Rental {Neighborhood $1} {Bedrooms $2} {Price $3}

شکل۱: نمونه‌ای از قاعده‌ی استخراج اطلاعات از متن در WHISK

پارامترهای ارائه شده در خروجی از عبارات مشخص شده در الگو بدست می‌آید. بطور مثال $1 به (Nghbr) و $3 به (Number) اشاره دارد که به ترتیب به نام محله و اجاره‌بها اشاره می‌کنند. همانطور که دیده می‌شود مجموعه‌ی خروجی می‌تواند پایگاه داده‌ای از اطلاعات موجود در متن بوجود آورد.

روش تولید قواعد در WHISK مانند روشی است که در Wrapper Induction [۹] استفاده می‌شود. به‌اضافه‌ی اینکه از اطلاعات لغوی، گفتاری و نحوی موجود در متن برای قوی‌تر کردن قواعد استفاده می‌شود. سیستم WHISK، یادگیری قواعد را با تعامل با کاربر انجام می‌دهد تا مثال‌هایی را برای مشخص کردن خروجی به کاربر ارائه کند که برای تصمیم‌گیری بعدی مفیدتر باشد و منجر به مجموعه قواعد استخراج کاراتر شود. برای تولید خروجی، قواعد بدست آمده بر اساس معیاری با هم مقایسه می‌شوند و قاعده‌ی مناسب‌تر انتخاب می‌شود. معیاری که در WHISK استفاده شده است میزان پوشش صحیحی است که قواعد بر روی مجموعه‌ی آموزشی دارند یعنی نسبت تعداد استخراج‌های درست بر کل استخراج‌هایی که بر روی مجموعه‌ی آموزشی داشته‌اند.

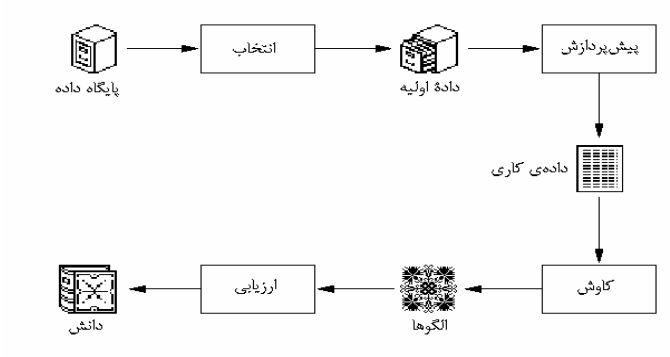
۲-۳ داده‌کاوی

هدف اصلی در داده‌کاوی کشف دانش شناخته نشده یا ضمنی در مجموعه‌های داده‌ای است[۱۰]. در داده‌کاوی یا کشف دانش از پایگاه داده (KDD)[۳] فرض می‌شود که داده در قالب پایگاه‌های داده رابطه‌ای می‌باشد. فرآیند عمومی در داده‌کاوی یا کشف دانش از پایگاه داده، در شکل ۲ آورده شده است[۱۱].

تکنیک‌های مورد استفاده در داده‌کاوی در سه بخش تقسیم‌بندی می‌شوند[۱۰]: دسته‌بندی، خوشه‌بندی و ساختن قواعد تداعی‌گر.

هدف در خوشه‌بندی پیشگویی دسته‌ی یک شیء جدید بر اساس مثال‌های آموزشی می‌باشد. بیشترین توجه‌ها در این قسمت بر روی الگوریتم‌های سازنده‌ی درخت‌های تصمیم و شبکه‌های عصبی می‌باشد. در خوشه‌بندی، مجموعه‌ای از اشیاء بر اساس شباهت مقادیر ویژگی‌هایشان خوشه‌بندی می‌شوند. یکی از کاربردهای مهم خوشه‌بندی ساخت رده‌بندی[۴] می‌باشد. در خوشه‌بندی اغلب از الگوریتم‌های خوشه‌بندی سلسله مراتبی جمع‌شونده[۵] استفاده می‌شود.

[3] Knowledge Discovery form Database
[4] Taxonomy
[5] Hierarchical Agglomerative Clustering



شکل۲: فرآیند داده‌کاوی یا کشف دانش از پایگاه داده

در ساختن قواعد تداعی‌گر یا پیشگو فرم قواعد بصورت زیر می‌باشد:

$$A_1 \wedge A_2 \wedge \ldots \wedge A_m \longrightarrow B_1 \wedge B_2 \wedge \ldots \wedge B_n$$

هر یک از A_i ها و B_i ها مقادیر ویژگی‌ها یا ستون‌های یک جدول می‌باشد. این قاعده نشان می‌دهد که اگر رابطه‌ی $A_1 \wedge A_2 \wedge \ldots \wedge A_m$ میان مقادیر ستون‌های جدول برقرار باشد، می‌توان نتیجه گرفت که رابطه‌ی $B_1 \wedge B_2 \wedge \ldots \wedge B_n$ نیز برقرار است. الگوریتم‌های مختلفی برای ساخت این قواعد وجود دارد که از جمله می‌توان به A-Priori، RIPPER و C4.5RULES اشاره کرد[۱۲].

در این مقاله برای کاراتر کردن سیستم استخراج اطلاعات، تکنیک ساختن قواعد تداعی‌گر را به آن اضافه خواهیم کرد که از این به بعد با نام قواعد پیشگو به آنها اشاره خواهیم کرد. برای این منظور از الگوریتم C4.5RULES استفاده می‌کنیم[۶].

در سال ۹۳، آقای Quinlan با گسترش الگوریتم استقرایی ID3، روش C4.5 را پیشنهاد کرد. در ID3 هدف، پیشگویی دسته هر ورودی بر اساس مقادیر ویژگی‌هایش بود برای این منظور درخت تصمیمی بدست می‌آمد که هر یک از گره‌های میانی حاوی شرطی بر روی یک ویژگی بود و برگ‌های درخت، دسته‌ای را مشخص می‌کرد که ورودی به آن تعلق داشت. هر یک از مسیرهای موجود از ریشه تا یکی از برگ‌ها نشان دهنده یک قاعده بود. پیمایش تمام مسیرهای مزبور منجر به ایجاد یک مجموعه اولیه از قواعد پیشگو می‌شد. سپس در C4.5 مجموعه قواعد اولیه هرس می‌شود و شروط اضافی از بخش مقدم قواعد حذف و سپس قواعد تکراری حذف می‌شوند. خروجی C4.5، مجموعه قواعد به مراتب کوچکتر و ساده‌تر نسبت به مجموعه قواعد اولیه می‌شد. C4.5RULES نسخه بهبود یافته نرم افزاری C4.5 می‌باشد که یکی از پرکاربردترین برنامه‌ها در داده‌کاوی به حساب می‌آید. پس از اعمال الگوریتم C4.5RULES بر روی یک پایگاه داده قواعدی بصورت زیر خواهیم داشت[۴].

$$A_1 \wedge A_2 \wedge \ldots \wedge A_m \longrightarrow B$$

هر یک از A_i ها و B می‌تواند بصورت $a \in Col_i$ یا $a \notin Col_i$ باشد که نشان دهنده‌ی این است که مقدار ویژگی ستون iام برابر a باشد یا خیر. بر حسب اینکه B کدام یک از این حالات را دارد، بترتیب قواعد پیشگو مثبت و منفی خواهیم داشت.

۳- استفاده از C4.5RULES برای بهبود کارآیی سیستم‌های استخراج اطلاعات

برای بکارگیری تکنیک‌های داده‌کاوی، خروجی سیستم استخراج اطلاعات باید به صورت پایگاه داده رابطه‌ای باشد.خروجی سیستم استخراج اطلاعات بصورت یک ساختار با چند قاب می‌باشد که به راحتی تبدیل به جدول با چند ستون خواهد شد.

سیستم استخراج اطلاعات در یک دامنه‌ی خاص فعالیت می‌کند. دامنه‌ای که به عنوان مثال در این مقاله در نظر گرفته شده است، دامنه‌ی آگهی‌های استخدامی مربوط به رشته‌ی کامپیوتر می‌باشد. این آگهی‌ها را می‌توان در گروه‌های خبری

[6] http://www.cs.uregina.ca/~dbd/cs831/notes/ml/dtrees/c4.5/c4.5rules.html

dataacademy.ir

اینترنتی یافت. سیستم استخراج اطلاعات از روی این آگهی‌ها، خروجی‌هایی مانند نمونه‌ای که در شکل ۳ آورده شده است، استخراج می‌کند.

نکته‌ی دیگر این است که بعضی از مقادیر گرچه از نظر لغوی با هم متفاوت هستند ولی از نظر ماهیت یکسانند. بطور مثال مقادیر VB، Visual Basic و Microsoft VB هر سه به زبان برنامه‌نویسی ویژوال بیسیک اشاره دارند. در این مواقع لازم است که به عنوان یک پیش پردازش، یکسان‌سازی مقادیر صورت گیرد. مرحله‌ی بعدی باینری کردن محتویات جدول است بطوری که مقادیر فقط صفر یا یک باشند[۴]. برای این منظور به جای هر یک از ستون‌های انتخابی، چند ستون به تعداد مقادیر ممکن آن قرار می‌دهیم. حال به ازای هر یک از مقادیر موجود در ستون پیشین، محتویات ستون همنام با آن مقادیر را یک قرار می‌دهیم و بقیه‌ی ستون‌ها را با صفر پر می‌کنیم. اینک جدول آماده اعمال الگوریتم C4.5RULES است. این الگوریتم در ابتدا با ساخت درخت‌های تصمیم و سپس هرس کردن و ترجمه‌ی آن‌ها مجموعه‌ای از قواعد بدست می‌آورد که وجود روابطی بین محتویات ستون‌های جدول را پیشگویی می‌کنند. هر چه مجموعه‌ی آموزشی بزرگتری داشته باشیم، مجموعه قواعد کاراتری خواهیم داشت. در شکل ۴ نمونه‌هایی از قواعد پیشگوی بدست آمده نشان داده شده است.

Title: DBMS Administrator
City: Austin
Area: EC, DB Applivations
Lanquage: VB, VC, C++, Delphi, Java
Platform: Windows, Unix
Application: SQL Server, Oracle
Desires years of experience: 5
Required degree: BS

شکل ۳: نمونه خروجی سیستم استخراج اطلاعات (دامنه‌ی آگهی‌های استخدامی رشته‌ی کامپیوتر)

$$\text{SQL Server} \in \text{Application} \wedge \text{DB} \in \text{Application Area} \longrightarrow \text{Oracle} \in \text{Application}$$

$$\text{VC++} \in \text{Language} \wedge \text{VB} \in \text{Lanquage} \longrightarrow \text{Windows} \in \text{Platform}$$

$$\neg(\text{Windows} \in \text{Platform}) \wedge \text{Unix} \in \text{Platform} \longrightarrow \neg(\text{VC} \in \text{Language})$$

شکل ۴: نمونه قواعد بدست آمده از C4.5RULES

در شکل بالا قاعده‌ی اول و دوم قواعد پیشگوی مثبت و قاعده‌ی سوم قاعده‌ی پیشگوی منفی می‌باشد. بطور مثال قاعده آخر، پیشگویی می‌کند که اگر در مهارت‌های خواسته شده در آگهی، سیستم عامل ویندوز وجود نداشت و در ضمن سیستم عامل یونیکس وجود داشت می‌توان نتیجه‌گیری کرد که احتمالا زبان ویژوال سی جزو مهارت‌های درخواستی نخواهد بود.

در مرحله‌ی بعد می‌خواهیم از این نوع قواعد برای بهبود کارآیی سیستم‌های استخراج اطلاعات استفاده کنیم. کارآیی سیستم‌های استخراج اطلاعات با دو معیار Precision و Recall اندازه‌گیری می‌شوند که به صورت زیر تعریف می‌شوند[۲]:

Recall = (تعداد استخراجهای درست) / (تعداد کل استخراجهای ممکن در متن)

Precision = (تعداد استخراجهای درست) / (تعداد کل استخراجهای انجام شده)

همچنین برای ترکیب این دو، معیار F-Measure با تعریف زیر استفاده می‌شود:

F-Measure = (2 * Precision * Recall) / (Precision + Recall)

معمولا در سیستم‌های استخراج اطلاعات مقدار Recall بسیار کمتر از مقدار Precision است[2,13]. با توجه به تعاریف آن‌ها مشخص است که مقادیر آن‌ها با هم مرتبط‌اند. اغلب روش‌هایی که برای بالا بردن یک معیار استفاده می‌شود باعث کاهش معیار دوم می‌شود و در نتیجه معیار F_Measure تغییر چندانی نمی‌کند. اما ما با استفاده از قواعد پیشگو مقادیر معیارهای Precision و Recall را بدون ضربه زدن به یکدیگر بهبود داده‌ایم. در زیر چند پیشنهاد برای استفاده از قواعد پیشگو آمده است و نتایج اعمال آنها بر روی سیستم WHISK در بخش ارزیابی بیان شده است.

۱- خود قواعد بدست آمده به عنوان یک خروجی باارزش در کنار خروجی‌های بدست آمده از سیستم استخراج اطلاعات قابل استفاده است. بطور مثال از روی قواعد بدست آمده در دامنه‌ی مزبور می‌توان به یک فرد توصیه کرد که اگر می‌خواهد در زمینه‌ی گرافیک در حوزه‌ی نوشتن بازی‌های کامپیوتری دارای بازار کار خوبی داشته باشد چه مهارت‌هایی را بیشتر مد نظر قرار دهد.

۲- قواعد پیشگوی مثبت: با استفاده از قواعد پیشگوی مثبت می‌توان پیشگویی کرد که اگر شرطی خاص بین مقادیر خروجی برقرار باشد باید موردی خاص در خروجی وجود داشته باشد. اگر اطلاعات پیشگویی شده، استخراج نشده باشد، می‌توان آن‌ها را به خروجی اضافه کرد. بدین وسیله معیار Recall بدون صدمه زدن به Precision افزایش یافته است. بطور مثال فرض کنید که قاعده $VC \in lanquage \longrightarrow Win \in platform$ را دشته باشیم. ازمتن ورودی VC بعنوان زبان برنامه‌نویسی استخراج شده است ولی Win بعنوان سیستم عامل استخراج نشده است. پیشگویی می‌شود که احتمالا Win در قاب platform جا افتاده است. برای اینکه گواهی برای این مدعا داشته باشیم در متن ورودی جستجو کرده و اگر کلمه‌ی Win پیدا شد، فرض می‌شود که پیشگویی صحیح بوده است و آن مورد را به خروجی اضافه می‌کنیم.

ممکن است که چند قاعده پیشگو وجود داشته باشد و بکارگیری آن‌ها با ترتیب‌های مختلف منجر به نتایج متفاوت شود. سؤال این است که چه ترتیبی بهتر است. پیشنهاد می‌کنیم قواعدی که دارای عبارت‌های نقیض‌دار بیشتری می‌باشند در الویت قرار بگیرند. زیرا می‌دانیم که بیشتر محتویات جداول با صفر پر شده است و هدف ما افزایش معیار Recall و استخراج بیشتر اطلاعات است.

۳- قواعد پیشگوی منفی: قواعد بدست آمده از C4.5RULES دارای درصد احتمال درستی هستند که به صورت آماری از روی پایگاه داده بدست آمده‌اند. در مورد قبل که از قواعد پیشگوی مثبت استفاده کردیم، می‌توانستیم شاهدی از متن برای تصدیق پیشگویی قاعده بیاوریم. اما درباره‌ی قواعد پیشگوی منفی چنین تصدیقی وجود ندارد. بطور مثال قاعده‌ی $\neg(Win \in area) \longrightarrow \neg(VC \in lanquage)$ پیشگویی می‌کند که در صورتیکه در خروجی Win در قاب area نباشد، نباید VC نیز در قسمت language آمده باشد. اگر با درست بودن فرض قاعده، در خروجی VC در قاب language آمده باشد بدین معنی است که در متن وجود داشته اشت و نمی‌توان تصدیقی برای پیشگویی قاعده‌ی مزبور از درون متن بدست آورد.

دو پیشنهاد برای استفاده از قواعد پیشگوی منفی داریم: اول اینکه، از قواعد با درصد احتمال درستی بالا استفاده کنیم و پیشگویی آن‌ها را با تکیه به بالا بودن احتمال درستی‌شان بپذیریم و اگر در خروجی موردی بر خلاف پیشگویی داشته باشیم، آن را تصحیح کنیم. در این روش تعداد پیشگویی‌ها خیلی کم می‌شود ولی در عین حال، پیشگویی، اکثر اوقات صحیح خواهد بود.

در روش دوم از قواعد پیشگوی منفی با درصد احتمال درستی مناسب استفاده می‌کنیم. وقتی متنی به سیستم استخراج اطلاعات داده می‌شود، اکثر اوقات بیش از یک قاعده‌ی استخراج اطلاعات می‌توانند بر روی آن بکار برده شوند. به عبارت دیگر چون هر قاعده یک الگو را در متن جستجو می‌کند، ممکن است متن ورودی با چند الگو مطابقت کند. برای انتخاب مناسب‌ترین قاعده، علاوه بر معیار مقایسه‌ی قواعد می‌توان از قواعد پیشگوی منفی استفاده کرد و امتیاز بیشتری برای قاعده‌ای در نظر گرفت که با پیشگویی‌های انجام شده مطابقت داشته باشد. با اعمال این روش صحت درستی استخراج‌ها بیشتر خواهد شد. به عبارت دیگر معیار Precision را بهبود داده‌ایم.

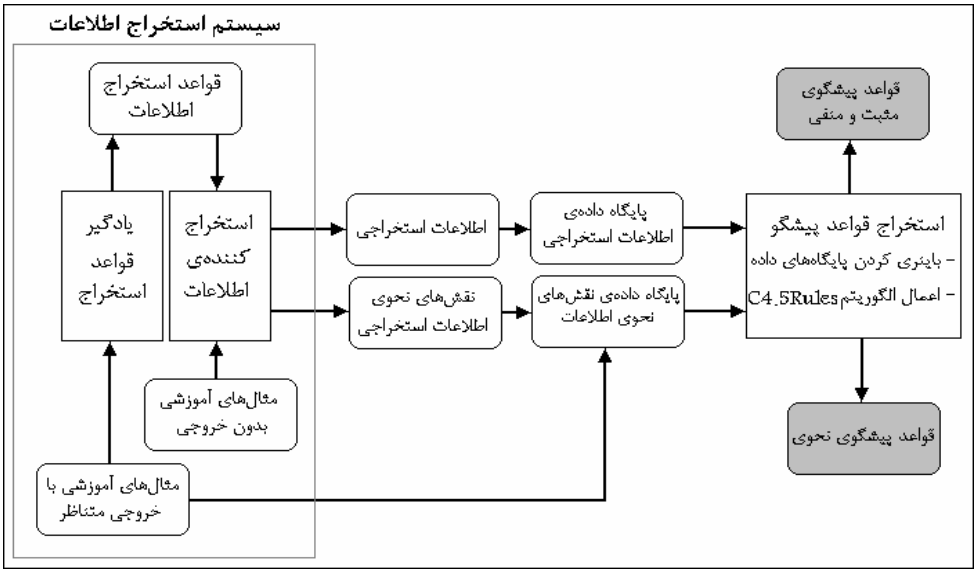
البته می‌توان از تلفیق این دو روش نیز استفاده کرد. بدین طریق که حدی را تعیین کرد و از قواعد با احتمال درستی بالاتر از آن مستقیم در تصحیح خروجی استفاده کرد و از بقیه قواعد برای انتخاب قاعده‌ی استخراج بهینه استفاده کرد. در قسمت ارزیابی از روش دوم استفاده شده است.

۴- قواعد پیشگوی نحوی: سیستم‌های استخراج اطلاعات که برای متون بدون ساختار طراحی شده‌اند بر روی متن ورودی پردازش نحوی انجام می‌دهند و از اطلاعات نحوی اضافه شده به متن برای قدرت‌مند کردن قواعد استخراج اطلاعات استفاده می‌کنند. در خروجی چنین سیستم‌هایی نقش نحوی مقادیر ظاهر شده در خروجی مشخص است و می‌توان پایگاه داده‌ای از نقش‌های مقادیر خروجی ساخت. چون تعداد نقش‌های نحوی محدود و شمارش‌پذیر هستند این پایگاه داده را می‌توان برای اعمال الگوریتم C4.5RULES استفاده کرد و روابط جدید نحوی بین مقادیر سازنده‌ی خروجی‌ها کشف کرد. روش کار در اینجا

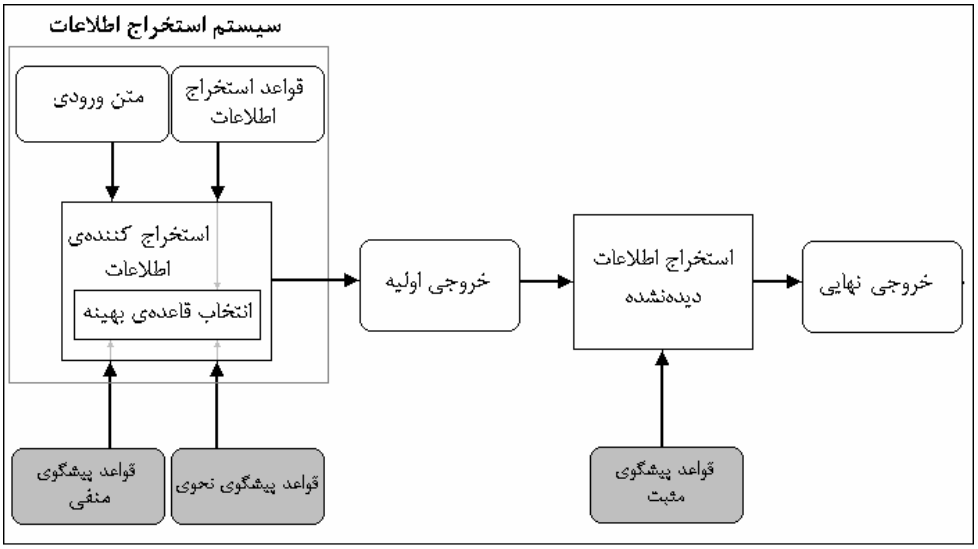
مانند مورد قبل است با این تفاوت که به جای پایگاه داده‌ی خروجی از پایگاه داده‌ی نقش‌های نحوی خروجی‌ها استفاده می‌کنیم. حاصل این کار بدست آوردن قواعدی است که وجود روابطی بین نقش‌های نحوی قاب‌های خروجی را پیشگویی می‌کنند.

نحوه‌ی استفاده از این قواعد نیز مانند مورد قبل است و برای انتخاب قاعده‌ی استخراج بهینه در هنگام تولید خروجی استفاده می‌کنیم. با این کار بهبودی دیگر در معیار Precision حاصل می‌شود. در شکل ۵ معماری پیشنهادی سیستم یادگیرنده‌ی قواعد پیشگو نشان داده شده است. این کار با استفاده از خروجی‌های بدست آمده ازسیستم استخراج اطلاعات و مثال‌های آموزشی (که قبلا خروجی‌هایشان به صورت دستی بدست آمده است) صورت می‌گیرد.

در شکل ۶ معماری سیستم استخراج اطلاعات در حالت استفاده از قواعد پیشگو طراحی شده است. از قواعد پیشگوی منفی و قواعدپیشگوی نحوی برای انتخاب قاعده‌ی بهینه برای تولید خروجی اولیه استفاده می‌شود و سپس از قواعد پیشگوی مثبت برای اضافه کردن اطلاعات جاافتاده‌ی احتمالی استفاده می‌شود.



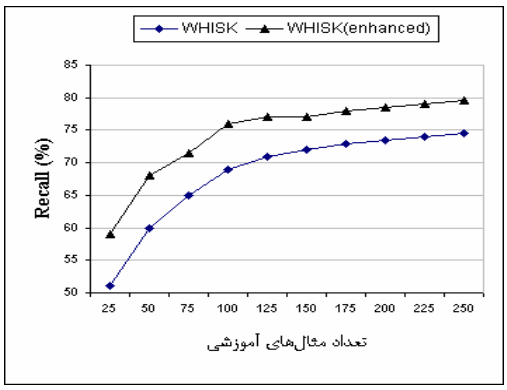
شکل ۵: معماری سیستم یادگیر قواعد پیشگو



شکل ۶: نحوه‌ی استفاده از قواعد پیشگو در سیستم استخراج اطلاعات

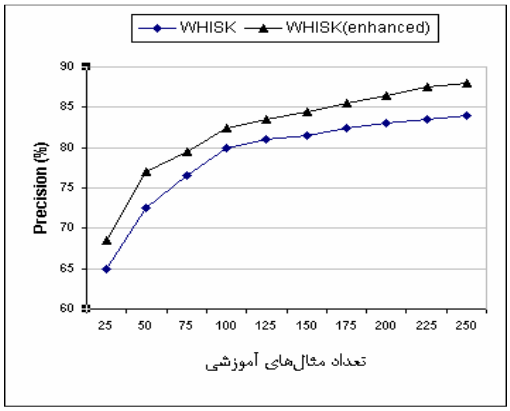
۴- ارزیابی

سیستم استخراج اطلاعاتی که برای ارزیابی استفاده شد، سیستم WHISK می‌باشد. کد برنامه و توضیحات آن در اینترنت موجود است[۷]. دامنه‌ی مورد نظر، آگهی‌های استخدام شغل‌های کامپیوتری می‌باشد. مثال‌های مورد نیاز از گروه‌های خبری اینترنتی مانند www.austinjobs.com جمع‌آوری شد. در مجموع ۳۰۰ مثال بدست آمد. سپس خروجی آن‌ها به صورت دستی محاسبه شد. از ۲۵۰ عدد آن‌ها برای آموزش سیستم WHISK و یادگیری قواعد استخراج استفاده شد. از همان مثال‌ها برای بدست آوردن قواعد پیشگو استفاده شد. به عبارت دیگر از مثال‌های آموزشی بدون خروجی که در شکل ۴ آمده است برای یادگیری قواعد استفاده نشده است. چون در عمل دیده شد زمانیکه از این مثال‌ها استفاده می‌شد، معیارهای کارآیی پایین‌تر می‌آمد.



شکل ۷: مقایسه‌ی معیار Recall

سپس از ۵۰ عدد مثال‌های قبلی به علاوه‌ی ۵۰ مثال باقیمانده برای تست سیستم و بدست آوردن معیارهای کارآیی استفاده شده است.برای یادگیری قواعد پیشگو از فیلدهای area، platform، application و lanquage استفاده شده است. چون ارتباط بیشتری با هم داشتند.

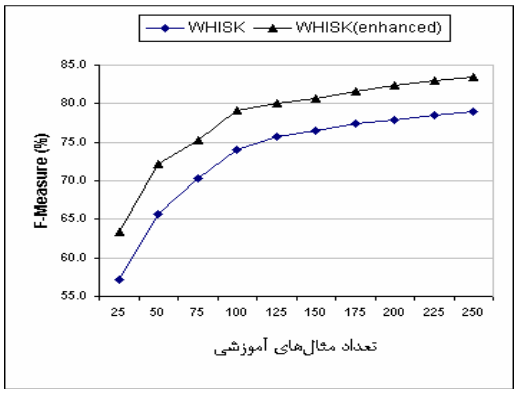


شکل ۸: مقایسه‌ی معیار Precision

اشکال ۷، ۸ و ۹ مقایسه مقادیر معیارهای کارآیی را نشان می‌دهد. مقایسه، توسط مقادیر معیارهای سه‌گانه‌ی ذکر شده در قبل صورت گرفته است. منحنی پایینی برای زمانی است که از سیستم WHISK به تنهایی استفاده شده است و منحنی بالایی برای زمانی است که تکنیک استفاده از قواعد پیشگو را به آن اضافه کرده‌ایم. هر چقدر مثال‌های آموزشی بیشتر شده است، معیارها بهبود بیشتری را نشان داده‌اند ولی از یک مرحله به بعد اضافه کردن آن‌ها بهبود چندانی حاصل نکرده است. بطور متوسط معیارهای Recall، Precision و F_Measure نسبت به قبل (استفاده از WHISK به تنهایی) بترتیب ۶٪، ۴٪ و

[7] http://www.cs.washington.edu/homes/soderlan/WHISK.tar.gz

۵٪ افزایش یافته است. با توجه به اینکه مقادیر معیارهای کارآیی سیستم WHISK در مقایسه با سیستم‌های دیگر استخراج اطلاعات در حد قابل قبولی قرار دارد [۱]، نتایج بدست آمده در جهت بهبود کارآیی سیستم‌های استخراج اطلاعات امیدوار کننده می‌باشد.



شکل ۹: مقایسه‌ی معیار F-Measure

۵- کارهای انجام شده

با توجه اینکه موضوع استخراج اطلاعات از متن در سال‌های اخیر مورد توجه بیشتر قرار گرفته است در گذشته کار چندانی بر روی استفاده از تکنیک‌های داده‌کاوی در استخراج اطلاعات از متن صورت نگرفته است. در سال ۹۵ آقایان Feldnom و Dagan سیستمی را با نام KDT ارائه کردند [۱۴] که از پایگاه داده‌ی متنی، دانش استخراج می‌کرد. سیستم مزبور متون را به صورت دستی و غیر خودکار دسته بندی می‌کرد. به عبارت دیگر KDT از یک دسته‌بندی کننده‌ی خودکار یا سیستم استخراج اطلاعات استفاده نمی‌کرد و در مقاله‌ی مربوطه نیز هیچ بحثی درباره‌ی استفاده از دانش استخراجی برای بهبود عملکرد سیستم نشده است. در سال‌های اخیر توجهات زیادی به مبحث متن‌کاوی جلب شده است. با این حال سیستم‌های عملیاتی و ارزیابی‌های تجربی کمی وجود دارد.

۶- زمینه‌های تحقیقاتی و نتیجه‌گیری

یکی از مراحلی که برای برای تبدیل خروجی سیستم‌های استخراج اطلاعات به پایگاه‌های داده‌ی مورد قبول الگوریتم C4.5RULES وجود داشت، یکپارچه سازی عبارات هم‌معنی بود. این کار در پیاده‌سازی صورت گرفته به صورت دستی انجام شد ولی می‌توان بر روی استفاده از روش‌های خودکار مانند محاسبه‌ی شباهت دو عبارت از نظر حروف تشکیل دهنده[۱۵] تحقیق کرد.

در بررسی صورت گرفته مقادیر قاب‌های خروجی همیشه گسسته فرض شد و برای حالتی که مقادیر پیوسته باشند مانند طول یا قیمت بحثی نشد. یافتن روش‌هایی برای پیشگویی مقادیر پیوسته می‌تواند موضوعی برای تحقیقات آتی باشد.

در این مقاله، اینکه از چه فیلدهایی از جدول برای تولید قواعد پیشگو استفاده شود به صورت ذهنی بدست آمد. اگر بتوان ارتباط میان ستون‌های جدول را از روی محتویات آن‌ها پیدا کرد، می‌توان انتخاب فیلدهای مناسب برای یادگیری قواعدپیشگو را خودکار انجام داد.

مطمئنا در آینده تحقیقات بیشتری بر روی استفاده از تکنیک‌های موجود در موضوعات دیگر هوش مصنوعی برای بهبود کارآیی سیستم‌های استخراج اطلاعات صورت خواهد گرفت.

نتیجه‌گیری اینکه، همکاری استخراج اطلاعات از متن و داده‌کاوی می‌تواند برای کاربردهای مختلفی مفید باشد. سیستم استخراج اطلاعات خروجی‌هایی را در اختیار قرار می‌دهد که داده‌کاوی می‌تواند از روی آن، دانش مفیدی برای کاربردهای مختلف بویژه متن‌کاوی تولید کند.

متن کاوی موضوع نسبتا جدیدی است که در آن پردازش‌های زبان طبیعی، یادگیری ماشین و بازیابی اطلاعات با هم همکاری دارند[۱۶]. ادغام این روش‌ها می‌تواند منجر به ایجاد تکنیک‌های جدید و مفید برای استخراج دانش از مجموعه‌های حجیم متنی شوند.

مراجع

[1] S. Soderland, *Learning Information Extraction Rules for Semi-Structured and Free Text,* Kluwer Academic Publishers, Boston, 1999.
[2] DARPA., ed. Proceedings of the Fifth DARPA Message Understanding Conference, San Mateo, CA: Morgan Kaufman, 1995.
[3] C. Cardie, and R. J. Mooney, *Machine learning and natural language (introduction to special issue on natural language learning)*, Machine Learning Magazine, No 34, pp 5-9, 1999.
[4] J. R. Quinlan, *C4.5: Programs for Machine Learning*, San Mateo, CA: Morgan Kaufman.
[5] F. Peng, *Models Development in IE Tasks*, CS685 Project, Computer Science Department, University of Waterloo, http://ai2.uwaterloo.ca/~f3peng, 2001.
[6] M. E. Callif, and R. J. Mooney, *Relational learning of pattern-match rules for information extraction,* In Proceedings of the Sixteenth National Conference on Artificial Intelligence, pp 328-324, 1999.
[7] H. Ahonen-myka, *Information Extraction from text*, Information Extraction Course Slides, Carnegie Mellon School of Computer Science, spring 2002.
[۸] ا. عبداله زاده، و م. شکری. **روشهای "مهندسی دانش" و "سیستمهای یادگیر" برای طراحی سیستمهای استخراج اطلاعات از متن**، مجموعه مقالات چهارمین کنفرانس دانشجوئی انجمن کامپیوتر ایران، نجف آباد. ۱۳۸۱.
[9] N. Kushmerick, D. Weld, and R. Doorenbos, *Wrapper induction for information extraction,* Proceedings of the Fifteenth International Joint Conference on Artificial Intelligence, pp 729-737, 1997.
[10] M. Holsheimer, M. Kerstenm, H. Mannila, and H. Toivonen, *A perspective on databases and data mining*, Proceedings of the First International Conference on Knowledge Discovery and Data Mining, AAAI Press, PP 150-155, 1995.
[11] G. J. Williams, and Z. Huang, *Modeling the KDD Process, A Four Stage Process and Four Element Model*, CSIRO Division of Information Technology, 1996.
[12] R. Agrawal, and R. Sirkant, *Fast algorithms for mining association rules in large databases*, VLDB'94, 1994.
[13] DARPA., ed. Proceedings of the Fifth DARPA Message Understanding Evaluation and Conference, San Mateo, CA: Morgan Kaufman, 1993.
[14] R. Feldman, and I. Dagan, *Knowledge discovery in textual database (KDT)*, In Proceedings of the First International Conference on Knowledge Discovery and Data Mining, 1995.
[15] E. S. Ristad, and P. N. Yianilos, *Learning string edit distance*, IEEE Transactions on Pattern Analysis and Machine Intelligence 20(5), 1998.
[16] M. Hearst, *Untangling text data mining*, In Proceedings of the 37th Annual Meeting of the Association for Computational Linguistic, pp 3-10, 1999.