شبکه‌ها و پردازش زبان‌طبیعی

دراگومیر رادف
دانشگاه میشیگان
radev@umich.edu
رادا میهالسی
دانشگاه تگزاس شمالی
rada@cs.unt.edu
ترجمه بهروز یل
کارشناسی ارشد علم اطلاعات
و دانش‌شناسی دانشگاه خوارزمی تهران

اشاره

در طول چند سال گذشته، برخی از حوزه‌های پردازش زبان طبیعی کار خود را با به‌کارگیری فنون تصویرمبنا آغاز کرده‌اند. این حوزه‌ها در کنار سایر موارد، شامل: خلاصه‌سازی متن، تجزیه نحوی، عدم ابهام معنای کلمه، ایجاد هستی‌شناسی، تجزیه‌وتحلیل احساسات، تجزیه‌وتحلیل ذهنیت و خوشه‌بندی متن را می‌باشد. در این مقاله، برخی از موفق‌ترین بازنمون‌ها و الگوریتم‌های تصویرمبنا را که در پردازش زبان مورد استفاده قرار می‌گیرد، ارائه می‌دهیم و سعی می‌کنیم سازوکار عمل آنها را شرح دهیم.

۱. مقدمه

واحدهای زبانی در یک متن یکپارچه ـ فارغ از اینکه کلمات، عبارات یا جملات کامل هستند ـ به طرق گوناگونی با هم مرتبط هستند که به مفهوم کلی متن کمک می‌کند و ساختار یکپارچه متن و یکپارچگی گفتار را حفظ می‌نماید. از روزهای نخستین هوش مصنوعی، شبکه‌های پیوندی و معنایی به‌عنوان بازنمون‌هایی پیشنهاد شدند که قادر به ذخیره‌سازی چنین واحدهای زبانی هستند و رابطه‌هایی که آنها را به هم متصل می‌کنند و انواع فرآیندهای استنباط و استدلال را امکان‌پذیر می‌سازد و برخی از عملکردهای ذهن انسان را شبیه‌سازی می‌کنند. ساختارهای رمزی که از این بازنمون‌ها پدیدار می‌شود، به‌طورطبیعی با نمودارها منطبق هستند که در آن سازه‌های متن به عنوان رأس‌ها]قلّه‌ها[(۱) ارائه می‌شوند و رابطه‌های مرتبط، لبه‌ها را در نمودار ایجاد می‌کنند.

در دهه گذشته، تعدادی مقاله پژوهشی قابل‌توجه نوشته شد که روش‌های نمودارمبنا را برای طیف گسترده‌ای از مسائل زبان طبیعی، از یادگیری واژگانی گرفته تا تجزیه جمله و عدم‌ابهام معنای کلمه و خلاصه‌سازی متن استفاده می‌کنند. در این مقاله، روش‌های متعدد و برنامه‌های کاربردی‌شان برای پردازش زبان طبیعی را بررسی می‌کنیم. برای نمایش این واقعیت که الگوریتم‌ها و بازنمون‌ها از جوامع متفاوت ـ پردازش زبان طبیعی و نظریه/ نمودار(۲) ـ سرچشمه می‌گیرند، از یک واژگان

دووجهی برای توصیف این روش‌ها استفاده خواهیم کرد؛ یعنی شبکه‌ها، نمودارها هستند و گره‌ها، رأس‌ها هستند و پیوندها، لبه‌ها هستند.

از نظر بازنمون‌های نمودار مبنا، بسته به برنامه‌های کاربردی پردازش زبان طبیعی از انواع گره‌ها و یال‌ها استفاده شده است. واحدهای متنی از اندازه‌ها و مشخصه‌های متفاوت می‌تواند به‌عنوان رأس‌ها به نمودار اضافه کند؛ برای مثال، کلمات، همایندها (۳)، معانی کلمات، جملات کامل یا حتی اسناد کامل.

توجه داشته باشید که گره‌های نمودار، لزوماً متعلق به همان دسته نیستند؛ به‌عنوان مثال، جملات و کلمات را می‌توان به صورت رأس به همان نمودار اضافه کرد. یال‌ها، هم‌ظهوری (مثلاً: دو کلمه‌ای که در همان جمله یا در همان تعریف واژه‌نامه ظاهر می‌شوند)، همایند (برای مثال: دو کلمه که فوراً در کنار یکدیگر ظاهر می‌شوند یا ممکن است توسط یک حرف ربط از هم جدا شوند)، ساختار نحوی (به‌عنوان مثال: والدین و کودک در یک وابستگی نحوی) و شباهت واژگانی (مثلاً: کسینوس بین بازنمون‌های بُردار از دو جمله) را ارائه دهند.

از نظر الگوریتم‌های نمودارمبنا، روش‌های اصلی مورد استفاده را می‌توان به ۴ دسته تقسیم می‌شود:

۱. طبقه‌بندی نیمه‌نظارتی(۴) (Zhu and Ghahramani, ۲۰۰۲; Zhu and Lafferty, ۲۰۰۵; Toutanova et al., ۲۰۰۴; Radev, ۲۰۰۴; Otterbacher et al., ۲۰۰۵): جایی‌که سیر یا آزادسازی تصادفی در مجموعه‌های مختلط از گره‌های برچسب‌گذاری‌شده و بدون برچسب اعمال می‌شود؛

۲. تحلیل شبکه (Masucci and Rodgers, ۲۰۰۶; Caldeira et al., ۲۰۰۶): که در آن، ویژگی‌های شبکه مانند: قُطر، مرکزیت(۵) و... محاسبه می‌شوند؛

۳. روش‌های خوشه‌بندی نمودارمبنا (Pang and Lee, ۲۰۰۴; Widdows and Dorow, ۲۰۰۲): از جمله روش‌های کمینه برش؛(۶)

۴. الگوریتم‌های درخت پوشای کمینه]درخت فراگیر حداقلی[(۷)(McDonald et al., ۲۰۰۵).

در این مقاله، ما روش‌های متعدد نمودار مبنا را برای وظایف پردازش زبان طبیعی بررسی می‌کنیم که به صورت کلی، به سه دسته اصلی تقسیم می‌شوند. ابتدا فعالیت پژوهشی انجام‌شده در حوزه نحو، از جمله: تجزیه نحوی، پیوند اضافه‌ای(۸) و وضوح هم‌ارجاعی(۹) را بررسی می‌کنیم. سپس، روش‌های مورد استفاده در معناشناسی واژگانی،(۱۰) از جمله عدم‌ابهام معنای کلمه، یادگیری واژگانی(۱۱) و تجزیه‌وتحلیل احساس و ذهنیت را توصیف می‌کنیم. در نهایت، برنامه‌های کاربردی پردازش زبان طبیعی متعددی را بررسی می‌کنیم که بر روش‌های نموداری، از جمله: خلاصه‌سازی متن،(۱۲) بازیابی متن(۱۳) و استخراج کلیدواژه(۱۴) متکی هستند.

۲. نحو

در این بخش، سه مقاله را بررسی می‌کنیم و روش‌هایی برای تجزیه نحوی (McDonald et al., ۲۰۰۵)، پیوند اضافه‌ای (Toutanova et al., ۲۰۰۴) و وضوح هم‌ارجاعی (Nicolae and Nicolae, ۲۰۰۶) ارائه می‌دهیم.

۲ـ۱. تجزیه وابستگی

مک‌دونالد و دیگران (McDonald et al., ۲۰۰۵)، رویکرد نامتعارفی را برای تجزیه جمله اتخاذ کردند. آنها با درک اینکه هر درخت وابستگی، از جمله یک نمودار فرعی جهت‌دار(۱۵) از پیوند تصویر کامل تمام کلمات در جمله است، آغاز به‌کارکردند؛ رویکردی مانند خودشان که روی درخت‌های سازه‌های خیلی شناخته شده کار نمی‌کند؛ همان‌طور آنها غیرپایانه‌ای‌ها(۱۶) را دربرمی‌گیرند. در تجزیه وابستگی، هر جمله به‌صورت یک درخت ارائه می‌شود. ریشه، به طور خاص، گزاره اصلی جمله است (یا آن یک گره ساختگی است و ریشه برچسب‌گذاری‌شده از آنچه گزاره اصلی sole child(۱۷) است) و در کدام یال‌ها برای اتصال هر کلمه به والدین وابستگی آن استفاده می‌شود؛ به‌عنوان مثال، در جمله John Likes green apples، گزاره اصلی likes است

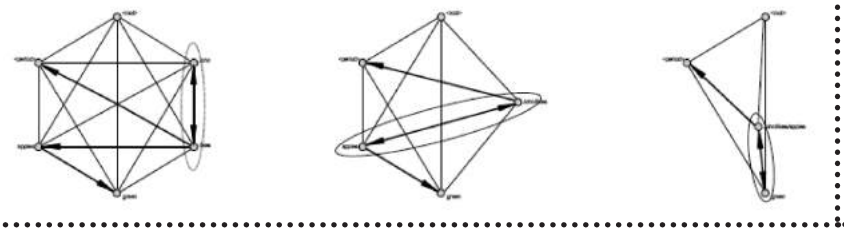
که دو استدلال را در نظر می‌گیرد: کسی که دوست می‌دارد؛ یعنی (John) و شیء دوست‌داشتنی یعنی(apples). در نهایت، از آنجا که سبز بودن، سیب‌ها را تغییر می‌دهد، به درخت به عنوان بچه سیب اضافه می‌شود. درخت نهایی Like به صورت زیر بود: مک‌دونالد و دیگران یک نمودار کامل از برون‌داد جمله ایجاد کردند و سپس، نمره را با هر زیرشاخه جهت‌دار بالقوه از آن نمودار که مساوی با مجموع نمرات تمام یال‌های دربرگیرنده آن است، همراه می‌کند. نمره هر یال در نمودار اصلی، محصول وزن‌بُردار w و بازنمون مشخصه یال f (I,j) است. هدف تجزیه‌کننده، یافتن درخت با بالاترین نمره است. توجه داشته باشید که نقطه‌گذاری،(۱۸) دوره نهایی از یک جمله است که در فرآیند تجزیه استفاده می‌شود.

الگوریتم چو ـ لیو ـ ادموندز(۱۹) (J. and liu ,۱۹۶۵; Edmond ,۱۹۶۷)، این الگوریتم برای یافتن درخت پوشای بیشینه(۲۰) در نمودارهای جهت‌دار استفاده می‌شود. روش این الگوریتم، بدین صورت است: هر گره، همسایه‌ای را انتخاب می‌کند که بالاترین نمره را دارد. نتیجه، یا درخت فراگیر است یا دربردارنده آن چرخه است. روش الگوریتم چو ـ لیو ـ ادموندز، چنین چرخه‌ای را درون یک گره تکی می‌ریزد و نمره‌های هر رویداد یال را از چنین چرخه‌ای دوباره محاسبه می‌کند و حاکی از آن است که الگوریتم چو ـ لیو ـ ادموندز، روی نمودار فروپاشی(۲۱) ایجاد می‌شود؛ همان‌طور الگوریتم چو ـ لیو ـ ادموندز، روی نمودار اصلی دوباره محاسبه می‌شود. این الگوریتم را می‌توان در O(n۲) اجرا کرد.

اجازه دهید به مثالی در مورد مک‌دونالد و دیگران اشاره کنیم. جمله‌ای که باید تجزیه شود John Likes green apples است: نمودار متناظر در نمودار دست‌چپی در شکل ۱ نشان داده شده است و هر گره در نمودار منطبق با کلمه در جمله است. پس از اوّلین تکرار الگوریتم، هیچ درختی پیدا نمی‌شود که تمام گره‌ها را پوشش دهد. بنابراین، دو تا از نزدیک‌ترین گره‌ها فرو می‌ریزند و منجر به نمودار دوم در شکل ۱ می‌شود. این فرآیند تا زمانی ادامه می‌یابد که کل نمودار به گره تکی از طریق مجموعه‌ای از تکرار کاهش یابد.

پس از اینکه تمام گره‌ها به یک گره فرو می‌ریزند، الگوریتم چو ـ لیو ـ ادموندز، با معکوس‌سازی(۲۲) روش ایجاد می‌شود و تمام گره‌ها را درون سازه‌شان بسط می‌دهند. نتیجه نهایی این مثال، در شکل ۲ ارائه شده است.

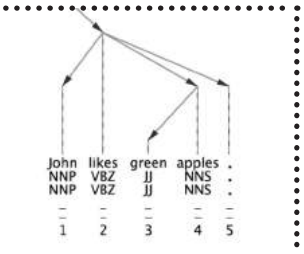
مک‌دونالد و دیگران به پیشرفته‌ترین نتایج با تجزیه‌کننده‌شان در مجموعه داده‌های انگلیسی استاندارد و بهتر از پیشرفته‌ترین نتایج در زبان چِک نائل شدند (زبان منظم کلمه آزاد).



شکل ۱: نمودارهای تهیه‌شده توسط مراحل میانی از الگوریتم چو – لیو – ادموندز

۲ـ۲. پیوند حرف اضافه

پیوند حرف اضافه، یکی از چالش‌برانگیزترین مسائل در تجزیه است. دستور زبان انگلیسی اجازه می‌دهد حرف اضافه، مانند with را به گزاره اصلی از جمله یا بی‌واسطه به عبارت اسمی قبل از آن اضافه کنیم.



شکل ۲: برونداد تجزیه‌کننده الگوریتم چو – لیو – ادموندز

برای مثال، I ate pizza with olives نمونه‌ای از پیوند (اسمی) کوتاه است؛ درحالی‌که I ate pizza with a knife نمونه‌ای از پیوند (فعلی) بلند است. به طور طبیعی، در متن انگلیسی اتفاق می‌افتد و به طور معمول، هر دو نوع پیوند دیده می‌شود.

تاوتانوا(۲۳) و دیگران (Toutanova et al., ۲۰۰۴)، به مشکل پیوند حرف اضافه با قالب‌گیری(۲۴) آن به‌عنوان فرآیند یادگیری نیمه‌نظارتی در نمودارها می‌پردازند. هر گره از نمودار، برابر با یک فعل یا اسم است. دو گره به هم متصل می‌شوند؛ اگر آنها در همان بافتار ظاهر شوند؛ به‌عنوان مثال، گره‌های فعلی hang و fasten به هم متصل می‌شوند؛ زیرا هر دو در عبارات، با nail(۲۵) ظاهر می‌شوند. در یک روش مشابه، گره‌های اسمی nail و rivet(۲۶) به یکدیگر متصل می‌شوند. انواع پیوندها (بیش از ۱۰ نوع، از جمله پیوندها بین کلمات با قالب ریشه مشابه، مترادف‌ها و غیره) در این مقاله توصیف می‌شوند. سپس، الگوریتم با یک سیر تصادفی روی نمودار تا همگرایی ادامه می‌یابد. ارزیابی روی مجموعه آزمون استاندارد پِن تری‌بَنک(۲۷) اجرا شده است. نتایج گزارش‌شده در مقاله عملکرد ۸۷,۵۴ دقت طبقه‌بندی را نشان می‌دهد که خیلی نزدیک به حد بالای، متناظر با عملکرد انسانی (۸۸,۲۰) است.

۲ـ۳. وضوح هم‌ارجاعی

وضوح هم‌ارجاعی، به‌عنوان مشکل شناسایی رابطه‌ها بین ارجاعاتِ هویت در یک متن تعریف می‌شود که آیا آنها توسط اسم‌ها یا ضمایر ارائه می‌شوند. الگوریتم‌های نمونه برای وضوح هم‌ارجاعی تلاش می‌کنند تا زنجیره‌ای از ارجاعات را با استفاده از سامانه‌های قاعده‌محور یا رده‌سازهای یادگیری ماشینی شناسایی کنند. در آثار اخیر (Nicolae and Nicolae, ۲۰۰۶)، روش نمودارمبنا برای وضوح هم‌ارجاعی معرفی شد که تلاش می‌کند تخصیص صحیح ارجاعات به موجودیت‌هایی در یک متن را با استفاده از الگوریتم برش ـ نمودار به طور تخمینی محاسبه کند.

یک نمودار مجزا برای هر نوع موجودیت مؤسسه ملی استاندارد و فناوری(۲۸) ـ مشخص، از جمله: شخص، سازمان، محل سکونت، مهارت و مشارکت جهانی برای آموزش(۲۹) ایجاد می‌شود. بعد، یال‌های وزن‌دار بین ارجاع‌های موجودیت ترسیم می‌شود؛ جایی که وزن‌ها برابر با اطمینان رابطه هم‌ارجاعی است. درنهایت، روش بخش‌بندی مبتنی بر برش ـ کمینه روی این نمودارها اعمال می‌شود که ارجاع‌های برابر با موجودیت یکسان را جدا می‌کند. هنگامی که معیارهای استاندارد برای وضوح هم‌ارجاعی ارزیابی شد، مشخص گردید که الگوریتم نمودارمبنا به عملکرد بسیارپیشرفته منجر می‌شود و به‌طرز قابل‌ملاحظه‌ای الگوریتم‌های قبلی را توسعه می‌دهد.

۳. معناشناسی واژگانی

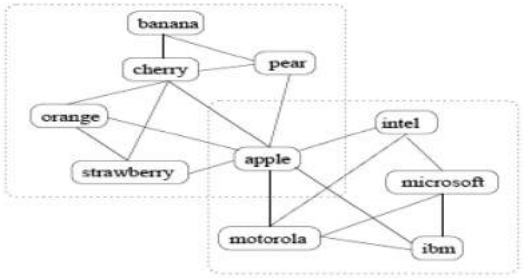
علاقه به تحلیل معناشناسی خودکار از متن برای پشتیبانی برنامه‌های کاربردی پردازش زبان طبیعی، قلمرو ترجمه ماشینی و بازیابی اطلاعات، سامانه پرسش پاسخ و یادگیری دانش در حال افزایش است. پژوهش‌های بسیاری در این حوزه، به‌ویژه روی عدم‌ابهام معنای کلمه، برچسب‌گذاری عملکرد معناشناسی، استلزام متنی(۳۰)، یادگیری واژگانی و روابط معناشناختی انجام شده است. در این بخش، ما روش‌های متعددی را بر اساس بازنمون‌ها و الگوریتم‌های نموداری بررسی خواهیم کرد که قبلاً برای پرداختن به وظایف متفاوت در تحلیل معنایی خودکار استفاده شده است.

۳ـ۱. شبکه‌های واژگانی

یکی از بزرگ‌ترین بازنمون‌های نموداری ایجادشده برای پشتیبانی وظیفه پردازش زبان طبیعی شاید مدل نموداری ارائه‌شده توسط ویدوز و دورو(۳۱) برای یادگیری واژگانی بدون‌نظارت(۳۲) باشد (Widdows and Dorow, ۲۰۰۲). هدف این اثر ایجاد طبقه‌های معنایی با استخراج خودکار از پیکره پردازش‌نشده تمام عناصر متعلق به یک دسته معنایی معین مانند میوه‌ها یا آلات موسیقی است.

این روش با ایجاد یک نمودار بزرگ متشکل از تمام اسم‌ها در یک پیکره بزرگ که توسط (پیکره ملّی بریتانیا، در مورد آن‌ها) توسط حرف ربط and یا or به هم متصل شده‌اند؛ آغاز می‌شود. مقدار قطع]برش[(۳۳) برای پالایش کلمات کمیاب استفاده

می‌شود که به ایجاد نمودار متشکل از تقریبا ۱۰۰۰۰۰ اسم منجر می‌شود که به بیش از نیم میلیون یال مرتبط است. جهت شناسایی عناصر طبقه معنایی، اول تعداد کمی از اسامی معرف به طور دستی انتخاب و برای تشکیل مجموعه دانه(۳۴) استفاده می‌شود. بعد، در یک فرآیند دیگر، گرهی که بیشترین تعداد پیوندها را با مجموعه دانه در نمودار هم‌ظهوری دارد، به‌عنوان «گره» به‌طور بالقوه درست انتخاب می‌شود و بدین‌ترتیب به مجموعه دانه اضافه می‌شود.
تا زمانی فرآیند تکرار می‌شود که هیچ عنصر جدیدی نتواند به طور قابل اطمینانی به مجموعه داده اضافه شود. شکل ۳ نمونه‌ای از نمودار ایجادشده برای استخراج طبقه‌های معنایی را نشان می‌دهد.



شکل ۳: شبکه واژگانی ایجادشده برای استخراج طبقه‌های معنایی

ارزیابی در برابر ده طبقه معنایی از وردنِت، دقت ۸۲ درصدی را نشان داد که با توجه به نویسندگان، نظم دامنه بهتر از کار قبلی در استخراج طبقه معنایی بود. نقطه‌ضعف روش آنها، پوشش کم است؛ با توجه به اینکه روش به آن کلماتِ پیداشده، در رابطه پیوند محدود می‌شود. باوجوداین، هر زمان قابل‌کاربرد است و بازنمون نمودار این توانایی را دارد که به‌دقت کلمات متعلق به طبقه معنایی را شناسایی کند.
حوزه پژوهشی دیگر مربوط به این اثر (Widdows and Dorow, ۲۰۰۲)، مطالعه مشخصه‌های شبکه واژگانی انجام‌شده توسط (Ferrer-i-Cancho and Sole) است. با ایجاد شبکه‌های واژگانی خیلی بزرگ، نزدیک نیم‌میلیون گره، با بیش از ده میلیون یال، ایجادشده توسط کلمات پیوندی در جملات انگلیسی با فاصله حداکثر دو کلمه ظاهر می‌شوند. آنها ثابت کردند که مشخصه‌های سامانه پیچیده در چنین شبکه‌های هم‌ظهوری حفظ می‌شوند.
به طور خاص، آنها اثر جهان ـ کوچک را با تعداد نسبتاً کمی از ۲ ـ ۳ گام مورد نیاز برای اتصال هر دو کلمه در شبکه واژگانی مشاهده کردند. علاوه‌براین، مشاهده شد که توزیع درجه گره درون شبکه، بی‌مقیاس(۳۵) است که تمایل یک پیوند به شکل‌گرفتن با یک کلمه قبلاً خیلی مرتبط را منعکس می‌کنند. شاید تعجب نکنید، مشخصه‌های جهان کوچک و بی‌مقیاس در شبکه‌های واژگانی، به‌طورخودکار، موردنیاز پیکره را مشاهده کردند. همچنین، روی شبکه‌های معنایی به‌طوردستی ایجادشده، مانند وردنت (Sigman and Cecchi, ۲۰۰۲; Steyvers and Tenenbaum, ۲۰۰۵) مشاهده کردند.

۳ـ۲. شباهت و ربط معنایی(۳۶)

الگوریتم‌های نمودار مبنا نیز با موفقیت در شناسایی شباهت و ربط کلمه استفاده می‌شود. گروه بزرگی از روش‌های شباهت معنایی، شامل متریک‌های محاسبه‌شده در شبکه‌های معنایی موجود مانند وردنت و راجِت(۳۷) وجود دارند؛ برای مثال، با استفاده از الگوریتم‌های کوتاه‌ترین مسیر(۳۸)، رابطه معناشناختی نزدیک بین دو مفهوم درون‌داد را شناسایی می‌کنند (Leacock et al., ۱۹۹۸).
به‌تازگی، الگوریتمی مبتنی بر سیر تصادفی توسط هیوز(۳۹) و رَمیج(۴۰) پیشنهاد شد (Hughes and Ramage, ۲۰۰۷). به طور خلاصه، در روش آنها، الگوریتم رتبه‌پیج برای محاسبه توزیع ثابت گره‌ها در نمودار وردنِت و گرایش روی هریک از کلمات درون‌داد در یک جفت کلمه مفروض استفاده می‌شود. بعد، واگرایی بین این توزیع‌ها محاسبه می‌شود که پیوند دو کلمه را نشان می‌دهد. وقتی این روش بر اساس مجموعه داده‌های پیوند کلمه استاندارد ارزیابی گردید، روشن شد که نسبت به الگوریتم‌های

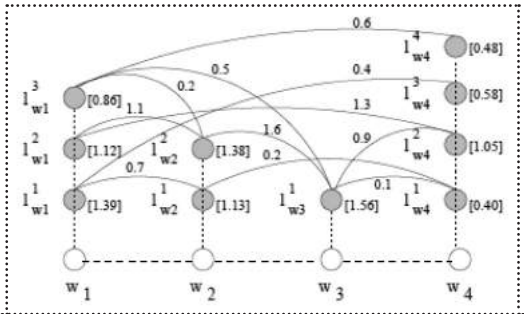
پیشنهادشده قبلی برای ربط معنایی، بسیار بهبود یافته است. در واقع، بهترین سنجه عملکرد آنها به حدّ بالای ارائه‌شده توسط توافق مفسر درونی به این مجموعه داده‌ها نزدیک می‌شود.

۳ـ۳. عدم‌ابهام معنای کلمه

موضوع جالب دیگر در معناشناسی واژگانی، عدم‌ابهام معنای کلمه است و به‌عنوان مشکل شناسایی مناسب‌ترین معنای کلمه باتوجه‌به بافتار آن تعریف می‌شود. بیشتر کارها در این حوزه، دسترس‌پذیری سیاهه معنا از پیش‌تعریف‌شده مانند وردنِت تلقی می‌شود و روش‌هایی را شامل می‌شوند که می‌تواند به‌طورگسترده‌ای به‌عنوان دانش‌مدار(۴۱)، نظارت‌شده یا نیمه‌نظارتی طبقه‌بندی شوند.

روش نمودار مبنا که به‌طور موفقیت‌آمیزی برای عدم ابهام معنای کلمه نیمه نظارتی استفاده می‌شود، الگوریتم انتشار برچسب است (Niu et al.، ۲۰۰۵). در اثرشان، نیو(۴۲) و همکاران، با ایجاد نموداری متشکل از همه برچسب‌ها شروع می‌کنند و نمونه‌های بدون برچسب برای کلمه مبهم مفروض تهیه می‌شوند. نمونه‌های معنای کلمه، به‌عنوان گره‌ها در نمودار استفاده می‌شوند و یال‌های وزنی با استفاده از متریک جفت‌جفت شباهت ترسیم می‌شوند. در این نمودار، همه نمونه‌های برچسب‌دار شناخته‌شده (مجموعه دانه) با برچسب‌های صحیحشان تعیین می‌شوند که پس از آن، در سراسر نمودار در پیوندهای وزنی منتشر می‌گردند. در این روش، تمام گره‌ها با مجموعه‌ای از برچسب‌ها، هریک با احتمال مشخص تعیین می‌شوند. الگوریتم از طریق همگرایی، تکرار می‌شود و با نمونه‌های برچسب‌دار شناخته‌شده با برچسب صحیح در هر تکرار مشخص می‌گردند. در ارزیابی انجام شده در مجموعه داده‌های عدم ابهام معنای کلمه استاندارد، عملکرد الگوریتم برای فراتر رفتن از یک الگوریتم به‌دست‌آمده با خودراه‌انداز(۴۳) یک زبانه یا دو زبانه مشخص شد. همچنین، الگوریتمی برای انجام بهتر نسبت به ماشین پشتیبان‌بُردار هنگامی که فقط تعداد کمی نمونه‌های برچسب‌دار قابل‌دسترس بودند، پیدا شد.

روش‌های نمودارمبنا برای عدم‌ابهام معنای کلمه دانش‌مدار استفاده می‌شود (Mihalcea et al.، ۲۰۰۴; Sinha and Mihalcea، ۲۰۰۷). میهالسی(۴۴) و دیگران، روشی بر پایه نمودارهای ایجادشده مبتنی بر وردنِت را پیشنهاد کردند. با توجه به متن درون‌داد، نمودار با افزودن تمام مفاهیم احتمالی برای کلمات در متن ایجاد و پس از آن، بر اساسِ روابط معنایی موجود در واژگان وردنِت ایجاد می‌شوند (به‌عنوان مثال: مترادف، تضاد معنایی و غیره). برای مثال، شکل ۴ نمونه‌ای از یک نمودار ایجادشده روی یک جمله کوتاه از چهار کلمه را نشان می‌دهد.



شکل ۴: نمودار ایجادشده روی معانی کلمه در یک جمله، برای پشتیبانی عدم‌ابهام معنای کلمه خودکار

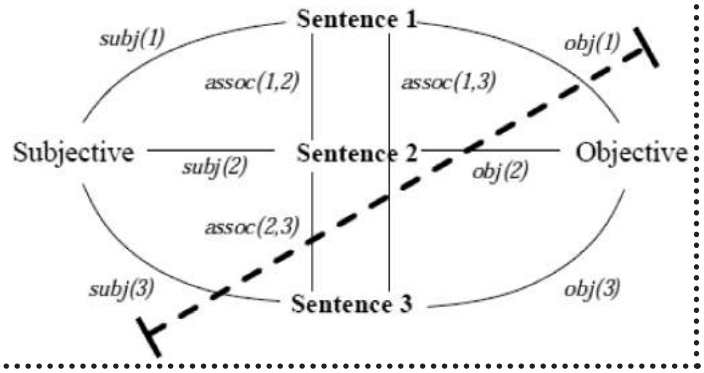
سیر تصادفی اعمال‌شده روی این نمودار، به مجموعه‌ای از نمره‌ها منجر می‌شود که «اهمیت» هر معنای کلمه را در متن مفروض نشان می‌دهد. بنابراین، معانی کلمه با بالاترین نمره به‌طوربالقوه صحیح انتخاب می‌شوند. ارزیابی داده‌های معنی ـ توضیحی نشان داد که این الگوریتم نمودارمبنا برای (انتخاب) روش‌های دانش‌مبنای جایگزین، عالی بود که استفاده از چنین بازنمون‌های قوی از روابط معنای کلمه را ایجاد نمی‌کنند.

در اثر بعدی، میهالسی روش نمودارمبنای کلی‌تری را توسعه داد که نیازی به دسترس‌پذیری روابط معنایی مانند موارد مشخص‌شده در وردنِت نیست. در عوض، او یال‌های وزنی اشتقاقی تعیین‌شده با استفاده از سنجه شباهت واژگانی در میان

تعاریف معنای کلمه را استفاده کرد (Mihalcea, ۲۰۰۵) که عمومیت را پدید می‌آورد؛ همان‌طور این روش به شبکه‌های معنایی همچون وردنِت محدود نمی‌شود؛ اما می‌توان آن را در هر واژه‌نامه الکترونیکی استفاده کرد.
نویگلی(۴۵) و لاپاتا(۴۶) به سبکِ (میهالسی و دیگران، ۲۰۰۴)، ارزیابی مقایسه‌ای از الگوریتم‌های ارتباط نمودار متفاوت به‌کارگرفته‌شده در نمودارهای معنای کلمه مشتق‌شده از وردنِت (Navigli and Lapata, ۲۰۰۷) را اجرا کردند. آنها دریافتند که بهترین دقت عدم‌ابهام معنای کلمه، با استفاده از یک سنجه شباهت برای الگوریتم‌های مرکزیت نمودار دیگر نظیر indegree(۴۷)، رتبه‌پیج و بینیت(۴۸)، عالی بود.

۳ـ٤. احساس و ذهنیت

تجزیه و تحلیل احساس و ذهنیت، حوزه‌ای مرتبط به معناشناسی و عمل‌گرایی(۴۹) است که توجه زیاد جامعه پژوهشی را به خود جلب کرد. روش مبتنی بر نمودار، توسط پانگ(۵۰) و لی(۵۱) (Pang and Lee, ۲۰۰۴) ارائه شد. در این سامانه، آنها نشان می‌دهند که الگوریتم نمودارمبنای بُرش کمینه می‌تواند به شکل کارآمدی برای ایجاد خلاصه‌های ذهنی از نقدهای فیلم استفاده شود.
ابتدا آنها نموداری را با افزودن همه جملات در یک بازبینی به‌صورت گره‌ها و با ترسیم یال‌هایی مبتنی بر همجواری جمله ایجاد می‌کنند. هر گره در نمودار ابتدا با نمره‌ای که نشان‌دهنده احتمال جمله متناظر ذهنی یا عینی است، بر اساس برآورد ارائه‌شده توسط رده‌بند(۵۲) ذهنیت نظارتی تعیین می‌شود. سپس، الگوریتم بُرش کمینه روی نمودار اعمال می‌شود و برای جداکردن جملات ذهنی از جملات عینی استفاده می‌شود. شکل ۵ نمودار ایجادشده روی جملات در یک متن را نشان می‌دهد که در آن، نمودار الگوریتم بُرش کمینه برای شناسایی و استخراج جملات ذهنی استفاده می‌شود.
دقت این رده‌بند ذهنیت نمودارمبنا، بهتر از برچسب‌گذاری به‌دست‌آمده با رده‌بند نظارتی اوّلیه بود. علاوه‌براین، رده‌بند پلاریته (قطب‌داری)، متکی بر خلاصه‌های جایگزین برش کمینه، دقیق‌تر از موارد به‌کارگرفته‌شده در کل بررسی‌ها بودند.



شکل ۵: طبقه‌بندی ذهنی با استفاده از الگوریتم بُرش کمینه است. نقطه‌چین انشعاب بین جملات ذهنی و عینی، همان‌طور به‌دست‌آمده با الگوریتم بُرش کمینه را نشان می‌دهد.

پژوهش اخیر در مورد تجزیه‌وتحلیل احساس و ذهنیت می‌باشد. همچنین، معانی و ذهنیت کلمه را نیز در نظر می‌گیرد (Wiebe and Mihalcea, ۲۰۰۶). در کار، تخصیص ذهنیت و برچسب‌های پولاریته به معانی وردنِت را هدف قرار دادند. اِسالی(۵۳) و سباستیانی(۵۴)، الگوریتم رتبه‌پیج اریب متمایل را در کل نمودار وردنِت به کار گرفتند. تا حدی شبیه به روش انتشار برچسب، الگوریتم سیر تصادفی با گره‌های برچسب‌گذاری‌شده ذهنیت و قطبیت ایجاد می‌شوند. هنگام مقایسه با روش طبقه‌بندی ساده، سیر تصادفی‌شان به یادداشت‌های توضیحی دقیق‌تر از معانی کلمه ذهنیت و قطبیت منجر می‌شود.

٤. برنامه‌های کاربردی دیگر

برخی از برنامه‌های کاربردی پردازش زبان طبیعی دیگر، مانند: خلاصه‌نویسی متن، بازیابی متن و استخراج کلیدواژه، برای فنون نمودارمبنا مستعد هستند.

٤ـ١. خلاصه‌سازی

یکی از نخستین روش‌های نمودارمبنا، برای خلاصه‌سازی توسط اَلِن(۵۵) و دیگران معرفی شد (Saltom et al.، ۱۹۹۴؛ Saltom et al.، ۱۹۹۷). این روش آنها، در مقاله‌هایی از دایره‌المعارف فانک(۵۶) و وَگنلز(۵۷) بروز یافت؛ به‌عنوان نمودارهایی که در آن، هر گره منطبق با پاراگراف است و پاراگراف‌های مشابه به‌لحاظ واژگانی با هم مرتبط هستند. خلاصه‌ای از سند و مسیرهای ذیل، با الگوریتم‌های متفاوت مشخص شده‌اند که همان‌اندازه از محتوای نمودار را که امکان‌پذیر است، پوشش می‌دهند. (Erkan and Radev, ۲۰۰۴; Mihalcea and Tarau, ۲۰۰۴) ایده خلاصه‌سازی نمودارمبنا را بیشتر با معرفی مفهوم مرکزیت واژگانی به کار گرفتند. مرکزیت واژگانی، سنجه‌ای از اهمیت (مرکزیت) گره‌ها در یک نمودار شکل‌گرفته توسط پیوند جملات یا اسناد مرتبط از لحاظ واژگانی است. پس، سیر تصادفی روی نمودار پیاده‌سازی می‌شود و گره‌هایی که مشاهده می‌شوند، اغلب به‌عنوان خلاصه‌ای از نمودار درون‌داد انتخاب می‌شوند (که در بیشتر موارد، اطلاعاتی از اسناد متعدد را در برمی‌گیرند). باوجوداین، ابتدا باید توجه داشته باشید که به‌منظور اجتناب از گره‌هایی با محتوای تکراری یا تقریباً تکراری، تصمیم نهایی در مورد اینکه شامل یک گره در خلاصه می‌شود، به حداکثر ربط حاشیه‌ای(۵۸) آن بستگی دارد؛ چنان‌که در (Carbonell and Goldstein, ۱۹۹۸) مشخص می‌شود. همچنین، (Erkan and Radev, ۲۰۰۴) روی فناوری خلاصه‌سازی پیشین، یعنی سامانه خلاصه‌سازی، اخبار قابل‌دسترس وب اوّل، NewsInEssence (Radev et al.، ۲۰۰۱) را ایجاد کردند.

نمونه‌ای از (Erkan and Radev, ۲۰۰۴)، در شکل ۶ نشان داده شده است که درون‌داد ۱۱ جمله از گزارش‌های خبری مختلف در موضوع‌های مرتبط را دربرمی‌گیرد. شکل ۷، شباهت‌های کسینوسی از تمام زوج‌های جمله‌ها را به نمایش می‌گذارد؛ درحالی‌که شکل ۸، توزیع کسینوس‌ها را نشان می‌دهد.

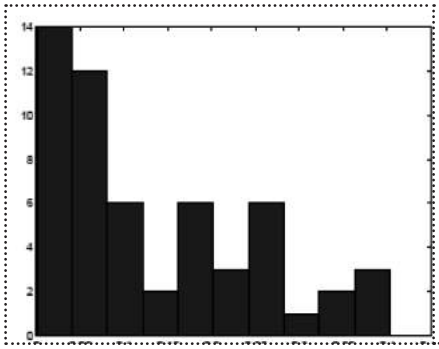
SNo	ID	Text
1	d1s1	Iraqi Vice President Taha Yassin Ramadan announced today, Sunday, that Iraq refuses to back down from its decision to stop cooperating with disarmament inspectors before its demands are met.
2	d2s1	Iraqi Vice president Taha Yassin Ramadan announced today, Thursday, that Iraq rejects cooperating with the United Nations except on the issue of lifting the blockade imposed upon it since the year 1990.
3	d2s2	Ramadan told reporters in Baghdad that "Iraq cannot deal positively with whoever represents the Security Council unless there was a clear stance on the issue of lifting the blockade off of it.
4	d2s3	Baghdad had decided late last October to completely cease cooperating with the inspectors of the United Nations Special Commission (UNSCOM), in charge of disarming Iraq's weapons, and whose work became very limited since the fifth of August, and announced it will not resume its cooperation with the Commission even if it were subjected to a military operation.
5	d3s1	The Russian Foreign Minister, Igor Ivanov, warned today, Wednesday against using force against Iraq, which will destroy, according to him, seven years of difficult diplomatic work and will complicate the regional situation in the area.
6	d3s2	Ivanov contended that carrying out air strikes against Iraq, who refuses to cooperate with the United Nations inspectors, "will end the tremendous work achieved by the international group during the past seven years and will complicate the situation in the region."
7	d3s3	Nevertheless, Ivanov stressed that Baghdad must resume working with the Special Commission in charge of disarming the Iraqi weapons of mass destruction (UNSCOM).
8	d4s1	The Special Representative of the United Nations Secretary-General in Baghdad, Prakash Shah, announced today, Wednesday, after meeting with the Iraqi Deputy Prime Minister Tariq Aziz, that Iraq refuses to back down from its decision to cut off cooperation with the disarmament inspectors.
9	d5s1	British Prime Minister Tony Blair said today, Sunday, that the crisis between the international community and Iraq "did not end" and that Britain is still "ready, prepared, and able to strike Iraq."
10	d5s2	In a gathering with the press held at the Prime Minister's office, Blair contended that the crisis with Iraq "will not end until Iraq has absolutely and unconditionally respected its commitments" towards the United Nations.
11	d5s3	A spokesman for Tony Blair had indicated that the British Prime Minister gave permission to British Air Force Tornado planes stationed in Kuwait to join the aerial bombardment against Iraq.

شکل ۶: خوشه‌ای از ۱۱ جمله مرتبط

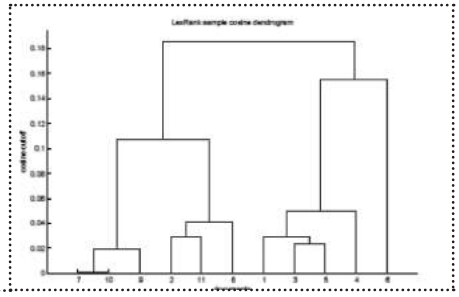
	1	2	3	4	5	6	7	8	9	10	11
1	1.00	0.45	0.02	0.17	0.03	0.22	0.03	0.28	0.06	0.06	0.00
2	0.45	1.00	0.16	0.27	0.03	0.19	0.03	0.21	0.03	0.15	0.00
3	0.02	0.16	1.00	0.03	0.00	0.01	0.03	0.04	0.00	0.01	0.00
4	0.17	0.27	0.03	1.00	0.01	0.16	0.28	0.17	0.00	0.09	0.01
5	0.03	0.03	0.00	0.01	1.00	0.29	0.05	0.15	0.20	0.04	0.18
6	0.22	0.19	0.01	0.16	0.29	1.00	0.05	0.29	0.04	0.20	0.03
7	0.03	0.03	0.03	0.28	0.05	0.05	1.00	0.06	0.00	0.00	0.01
8	0.28	0.21	0.04	0.17	0.15	0.29	0.06	1.00	0.25	0.20	0.17
9	0.06	0.03	0.00	0.00	0.20	0.04	0.00	0.25	1.00	0.26	0.38
10	0.06	0.15	0.01	0.09	0.04	0.20	0.00	0.20	0.26	1.00	0.12
11	0.00	0.00	0.00	0.01	0.18	0.03	0.01	0.17	0.38	0.12	1.00

شکل ۷: شباهت‌های کسینوسی در تمام زوج‌های جمله در خوشه‌ای از ۱۱ جمله

مهم است بدانیم که ماتریس کسینوس در خود تعداد نامحدودی از نمودارها را برای هر مقدار از انقطاع کسینوس، t مخفی می‌کند. این را می‌توان در دو شکل بعدی دید: شکل‌های ۱۰ ـ ۹. به‌عنوان مثال، اگر یکی آستانه را بیش از حد کم کند، نمودار تقریباً کاملاً مرتبط (متصل) است. برعکس، با بالا بردن آستانه بالأخره نمودار را به مجموعه‌ای از مؤلفه‌های غیرمرتبط تبدیل می‌کنند. سیر تصادفی به‌طور خاص، در مقدار t که در آن تقریباً نیمی از زوج‌های گره از طریق یال‌ها متصل می‌شوند، انجام می‌گیرد.



شکل ۸: نمودار ستونی کسینوس لکس‌رنک



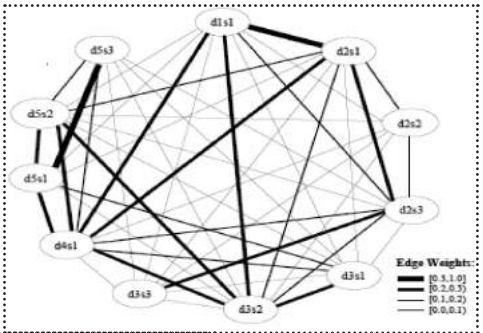
شکل ۹: درختواره‌نگار(۵۹) نمونه لکس‌رنک

شکل ۱۱، رابط کاربری جاوا لکس‌رنک استفاده‌شده برای خلاصه‌سازی متن را نشان می‌دهد

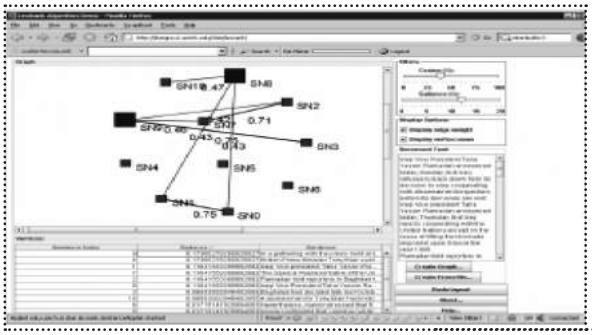
٤ـ۲. بازیابی متن نیمه‌نظارتی

آترباچر(۶۰) و دیگران (Otterbacher et al.، ۲۰۰۵)، ابتدا نظریه (Erkan and Radev, ۲۰۰۴) را با معرفی مفهوم سیر تصادفی اریب برای پرداختن به مسئله بازیابی متن پرسش‌محور بسط دادند. در آن مشکل، کاربر پرسش را در قالب پرسش زبان طبیعی وارد می‌کند و انتظار دارد مجموعه از متن‌های اسناد درون‌داد که شامل پاسخ به این سؤال می‌شود، به او جواب

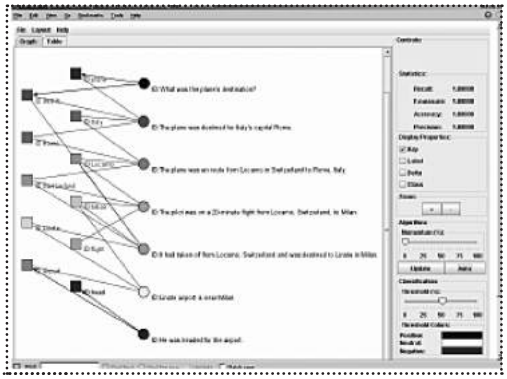
دهد. سیر تصادفی اریب، روی نموداری اجرا می‌شود که قبلاً با نمونه‌های مثبت و منفی شناخته‌شده ایجاد شده‌اند. پس، هر گره متناسب با درصد دفعات سیر تصادفی روی یال‌های نمودار در آن گره، برچسب‌گذاری می‌شوند. در آغاز با توجه به حضور گره‌های برچسب‌گذاری‌شده و درنهایت گره‌ها دارای بالاترین نمره، آن‌هایی هستند که هر دو شبیه به گره‌های اوّلیه (هسته) و مرکزی برای مجموعه سند هستند؛ به‌عبارت دیگر، درنهایت، آنها به‌عنوان مجموعه با یک مدل ترکیبی که با توجه به دانه‌های شناخته‌شده (مثبت یا منفی) و نمره مرکزیت، همان‌طورکه در بخش قبلی است، انتخاب می‌شوند. نمودار، شامل جملات (پاراگراف‌ها) و مشخصه‌ها (کلمات محتوا که در این جمله‌ها ظاهر می‌شوند) بودند. نمودار دوبخشی است؛ همان‌طورکه یک جمله فقط می‌تواند به یک مشخصه و برعکس، پیوند یابد.



شکل ۱۰: نمودار شباهت کسینوس وزنی برای خوشه در شکل ۶



شکل ۱۱: رابط کاربری لکس‌رنک



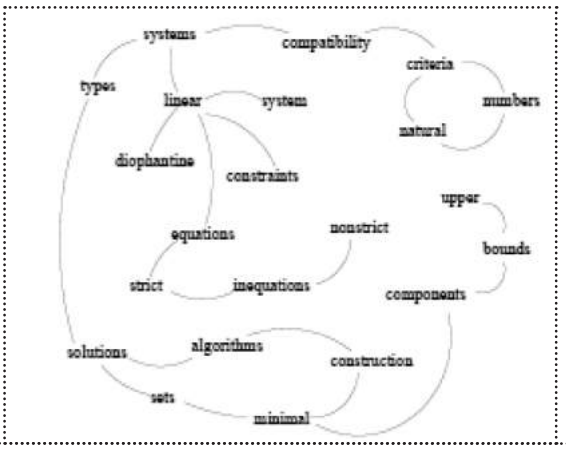
نمودار ۱۲: لکس‌رنک اریب؛ همان‌طور برای بازیابی متن نیمه‌نظارتی استفاده می‌شود

در مثالِ نشان‌داده‌شده در شکل ۱۲، گره بالا سمت راست، در ابتدا به‌عنوان مثبت (سیاه) برچسب‌گذاری می‌شود؛ درحالی‌که گره پایین سمت راست، به‌عنوان منفی (گره روشن) برچسب‌گذاری می‌شود. در طول برچسب‌گذاری (اجراشده با استفاده از روش آزادسازی)، عمق رنگ گره تغییر می‌کند تا فرآیند همگرا شود. بالأخره، گره‌های تیره به‌عنوان مرتبط به سؤال کاربر بازمی‌گردند. توجه داشته باشید که برخی از آنها، هیچ کلمه‌ای هماهنگ با پرسش اصلی را دربرنمی‌گیرند.

٤ـ٣. استخراج کلیدواژه

وظیفه برنامه کاربردی استخراج کلیدواژه، این است که به‌طورخودکار در یک متن، مجموعه اصطلاح‌هایی را که بهترین توصیف سند هستند، شناسایی کند. چنین کلیدواژه‌هایی، مدخل‌های مفید برای ایجاد نمایه خودکار برای مجموعه سند را به وجود می‌آورند و می‌توانند برای طبقه‌بندی متن یا به‌عنوان یک خلاصه کسینوسی برای سند مفروض به کار گرفته شوند. سامانه‌ای برای شناسایی خودکار اصطلاح‌های مهم، همچنین می‌تواند برای مشکل استخراج مجموعه اصطلاحات و ایجاد واژه‌نامه‌های خاص ـ حوزه‌ای استفاده شود. الگوریتم سیر تصادفی برای استخراج کلیدواژه در (Mihalcea and Tarau, ۲۰۰۴) پیشنهاد شد که در آن نمودار روی متن درون‌داد با افزودن تمام کلمات در متن به‌عنوان گره‌هایی در نمودار ایجاد می‌شود و ارتباط آنها با رابط هم‌ظهوری با فاصله بین کلمات محدود می‌شود.

شکل ۱۳، نمودار نمونه ایجادشده برای متن علمی کوتاه را نشان می‌دهد. سیر تصادفی روی چنین نموداری، از هم‌ظهوری‌ها اجرا می‌شود که به رتبه‌بندی در اهمیت کلمات در متن منجر می‌شود. در مرحله پس‌پردازش، کلمات مهم با رتبه‌بندی الگوریتم انتخاب می‌شوند و در مجاورت یکدیگر در متنی که درون یک عبارت جداگانه فرومی‌ریزند، یافت می‌شوند. جالب اینکه مقایسه آزمایش‌های این رتبه‌بندی با فراوانی اصطلاح ـ فراوانی سند معکوس سنّتی، نشان داد که نمره‌های اختصاص‌یافته به سیر تصادفی می‌تواند به‌طور قابل‌توجهی متفاوت باشد. درواقع، ارزیابی‌ها روی مجموعه داده چکیده‌های علمی نشان داد که روش سیر تصادفی، از روش فراوانی اصطلاح ـ فراوانی سند معکوس برای استخراج کلیدواژه، بهتر است و همچنین، آن نسبت به روش‌های نظارتی بسیار پیشرفته قبلاً منتشرشده برای استخراج کلیدواژه، بهبودیافته است.



شکل ۱۳: نمودار نمونه ایجادشده برای استخراج کلیدواژه

٤ـ٤. مطالعه بیشتر

کتاب‌شناسی قابل‌توجهی روی وبگاه مؤلف و روی پایگاه www.textgraphs.org دیده می‌شود.

سپاسگزاری‌ها

این مقاله، تا حدودی مبتنی بر اثر انجام‌شده قبلی توسط دو نویسنده بود. آن اثر تا حدودی توسط کمکِ مالی بنیاد ملی علوم به شماره IIS ۰۵۳۴۳۲۳ " پژوهش مشارکتی: بلاگوسنتر ـ زیرساختاری برای جمع‌آوری، داده‌کاوی و دسترسی به بلاگ‌ها" به شماره ۰۳۲۹۰۴۳ " روش‌های احتمالاتی و پیوندمحور برای به‌کارگیری انباره‌های متنی خیلی بزرگ" و BCS ۰۵۲۷۵۱۳، " DHB: پویاشناسی بازنمون سیاسی و سخنوری سیاسی " و توسط کمکِ مالی مؤسسه ملی سلامت(۶۱) R۰۱LM۰۰۸۱۰۶ " بازنمون و فراهم‌آوری دانش قاعده ژنوم" و DA۰۲۱۵۱۹ U۵۴ " مرکز ملی زیست‌داده‌ورزی(۶۲) یکپارچه که همه برای دراگومیر رادف است، تأمین مالی می‌شود. این اثر نیز در بخشی توسط کمک‌هزینه پژوهشی #۰۰۳۵۹۴ در «پردازش زبان طبیعی نمودارمبنا " برنامه پژوهشی پیشرفته تگزاس و با کمک‌هزینه گوگل در «یافتن اطلاعات مهم در متن بدون ساختار" که هردوی آنها به رادا میهالسی اهدا شد، پشتیبانی می‌شود.

هر نظر، یافته و نتیجه‌گیری یا توصیه بیان‌شده در این منبع، متعلق به نویسندگان است و لزوماً دیدگاه‌های بنیاد ملی علوم یا حامیان دیگر نیست.

پی‌نوشت‌ها:

1. Vertices.
2. Graph-Theory.
3. Collocation.
4. Semi-Supervised Classification.
5. Centrality.
6. Min-cut.
7. Minimum Spanning-Tree.
8. Prepositional Attachment.
9. Co-Reference.
10. Lexical Semantics.
11. Lexical Acquisition.
12. Text Summarization.
13. Passage Retrieval.
14. Keyword Extraction.
15. Driected Subgraph.
16. Non-Terminal.

۱۷. کف‌پای کودک.

18. Punctuation.
19. Chu-Liu-Edmonds(CLE) Algorithm.
20. Maximum Spanning Tree (MST).
21. Collapsed Graph.
22. Reversing.
23. Toutanova.

24. Casting.

۲۵. میخ.

۲۶. میخ پرچ.

27. Penn Treebank.
28. National Institute of Standards and Technology.
29. Global Partnership for Education(GPE).
30. Textual Entailment.
31. Widdows and Dorow.
32. Unsupervised.
33. Cutoff Value.
34. Seed Set.
35. Scale free.
36. Semantic Similarity and Relatedness.
37. Roget.
38. Shortest Path.
39. Hughes.
40. Ramage.
41. Knowledge-Based.
42. Niu.
43. bootstrapping.
44. Mihalcea.
45. Navigli.
46. Lapata.

۴۷. تعداد لبه‌های جهت‌دار که به سوی یک گره اشاره می‌کند یا درجه ورودی.

48. Betweenness.
49. Pragmatics.
50. Pang.
51. Lee.
52. Classifier.
53. Esuli.
54. Sebastiani.
55. Allan.
56. Funk.
57. Wagnalls.
58. Maximal Marginal Relevance.
59. Dendrogram.
60. Otterbacher.

61. National Institutes of Health.
62. Bioinformatics.

منابع:

1. Silvia M. G. Caldeira, Thierry C. Petit Lob ao, R. F. S.Andrade, Alexis Neme, and J. G. V. Miranda. 2006.The network of concepts in written texts. European Physical Journal B, 49(4):523–529, February.
2. Jaime G. Carbonell and Jade Goldstein. 1998. The use of MMR, diversity-based reranking for reordering documents and producing summaries. In Proceedings of the Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pages 335–336.
3. J. Edmonds. 1967. Optimum branchings. Journal of Research of the National Bureau of Standards, 71B:233– 240.
4. G¨unes¸ Erkan and Dragomir Radev. 2004. The university ofMichigan at duc 2004. In Document Understanding Conference (DUC), Boston, Massachusetts, May.
5. A. Esuli and F. Sebastiani. 2007. PageRanking wordnet synsets: An application to opinion mining. In Proceedings of the Annual Meeting of the Association of Computational Linguistics, Prague, Czech Republic.
6. Ramon Ferrer-i-Cancho and Ricard V. Sole. 2001. The small world of human language. Proceedings of The Royal Society of London. Series B, Biological Sciences,268(1482):2261–2265, November.
7. T. Hughes and D. Ramage. 2007. Lexical semantic relatedness with random graph walks. In Proceedings of EMNLP 2007, Prague, Czech Republic.
8. Y. J. and T. H. Liu. 1965. On the shortest arborescence of a directed graph. Science Sinica, 14:1396–1400.
9. C. Leacock,M. Chodorow, and G.A.Miller. 1998. Using corpus statistics andWordNet relations for sense identification.Computational Linguistics, 24(1):147–165.
10. A. P. Masucci and G. J. Rodgers. 2006. Network properties of written human language. Physical Review E, 74, August 2,.
11. Ryan McDonald, Fernando Pereira, Kiril Ribarov, and Jan Hajic. 2005. Non-projective dependency parsing using spanning tree algorithms. In Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing, pages 523–530,Vancouver, British Columbia,Canada, October.
12. Rada Mihalcea and Paul Tarau. 2004. Textrank: Bringing order into texts. In Proceedings of EMNLP 2004,pages 404–411, Barcelona, Spain, July. Association for Computational Linguistics.

13. R. Mihalcea, P. Tarau, and E. Figa. 2004. PageRank on semantic networks, with application to word sense disambiguation.In Proceedings of the 20st International Conference on Computational Linguistics (COLING 2004), Geneva, Switzerland.
14. Rada Mihalcea. 2005. Unsupervised large-vocabulary word sense disambiguation with graph-based algorithms for sequence data labeling. In Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing, pages 411–418,Vancouver, British Columbia,Canada, October.
15. Roberto Navigli and Mirella Lapata. 2007. Graph connectivity measures for unsupervised word sense disambiguation. In Proceedings of the 20th International Joint Conference on Artificial Intelligence, Hyderabad,India.
16. Cristina Nicolae and Gabriel Nicolae. 2006. Bestcut: A graph algorithm for coreference resolution. In Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing, pages 275–283,Sydney, Australia, July.
17. .Z.Y. Niu, D.H. Ji, and C.L. Tan. 2005. Word sense disambiguation using label propagation based semisupervised learning. In Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics, Ann Arbor, Michigan. Association for Computational Linguistics.
18. Jahna Otterbacher, G¨unes¸ Erkan, and Dragomir Radev. 2005. Using random walks for question-focused sentence retrieval. In Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing, pages 915– 922, Vancouver, British Columbia, Canada, October. Association for Computational Linguistics.
19. Bo Pang and Lillian Lee. 2004. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In Proceedings of the 42ndMeeting of the Association for Computational Linguistics (ACL'04), Main Volume, pages 271–278, Barcelona, Spain, July.
20. Dragomir R. Radev, Sasha Blair-Goldensohn, Zhu Zhang, and Revathi Sundara Raghavan. 2001. NewsInEssence: A system for domain-independent, realtime news clustering and multi-document summarization. In Proceedings of Human Language Technology Conference (HLT 2001).
21. Dragomir R. Radev. 2004. Weakly supervised graphbased methods for classification. Technical Report CSE-TR-500-04, University of Michigan. Department of Electrical Engineering and Computer Science.
22. Gerard Salton, James Allan, Chris Buckley, and Amit Singhal. 1994. Automatic analysis, theme generation, and summarization of machine-readable texts. Science, 264(5164):1421–1426.
23. Gerard Salton, Amit Singhal, Mandar Mitra, and Chris Buckley. 1997. Automatic text

structuring and summarization. 33(2):193–207,March.
24. Mariano Sigman and Guillermo A. Cecchi. 2002. Global organization of the Wordnet lexicon. Proceedings of the National Academy of Sciences of the United States of America, 99(3):1742–1747, February 5,.
25. R. Sinha and R. Mihalcea. 2007. Unsupervised graphbased word sense disambiguation using measures of word semantic similarity. In Proceedings of the IEEE International Conference on Semantic Computing (ICSC 2007), Irvine, CA.
26. M. Steyvers and J.B. Tenenbaum. 2005. Graph theoretic analyses of semantic networks: Small worlds in semantic networks. Cognitive Science, 29:41–78.
27. Kristina Toutanova, Christopher D. Manning, and Andrew Y. Ng. 2004. Learning random walk models for inducing word dependency distributions. In ICML '04: Proceedings of the twenty-first international conference on Machine learning, page 103, New York, NY, USA.
28. D.Widdows and B. Dorow. 2002. A graph model for unsupervised lexical acquisition. In Proceedings of the 19th International Conference on Computational Linguistics, Taipei.
29. J.Wiebe and R. Mihalcea. 2006. Word sense and subjectivity. In Proceedings of the AnnualMeeting of the Association for Computational Linguistics, Sydney, Australia.
30. Xiaojin Zhu and Zoubin Ghahramani. 2002. Learning from labeled and unlabeled data with label propagation.Technical Report CMU-CALD-02-107, Carnegie Mellon University.
31. Xiaojin Zhu and John Lafferty. 2005. Harmonic mixtures: Combining mixture models and graph-based methods for inductive and scalable semi-supervised learning. In Saso Dzeroski, Luc De Raedt, and Stefan Wrobel, editors, Proceedings of the Twenty-Second International Conference on Machine Learning (ICML '05), Bonn, Germany, August 7-11,. ACM Press.