استفاده از زنجیره های واژگانی برای استخراج لغات کلیدی

چکیده

لغات کلیدی را می توان به صورت نسخه هایی متراکم از اسناد و اشکال کوتاهی ازچکیده در نظر گرفت. در این مقاله، مسئله استخراج خودکار لغات کلیدی از اسناد به صورت یک کار یادگیری نظارت شده در نظر گرفته می شود. یک زنجیره واژگا نی به صورت مجموعه ای از کلمات مرتبط از نظر معنایی از یک متن بوده و می توان گفت که زنجیره واژگا نی بیانگر محتوی معنایی یک بخش از متن است. اگرچه زنجیره واژگا نی به طور گسترده ای در خلاصه سازی متن مورد استفاده قرار گرفته است، کاربرد آن ها برای مسئله استخراج کلیدی به طور کامل بررسی نشده است. در این مقاله، یک روش استخراج لغات کلیدی که از زنجیره واژگا نی استفاده می کند توصیف شده و نتایج بدست می اید.

لغات کلیدی: استخراج لغات کلیدی، زنجیره واژگانی، پردازش زبان طبیعی، یادگیری ماشینی

1-مقدمه

لغات کلیدی را می توان به صورت خلاصه های کوتاهی از یک متن در نظر گرفت. از این روی می توان آن ها را به صورت مجموعه ای از عبارات در نظر گرفتن که بیشتر متن را پوشش می دهند. اگرچه یک خلاصه ای از متن ا قادر است تا اطلاعاتی در مورد متن بیشتر از لغات کلیدی متن ارایه کند، با این حال این خلاصه برای برخی از کاربرد ها به دلیل ساختار پیچیده جملات مناسب نیست. لغات کلیدی، جایگزینی برای خلاصه نمی باشند بلکه به صورت خلاصه های جایگزینی هستند گه توسط برخی برنامه های دیگر به آسانی مورد استفاده قرار می گیرند. از آن جا که آن ها مدل های فشرده تری از متن اصلی هستند، با این حال امکان استفاده از آن ها در برنامه های مختلف نظیر نمایه بندی در موتور های جست و جو و یا طبقه بندی متن وجود دارد.

لغات کلیدی خوانندگان را قادر به تصمیم گیری در مورد این موضوع می کنند که آیا یک سند برای آن ها مناسب است یا خیر. آن ها را می توان به عنوان شاخص های تشابه کم هزینه بین اسناد مورد استفاده قرار داد. با در نظر

گرفتن این که تخصیص لغات کلیدی به اسناد سخت است، می توان این کار را با یادگیری ماشینی و پردازش زبان طبیعی به صورت خودکار دراورد.

محققان می توانند عبارات کلیدی را برای اسناد خود تخصیص دهند و این عبارات کلیدی می توانند درون متن باشند یا نباشند. در استخراج عبارات کلیدی خودکار، شاخص ترین عبارات در یک سند به صورت عبارت کلیدی برای آن سند استفاده می شوند. از این روی الگوریتم های استخراج عبارات کلیدی خودکار با عبارات ظاهرا شده در متن محدود می شوند. شکل کلی تر استخراج عبارت کلیدی، تولید عبارت کلیدی است که عبارات را از سند انتخاب نمی کند با این حال عبارات کلیدی را برای سند تولید کرده و تخصیص می دهد. در این مقاله ما به جای عبارات کلیدی بر لغات کلیدی تاکید دارم تا اثبات شود که عبارت کلیدی می تواند متشکل از بیش از یک کلمه باشد و ما تنها لغات کلیدی را استخراج می کنیم.

ما باور داریم که لغات کلیدی یک متن بایستی از نظر معنایی مشابه با لغات متن باشد. تعداد کلمات و تعداد روابط معنایی میان آن ها می تواند برای زنجیره واژگانی متفاوت باشد. پوشش و اندازه یک زنجیره واژگانی نشان می دهد که به چه میزان زنجیره واژگانی نشان دهنده محتوی معنایی متن است. از این روی، ما باور داریم که لغات کلیدی که نشان دهنده محتوی معنایی متن است بایستی از کلمات یک زنجیره واژگانی انتخاب شود که بیشتر محتوی معنایی متن را در بر می گیرد. در این مقاله، یک روش استخراج کلیدی را ارایه می کنیم که از ویژگی های مبتنی بر زنجیره های واژگانی در انتخاب لغات کلیدی برای یک متن استفاده می کند.

استخراج لغات کلیدی ارتباط نزدیکی با خلاصه سازی خودکار متن دارد. در خلاصه سازی متن، شاخص ترین جملات برای نمایش متن استخراج می شوند. در استخراج لغات کلیدی شاخص ترین لغات کلیدی برای نشان دادن متن استخراج می شوند. در هر دوی این مسائل، ویژگی هایی نظیر فراوانی های لغات، عبارات کلیدی، موقعیت در متن، زنجیره های واژگانی و ساختار گفتمان برای کشف الگو استفاده می شوند. در این مقاله، هدف ما کشف اثر زنجیره های واژگانی در استخراج لغات کلیدی می باشد به خصوص زمانی که مسئله به صورت یک کار یادگیری ماشینی نظارت شده در نظر گرفته شود. این یادگیری از ویژگی های مبتنی بر زنجیره های واژگانی کلمات استفاده

می کند. چون می توان زنجیره های واژگانی را برای کلمات تنها با استفاده از انتولوژی ورد نت ایجاد کرد، ما بر مسئله استخراج لغات کلیدی به جای استخراج عبارات کلیدی تاکید میکنیم.

اگرچه ما کلاسیفایر های مختلف نظیر نیو بایس را ازمایش کرده ایم، نتایج بهتری با الگوریتم القای ددرخت تصمیم گیری C4.5 بدست امد. به همین دلیلفما از C4.5 برای نشان دادن مسئله استخراج لغات گلیدی به صورت یک وظیفه یادگیری استفاده کردیم. ما از C4.5 با دو مجموعه از ویژگی ها استفاده کردیم. در سیستم معیار، تنها ویژگی های متن استفاده شد. در دومین مورد، C4.5 با ویژگی های بر اساس زنجیره های واژگانی علاوه بر ویژگی های مورد استفاده در سیستم معیار استفاده شد. سپس نتایج دو نسخه مقایسه شد. نتایج بهتر زمانی حاصل شد که ویژگی های مبتنی بر زنجیره های واژگانی استفاده شد.

ما در ابتدا به مرور منابعی در خصوص استخراج لغات کلیدی و زنجیره های واژگانی در بخش دوم می پردازیم. زنجیره های واژگانی و ایجاد زنجیره های واژگانی در بخش 3 ارایه شده اند. در بخش چهارم، ویژگی های مبتنی بر زنجیره واژگانی در سیستم استخراج لغات کلیدی استفاده می شوند. و به این ترتیب در مورد نتایج روش استخراج لغات کلیدی در بخش 5 صحبت می کنیم. در نهایت در بخش 6، نتیجه گیری ارایه می شود.

2-مرور منابع

سیستم خلاصه سازی سعی می کند نا اطلاعات مهمی را شناسایی کند که در خلاصه بسیار مهم است. برای شناسایی موضوعات و ایده های مهم در سند، سیستم خلاصه سازی سعی در استفاده از علایم و نشانه ها در سند دارد. این علایم و نشانه ها شامل ویژگی های سطح جملات نظیر موقعیت و طول جمله هستند و ویژگی های سطح کلمات نظیر فراوانی کلمات و موقعیت کلمات در نظر گرفته شد. روش های خلاصه سازی که از فنون یادگیر ینظارت شده استفاده می کنند قادر به شناسایی جملات حاوی اطلاعات مهم با استفاده از نشانه ها به صورت شاخص اهمیت بوده و این علایم را می توان از داده های اموزشی استخراج کرد. برای مثال، کاپک، پدرسون و چن 1995، توسفل و مونز 1997 از ویژگی های سطح جمله برای طبقه بندی جملات مهم استفاده می کند. کوپک در عین حال از اطلاعات فراوانی کلمات برای کار طبقه بندی استفاده می کند. آن ها از هیچ ویژگی ای بر مبنای

معنای کلمات استفاده نکرده اند. در سیستم ما، ما از ویژگی های معنایی بر اساس زنجیره ای برای تعیین مهم ترین لغات کلیدی در سند استفاده می کنیم.

استخراج لغات کلیدی به صورت یک الگوریتم یادگیری نظارت شده در مطالعات فرانک، پتیر، ویتن، گوتوین و نویل مانینگ 1999، قلبوم 1998فهولت 2004، تورنی 2000 مطالعه شده است. الگوریتم درخت تصمیم C4,5 و الگوریتم کلی برای استخراج لغات کلیدی استفاده شد(ترنی 2000). این الگوریتم از دو ویژگی برای این الکوریتم ها استفاده می کند. موقعیت لغات در متن و فراوانی کلمات. ترنی نشان داده است که استفاده از درخت تصمیم در C4.5 موجب بهبود عملکرد C4,5 برای استخراچ لغات کلیدی می شود. تاب تناسب الگوریتم ژنتیکی GenEx صحت استخراج کننده است و جمعیت متشکل از مقادیر پارامتر است. با این حال GenEx از نظر محاسباتی در زمان اموزش هزینه بردار است. ترنر نشان می دهد که GenEx قابل تعمیم به حوزه های مختلف است و از این روی لزوم اموزش را برای حوزه های مختلف رد کرده است.

کی(فرانگ و همکاران 1999) یک الگوریتم دیگر است که استخراج لغات کلیدی را به صورت یک کار یادگیری نظارت شده در نظر می گیرد. ایشان از TFxIDF(فراوانی کلمه ضرب در فراوانی اسناد معکوس) و موقعیت نرمال کلمات به صورت ویژگی استفاده می کند. TFxIDF فراوانی کلمات در اسناد نرمال شده توسط فراوانی دامنه است و به صورت ویژگی خاص دامنه مطرح شده است. کی نتایج متناقضی را با GenEx را در اموزش عمومی بدست اورد. کی بسیار سریع تر از GenEx بوده و هنگام اموزش با داده های وابسته به حوزه عملکرد بهتری دارد. کی از الگوریتم بیزی ساده برای یادگیری جهت طبقه بندی عبارات استفاده می کند. اگرچه کی برای شناسایی عبارات از پارسر استفاده نمی کند، همه توالی کلمات تا سه کلمه به صورت عبارات احتمالی در نظر گرفته می شود. برخی محدودیت ها برای فیلتر برخی از توالی های کلمات استفاده می شود. برای مثال، علایم نقطه ای به صورت لغات در نظر گرفته نمی شوند و عبارات با لغات توقف شروع نمی شوند. در سیستم، ما لغات کلیدی را استخراج کرده و همه اسامی در اسناد به صورت لغات کلیدی احتمالی در نظر گرفته می شود.

چون استخراج لغات کلیدی و خلاصه سازی متن خودکار، از کار های بسیار مشابه می باشند، برخی از فنون مورد استفاده در الگوریتم های خلاصه سازی متن را می توان در الگوریتم های استخراج لغات کلیدی استفاده کرد.

بارازیلی و الدهاد 1997 نشان داده اند که ویژگی های مبتنی بر زنجیره های واژگانی، ویژگی های خوبی برای خلاصه سازی متن هستند. زنجیره های واژگانی را می توان برای تحلیل انسجام واژگانی استفاده کرد(موریس و هیرت 1991). هر متن دارای برخی از سطوح انسجام واژگانی است زیرا متشکل از روابط معنایی بین لغات است. زنجیره های واژگانی با استفاده از روابط معنایی بین کلمات ساخته می شوند. زنجیره های واژگانی برای خلاصه سازی متن با محققان دیگر استفاده می شوند. سیلبر و مک کوی(2000) روش کارامدی را برای ایجاد زنجیره های واژگانی توصیف می کنند. این کار موجب افزایش استفاده از زنجیره های واژگانی در زمینه های مختلف می شود زیرا یک روش کارامد را برای ایجاد زنجیره های واژگانی ارایه می کند که موسوم به مسئله نمایی است. زنجیره های واژگانی در مسائل NLP مختلف نظیر ابهامزدایی، خلاصه متن، تقسیم بندی متن و ردیابی موضوع استفاده شده اند.

هم چنین ما مسئله استخراج لغات را به صورت یک وظیفه یادگیری نظارت شده در نظر می گیریم و از الگوریتم درخت تصمیم C4.5 به عنوان دسته بند در سیستم استخراج لغات کلیدی استفاده می کنیم. اگرچه زنجیره های واژگانی در خلاصه سازی متن استفاده می شوند، با این حال اولین بار زنجیره های واژگانی را در استخراج لغات کلیدی پیشنهاد می کنیم. چون زنجیره های واژگانی کلمات بیانگر یک انسجام واژکانی در متن هستند، آن ها اطلاعات مهمی را در مورد محتوی معنایی متن ارایه می کنند. در این فصل، نشان می دهیم که زنجیره های واژگانی نقش مهمی در انتخاب لغات کلیدی ایفا می کنند که به طور معنایی متن را تعریف می کند. ما از رویکرد سیلبر استفاده می کنیم به خصوص زمانی که زنجیره های واژگانی برای متن معین ایجاد شود.

3- ایجاد زنجیره های واژگانی

انسجام واژگانی(موریس و هیرست 1991) در میان یک توالی ای از کلمات مرتبط ایجاد می شود و زنجیره های واژگانی یک متن ساختار منسجم واژگانی یک متن را پوشش می دهد. زنجیره های واژگانی با استفاده از روابط میان کلمات تعریف می شود. برای ایجاد زنجیره های واژگانی، روابط معنایی و حس کلمات بین کلمات مشخص است. وردنت (فلبوم 1998) یک دیتابیسی است که این دانش را ارایه می کند و ما از وردنت در الکوریتم ایجاد زنجیره های واژگانی استفاده می کنیم.

ما از مجموعه های مترادف از درختان ورد نت برای یافتن روابط بین دو حس کلمه در ایجاد زنجیره های واژگانی استفاده می کنیم. ایجاد زنجیره های واژگانی یک روش جامع است. دلیل این است که کلمه داری معانی زایدی است. حس درک صحیح کلمه انتخاب شده است. به عبارت دیگر، بایستی کلمات را ابهام زدایی کنیم.

برازیلی و الهداد(1997) و برایلی (1997) یک الکوریتم جامع را برای ابهام زدایی کلمات ضمن ایجاد زنجیره های واژگانی ارایه کرده اند. در متن معنی دار و منسجم، حس کلمه مربوط به تعداد کلمات بیشتری است که بایستی در نظر گرفته شوند. برازیلی پیشنهاد می کند تا همه زنجیره های زنجیره های واژگانی احتمالی برای متن و انتخاب بزر گترین تفسیر متن ایجاد شود. با این حال محاسبه سخت است زیرا تعداد تفاسیر به طور معنی داری با هر کلمه افزایش می یابد. از سوی دیگر، استفاده از رویکرد حریص منجر به سو تفسیر می شود. برازیلی از تفاسیر استفاده می کند زیرا با افزایش تعداد تفاسیر، این فرایند از نظر محاسباتی سخت خواهد بود. روش دیگر مورد استفاده، بخش های متنی برای کاهش پیچیدگی الکوریتم است.سیلبر و مک کوی(2002)ف روش کارامد دیگر را توصیف می کند که زنجیره های واژگانی را با ابهام زدایی کلمات ارزیابی می کند. سیلبر فر زنجیره هایی را برای یک متن با یافتن روابط بین کلمات ایجاد کرده است با این حال زنجیره های واژگانی را ایجاد نکرده است. برای هر کلمه، حس کلمه بیش از رابطه با کلمات انتخاب شده دیگر است. بعد از ابهام زدایی همه کلمات در متن، زنجیره های واژگانی ساخته می شود. در هر دو الگوریتم های برازیلی و سیلبر، زنجیره های واژگانی در دو بخش متفاوت ترکیب می شود به خصوص اگر تکرار کلمات یا مترادف آن ها وجود داشته باشد.

در صورتی که تنها یک زنجیره واژگانی برای یک متن وجود داشته باشد این بدین معنی است که همه کلمات موجود در متن ارتباط دارند. به طور کلیف متن می تواند با بیش از یک زنجیره های واژگانی ارتباط داشته باشد. در این رابطه کلمات متن بر اساس ارتباط باز یر مجموعه ها گروه بندی می شوند. ما از رویکرد سیلبر و مک کوی با تغییرات حداقل استفاده کرده این. در مطالعه سیلبر و مک کوی 2002، دیتابیس ورد نت برای دسترسی کارامد تر نمایه بندی شد. این روش به ما امکان بررسی روابط را در زمان خطی داده و ما را قادر به یافتن رابطه ای بین دو کلمه می کند. قبل از ایجاد زنجیره های واژگانی، ما اقدام به شناسایی اسامی در متن می کنیم. ما تنها از روابط بین اسامی استفاده کردیم زیرا اطلاعات بیشتری در مورد موضوع ارایه می کنند و لغات کلیدی در متن به صورت

اسم وجود دارند. بخشی از تگر گفتاری که برای شناسایی اسامی در متن استفاده می شود، ماکستت تاگر است(تاتونوا و مانینک 2000، تاتونوا، کلین و سنیگر 2003).

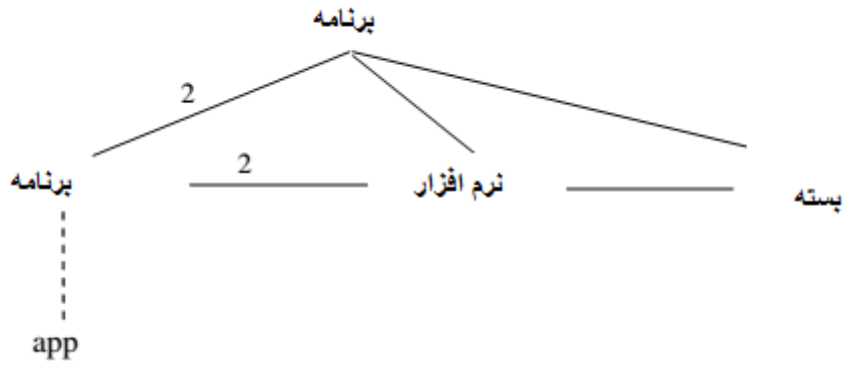
روابط وردنت مورد استفاده در نسخه سازنده زنجیره های واژگانی به صورت مترادف می باشند این الگوریتم از روابط یکسان به جز مرونیم استفاده می کند و نیز مترادفات در ساخت زنجیره های واژگانی استفاده می شوند. این سیستم دقیقا از ویژگی های یکسان مورد استفاده بهره می برد.

اگرچه ما از زنجیره های واژگانی برای استخراج لغات کلیدی از اسناد استفاده می کنیم، زنجیره های واژگانی را می توان در ابهام زدایی کلمات استفاده کرد. شاخص عملکرد سازنده زنجیره های واژگانی برای ابهانم زدایی استفاده نشد. با این حال، چون زنجیره های واژگانی مشابه با رویکرد سیبلر است(سیبلر و مک کوی 2002)، و ویژگی های وردنت مورد استفاده مشابه با ویژگی های برازیلی استف انتظار می رود که صحت سیستم مشابه با نتایج سیستم باشد. گالی و مک کون (2003) کزارش کرده است که الگوریتم سیلبر دارای صحت 53 درصد برای مسئله WSD می باشد و الگوریتم برازیلای دارای صحت 57 درصدی است.

انگلیسی یک زبان مولد برای اسامی ترکیبی است. برازیالیف ترکیبات اسمی را در متن جست و جو کرده و در ورد نت نیز به دنبال عبارت می گردد. در صورتی که کلمه در ورد نت ظاهر نشود، از یک اسم در عبارت شناسایی شده استفاده می کند. سیلبر، ترکیبات اسمی را مدیریت نمی کند. به همین دلیل، تنها بر استخراج عبارات کلیدی به جای استخراج عبارت های کلیدی تاکید داریم.

هر گره در زنجیره های واژگانی یک مفهومی از کلمه است و هر لینک رابطی بین دو کلمه است. برای مثال، زنجیره های واژگانیساده در شکل 1 روابط بین معانی کلمات انتخاب شده را در زنجیره های واژگانی نشان می دهد. در شکل 1، خطوط پر رئنگ روابط هایپو نایم و خط چین ها روابط ساینونایم هستند.

شکل 1 : زنجیره واژگانی ساده



4- استخراج ویژگی

هر کلمه در یک متن مربوط به ویژگی های مختلف است. ما از این ویژگی ها با الگوریتم القای درخت تصمیمC4.5 استفاده می کنیم. ویژگی های مورد استفاده در سیستم معیار، اولین تکرار از کلمات در متن، تعداد فراوانی کلمات در متن و اخرین فراوانی کلمات هستند. دو ویژکی اول توسط ترنی 2000 استفاده شده است. اگرچه ترنی از برخی ویژگی های دیگر استفاده می کنند،ما تنها این ویژگی ها را ترجیح می دهیم که با عبارت های کلیدی ارتباط دارند. استخراج این سه ویژگی ساده است. اخرین و اولین ویژگی ها بیانگر ویژگی های طول زمان وجود کلمات در متن است. این سه ویژگی مربوط به مدل سازی نمادین کلمه است. وقتی C4.5 با این سه ویژگی استفاده شود، سیستم معیار ما را تشکیل می دهد. به منظور مشاهده اثرات ویژگی های مبتنی بر زنجیره های واژگانی، نتایج سیستم معیار با سیستم با ویژگی های زنجیره های واژگانی مقایسه می شود.

علاوه بر این سه ویژگی در سیستم معیار، ما به بررسی ویژگی های مختلف بر گرفته از زنجیره های واژگانی در متن می پردازیم. بعد از بررسی ویزگی های مختلف بر اساس زنجیره های واژگانی، از چهار ویژگی زیر بر اساس زنجیره های واژگانی در سیستم اصلی استفاده می کنیم.

امتیاز زنجیره های واژگانی از یک کلمه: کلمه می تواند یک عدد بزرگ تر از یک باشد که در یک متن با معانی مختلف دیده می شود. زنجیره وازگانی معرف برای این کلمه انتخاب شد. به عبارت دیگر، بعد از امتیاز بندی هر زنجیره زنجیره های واژگانی کلمه، امتیاز زنجیره با بالاترین امتیاز به صورت یک زنجیره واژگانی از کلمه انتخاب می شود. امتیاز زنجیره واژگانی بستگی به روابط موجود در زنجیره واژگانی دارد. برای هر رابطه بین کلمات، اوزان به طور غیر مستقیم تخصیص داده می شوند. این اوزان برای روابط را می توان در جدول 1 مشاهده کرد.

این ویژگی با امتیاز بندی همه روابط زنجیره های واژگانی محاسبه می شود . برای مثال، اگر زنجیره واژگانی در شکل 1 به صورت زنجیره های واژگانی با ماکزیمم امتیاز از یک کلمه برنامه باشد، امتیاز زنجیره های واژگانی به صورت امتیاز زنجیره های واژگانی از کلمه برنامه استفاده می شود.

امتیاز زنجیره واژگانی مستقیم یک کلمه: این ویژگی با رتبه بندی روابط مربوط به کلمه محاسبه می شود. به عبارت دیگر می توان از قوی های متصل به هر کلمه استفاده کرد. امتیاز زنجیره های واژگانی مستقیم برای کلمه برنامه در زنجیره های واژگانی برابر با 28 است زیرا این ها چهار قوس هیپونیم مرتبط با کلمه برنامه در شکل 1 هستند.

امتیاز طیف زنجیره های واژگانی از یک کلمه: امتیاز طیف زنجیره های واژگانی از یک کلمه بستگی به بخشی از متن دارد که با زنجیره واژگانی پوشش داده می شود. بخش پوشش دهی متن فاصله بین اولین بخش از عضو زنجیره واژگانیاست که در فراوانی اول و اخر قرار دارد. زنجیره واژگانی با تفاضل بین دو موقعیت محاسبه می شود. برای مثال، به منظور ارزیابی امتیاز طیفی زنجیره واژگانی در شکل 1، همه موقعیت های لغات در زنجیره واژگانی در نظر گرفته می شود. در صورتی که P first اولین و plast اخرین موقعیت در میان این موقعیت ها باشد، امتیاز طیف زنجیره برابر با plast – pfirst است. امتیاز طیف زنجیره های واژگانی از یک کلمه یک امتیاز از زنجیره واژگانی با ماکزیمم امتیاز در برگیرنده یک کلمه است.

امتیاز طیف زنجیره های واژگانی مستقیم از یک کلمه: برای ارزیابی امتیاز طیف زنجیره های واژگانی مستقیم از یک کلمه، ما از امتیازات همه زنجیره های واژگانی حائی یک کلمه را ارزیابی کردیم. امتیاز زنجیره های واژگانی با ماکزیمم امتیاز یک امتیاز طیف زنجیره های واژگانی مستقیم از یک کلمه می باشد. امتیاز طیف زنجیره های واژگانی مستقیم از یک کلمه همانند امتیاز طیف زنجیره واژگانی محاسبه می شود و در آن تنها کلمات ارتباط مستقیم با زنجیره واژگانی دارند. برای مثال، زنجیره واژگانی در شکل 1 یک زنجیره واژگانی با ماکزیمم امتیاز مستقیم برای کلمه برنامه است. برای ارزیابی یک امتیاز طیف زنجیره های واژگانی مستقیم در شکل 1، تنها بخشی از متن پوشش داده شده توسط کلمه برنامه، نرم افزار و متن را در نظر می گیریم که با کلمه program ارتباط دارد. لازم به ذکر است که این ویژگی ها را در صورتی می توان استخراج کرد که کلمه در یک ورد نت قرار داشته

باشد. در صورتی که کلمه در ورد نت ظاهر نشود، این ویژگی ها به صورت ویژگی های حل نشده باقی می مانند. استخراج این ویژگی ها برای اسم سخت است زیرا ورد نت فاقداسامی ترکیبی است.

5- یادگیری برای استخراج لغات کلیدی

ما الگوریتم القای درخت تصمیم C4.5 را با ویژگی های مختلف برای استخراج لغات کلیدی در اسناد بررسی کردیم.الگوریتم یادگیری برای یافتن لغات کلیدی در عبارات کلیدی با اشکال صحیح اموزش داده می شوند. برای مثال اگر " زمان واکنش" یک عبارت کلیدی تعیین شده توسط نویسنده باشد، الگوریتم یادگیری برای استخراج واکنش و زمان به صورت لغات کلیدی اموزش داده می شود. این متفاوت از کار های فرانک و همکاران 1999 است زیرا آن ها برای استخراج زمان واکنش دقیق در الگوریتم ها اموزش می بینند.در مرحله ازمون، الگوریتم در صورتی به صورت صحیح است که لغات کلیدی به شکل یک عبارت کلیدی کاربر محور باشد.

تغییرات C4.5 در منابع وجود دارد. از نظر ازمایشی، معمولا C4.5 نتایج بهتری را برای مسئله استخراج لغات کلیدی ارایه می کند. به همین دلیل، نسخه C4.5 از درخت تصمیم استفاده می کند. فرایندب سته بندی، فرایند طبقه بندی نمونه ها با دسته بند های مختلف است. در این روش، نمونه با دسته بند های مختلف طبقه بندی می شود و احتمال طبقه بندی متوسط برای طبقه بندی نمونه استفاده می شود. بسته بندی موجب کاهش واریانس، افزایش صحت می شود. بر این اساس، تارنی(2000) گزارش کرده است که بسته بندی موجب بهبود عملکرد الگوریتم درخت تصمیم برای استخراج عبارت کلیدی می شود.

برای ارزیابی اثر ویژگی های زنجیره واژگانی، از دو مجموعه از ویژگی های متفاوت استفاده شد. در سیستم معیار، لغات گلیدی با استفاده از سه ویژگی استخراج شده اند.

موقعیت وقوع اول: موقعیت وقوع اول کلمه در متن

فراوانی کلمه: فراوانی کلمه

موقعیت وقوع احر: موقعیت اخر کلمه در متن

سپس ما نتایج را با نتایج بدست امده با افزودن .ویژگی های زنجیره های واژگانی قیاس کردیم. از چهار ویژگی زنجیره های واژگانی در بخش قبلی استفاده شد.

امتیاز زنجیره های واژگانی یک کلمه

امتیاز زنجیره واژگانی مستقیم یک کلمه

امتیاز طیف زنجیره واژگانی یک کلمه

امتیاز مستقیم طیف زنجیره واژگانی یک کلمه

از این روی الگوریتم یادگیری اصلی از هفت ویژگی استفاده می کند. الگوریتم یادگیری با زیر مجموعه های مختلف این هفت ویزگی بررسی شد و نتایج در این بخش ارایه شد.

یکی از مجموعه های ما توسط فرانک و همکاران 1999، ترنی 2000 بررسی شده است. این مجموعه متشکل از 75 مجله ژورتالی است و 50 مورد از این مقالات برای اموزش 25 مورد برای ازمون استفاده شد. متون مجموعه متون کاملی هستند. 1، 5، 10 و 15 لغات کلیدی برای هر سند استخراج شد. همه ویژگی های زنجیره های واژگانی موجب بهبود دقت شدند.

جدول 2 نتایج ازمایش را نشان می دهد. ردیف اول جدول نتایج را در زمانی نشان می دهد که هیچ ویژگی زنجیره های واژگانی استفاده نشود و ردیف دوم نتایج را در زمانی نشان می دهد که همه هفت ویژگی از جمله چهار ویژگی زنجیره های واژگانی استفاده شود. دقت دسته بند به طور معنی داری با ویژگی های زنجیره های واژگانی افزایش یافت. برای مثال، تنها 17 درصد دقت زمانی حاصل شد که 5 کلمه کلیدی استخراج شد. از سوی دیگر،45درصد زمانی حاصل شد که از ویژگی های زنجیره های واژگانی استفاده شد. این نتایج نشان می دهد که ویژگی های زنجیره زنجیره های واژگانی موجب بهبود دقت به طور معنی دار می شود.

جدول 2: نتایج مربوط به دقت: اثرات ویژگی های زنجیره های واژگانی با مجموعه از متون کامل

تعداد لغات کلیدی در هر سند 1 5 10 15

دقت برای سیستم معیار 9 17 18 14
دقت برای سیستم با همه هفت ویژگی 64 45 30 26

اگرچه یک مجموعه یکسان توسط تارنی 2000 استفاده شده است، ما قادر به مقایسه مستقیم نتایج ترنی د با نتایج جدول 2 نمی باشیم. سیستم ترنی مشابه با سیتسم معیار است ولی از ویژگی های بیشتری نسبت به سیستم معیار استفاده می کنند. این بهتر از نتایج سیستم معیار است ولی بدتر از نتایج سیستم اصلی است.

جدول 3، نتایج دقیق را برای دو مجموعه متفاوت از ویژگی های زنجیره های واژگانی نشان می دهد. اگرچه همه ویژگی های زنجیره های واژگانی موجب بهبود دقت نتایج می شود، ویژگی های طیف بیشتر بهبود می یابند.

برخی از ویژگی های مبتنی بر زنجیره های واژگانی با چهار ویژگی اصلی در این مقاله بررسی شدند. مشاهده شد که این چهار ویژگی با ویژگی های سیستم معیار، بهترین ترکیب در میان ویژگی هایی هستند که برای استخراج لغات کلیدی استفاده می شوند. اگرچه سایر ویژگی های زنجیره های واژگانی وجود دارند که منعکس کننده انسجام واژگانی متن هستند، ما به دقت بهتری از زنجیره های واژگانی دست پیدا نکردیم.

بر طبق نتایج جدول 2 و 3، مقدار دقت زمانی کاهش می یابد که تعداد لغات کلیدی افزایش یابد. دو دلیل اصلی برای این کاهش دقت وجود دارد. در ابتدا،تعداد متوسط لغات کلیدی در غبارت کلیدی تعیین شده 5.8 عبارت کلیدی به ازای هر سند در مجموعه است و برخی از لغات کلیدی در سند ظاهر نمی شوند. وقتی که تعداد لغات کلیدی استخراج شده بیش از 5 باشد، دقت کاهش می یابد زیرا مجموعه ای از لغات کلیدی استخراج شده به طور معرف دارای لغات لکیدی ای هستند که توسط محقق تعیین نشده است. دوما، لغات کلیدی استخراج شده بسته به احتمالات تعیین شده توسط دسته بند انتخاب می شود. اولین لغت کلیدی انتخاب شده دارای بالاترین احتمال است و احتمالات زمانی کاهش می یابد که به سمت پایین لیست برویم. وقتی تعداد لغات کلیدی افزایش می یابد، لغات کلیدی با احتمالات پایین تر در مجموعه نتایج قرار می گیرد و این تعداد حذفیات را افزایش می دهد. این دو دلیل منجر به کاهش دقت شده اند.

مجموعه دیگر آزمایش شده، مجموعه مورد استفاده در (ویتن، پینتر، فرانگ، کارتوین و نویل مانینگ 1999) است. این مجموعه متشکل از 155 چکیده است. ما با 110 مورد از این چکیده اموزش دیده و 45 مورد را ازمایش کردیم. نتایج در جدول 4 نشان داده شده است. اگرچه زنجیره های واژگانی نتایج بهتری را ارایه می کنند، این اثر متناسب با ازمایش اصلی نیست. بر طبق نتایج جدول 4، دقت تنها 4 درصد در استخراج 5 لغت کلیدی در مجموعه ای از چکیده ها بهبود می یابد. از سوی دیگر، دقت تا 28 درصد بهبود می یابد. که در جدول 2 نیز نشان داده شده است. این بدین معنی است که ویژگی های مبتنی بر زنجیره های وا ژگانی موجب بهبود عملکرد در استخراج لغات کلیدی از همه متون می شوند در حالی که بهبود عملکرد در استخراج لغات کلیدی از چکیده ها کم تر معنی دار است. دلیل این مشاهده این است که نشانه های انسجام واژکانی در متون با طول کوتاه تر، کم تر است.

جدول 3: نتایج دقت: استفاده از ویژگی های زنجیره واژگانی با مجموعه ای از متون کامل

تعداد لغات کلیدی در هر سند	1	5	10	15
دقت: تنها زنجیره واژگانی و ویژگی های زنجیره واژکانی مستقینم	27	28	22	17
دقت: تنها طیف زنجیره واژگانی و ویژگی های طیف واژکانی مستقیم	64	30	22	20

جدول 4: نتایج دقت: اثرات ویژگی های زنجیره واژگانی با مجموعه ای از چکیده ها

تعداد لغات کلیدی در هر سند 1 5 10 15
دقت برای سیستم معیار 20 16 13 14
دقت برای سیستم با همه هفت ویژگی 27 20 17 16

زنجیره ها زمانی ضعیف تر هستند که به طور گسترده در متون با طول کوتاه استفاده نشوند. در صورتی که متون طول کامل داشته باشند، یک مفهمم با استفاده از چند کلمه با تشابه در نظر گرفته می شود. در صورتی که لغات مشابه از نظر معنایی وجود داشته باشد، زنجیره های واژگانی تست قوی تر خواهد بود زیرا روابط بیشتری بین کلمات وجود دارد. از این روی ویژگی های مبتنی بر زنجیره واژکانی نقش مهمی در انتخاب لغات کلیدی دارند.

چکیده سند نمونه در جدول 5 نشان داده شده است. عبارت کلیدی در جدول 6 دیده می شود. پنج لغت کلیدی استخراج شده برای الکوریتم از متن چکیده در جدول 7 نشان داده شده است. اگرچه کلمه پارکینسون به صورت انتزاعی نوشته نشده است ولی یک نتیجه صحیح نیست. این بهترین نمونه برای ویژگی های واژگانی است. سیستم اصلی با هفت ویژگی دارای دقت 45 درصدی برای کل متن و 20 درصد برای 5 لغت کلیدی است.

جدول 5: چکیده یک سند نمونه

چکیده : افزونگی سیستم کنترل موتور، گزینه های ساختار کنترل مرکزی را برای حل مسائل روزمره موتور ارایه می کند. انتخاب الگو های کنترل بر اساس اولویت هستند. الگو های حرکتی مشاهده شده در بزرگ سالان منعکس کننده این اولویت ها هستند. فرض بر این است که در شرایط خاص، اختلال در ادراک محیط و تصمیم گیری، تغییرات در سیستم عصبی و تغییرات ساختاری در نظر گرفته می شود. مجموعه جدیدی از اولویت ها منعکس کننده وضعیت فعلی سیستم بوده و منجر به الکوهای مختلفی از جنبش های اختیاری می شوند. در این شرایط الگوی حرکتی به صورت پاتولوژیکی نیست بلکه تطبیقی است.

جدول 6: لغات کلیدی سند نمونه

حرکت اختیاری

کنترل حرکتی

اختلالات حرکتی

هماهنگ سازی

موقعیت

پیش برنامه نویسی

بیماری پارکینسون

سندرم داون

جدول 7: لغات کلیدی استخراج شده سند نمونه

حرکتی	سندرم
قاعده	پارکینسون
چکیده	حرکت

کنترل	جابه جایی
مسئله	کنترل
سیستم معیار با سه ویژگی	سیستم با هفت ویژگی

6-نتیجه گیری و کار های اینده

این مقاله به توصیف روش استخراج لغات کلیدی برای بررسی مزیت های استفاده از ویژگی های زنجیره واژگانی در استخراج لغات کلیدی می پردازد. بر اساس نتایج، ویژگی های زنجیره واژگانی موجب بهبود دقت در فرایند استخراج لغات کلیدی می شود.اگرچه زنجیره های واژکانی در حوزه های کاربردی مختلف استفاده شده اند، اولین بار از زنجیره های واژکانی در مسئله استخراج لعات کلیدی استفاده می شود. در این مقاله، ما سعی کرده یم تا لغات کلیدی را استخراج کنیم زیرا ورد نت فاقد تعداد زیادی از عبارات است. ما به بررسی این موضوع می پردازیم که چگونه ابیستی عبارات کلیدی را علاوه بر لغات کلیدی استخراج کرد. برای انجام این کار ما به بررسی شیوه قرار دادن این عبارات در زنجیره های واژگانی می پردازیم. در حقیقت، برخی عبارت ها در ورد نت وجود دارند. برای عبارت هایی که در ورد نت وجود ندارند، رویکرد احتمالی استفاده از اسم عبارت باری نشان دادن آن عبارت و قرار دادن نام در زنجیره های واژگانی است. در حقیقت، برازیلی از روش مشابه در زمان ایجاد زنجیره های واژکانی استفاده می کند. برازیلی از ویژگی های مبتنی بر زنجیره های واژگانی برای انتخاب جملات استفاده می کند. چون قصد ما استفاده از آن ها برای انتخاب عبارات است، مدل زنجیره واژکانی باید دقیق تر باشد. هم چنین ما استفاده از منابع دانشی دیگر نظیر ویکیپدیا برای ارتباط عبارات در زنجیره های واژگانی را در نظر می گیریم.