رویکردی برای خلاصه سازی متن با استفاده از الگوریتم یادگیری عمیق

چکیده

امروزه تحقیقات بسیاری بر روی تلخیص یا خلاصه سازی متن در حال انجام است. به دلیل افزایش اطلاعات در اینترنت، این انواع تحقیقات در حال کسب توجه بیشتری در میان محققان می باشند. خلاصه سازی متن های استخراجی ایجاد یک خلاصه مختصر با استخراج مجموعه مناسبی از جملات از یک سند یا چند سند با یادگیری عمیق می کند. این روش توسط الگوریتم ماشین بولتزمن محدود(RBM) برای کارایی بهتر با حذف جملات افزونه اصلاح شده است. روش فوق متشکل از سه لایه ورودی، مخفی و خروجی است. داده های ورودی توزیع یکنواختی در لایه مخفی برای عملیات دارند. آزمایشات انجام شده و خلاصه ای برای سه سند متفاوت از دامنه دانش متفاوت ارایه شد. مقدار شاخص F، شناساگر و معیاری برای عملکرد روش خلاصه سازی متن است. پاسخ های سه حوزه دانشی متفاوت بر طبق معیار f به ترتیب برای سه مجموعه سند به صورت 1.42،0.85 و 1.97 می باشد.

لغات کلیدی: چند اسنادی، خلاصه، افزونگی، RBM، مجموعه داده های کنفرانس درک اسناد 2002

1-مقدمه

به مدت سالیان متمادی، خلاصه سازی به طور دستی توسط انسان انجام شده است. در حال حاضر، مقدار اطلاعات به تدریج از طریق اینترنت و منابع دیکر در حال افزایش است. برای غلبه بر این مسئله، خلاصه سازی متن برای کاهش انباشت بیش از حد اطلاعات لازم است. خلاصه سازی متن به حفظ داده های متنی با قواعد و مقررات خاص برای استفاده موثر از داده های متنی کمک می کند. برای مثال استخراج خلاصه از یک سند برای استخراج محتوی معینی از اسناد و چند اسناد صورت می گیرد. خلاصه سازی متن مربوط به فرایند دست یابی به یک سند متنی است که محتوا از آن گرفته شده و از این روی محتوی لازم را برای کاربران درشکل کوتاه و به شکلی پذیرا برای رفع نیاز های کاربر فراهم می کند. خلاصه سازی خودکار ارتباط تنگاتنگی با درک متن دارد که می تواند با چالش هایی در ارتباط باشد که شامل تغییرات در فرمت، شگل و ویاریش متن است که موجب افزایش ابهام می

شود(شریف و همکراان 2013). محققان بخش خلاصه سازی متن این مسئله را از جهات بسیاری نظیر پردازش زبان طبیعی(زانگ و همکاران 2011)، آماری(دارلینگ و سانگ(2011) بررسی کرده و یادگیری ماشینی و تحلیل متن از اهمیت ویژه ای برای شناسایی اهداف متن برخوردار است.

تخلیص یا خلاصه سازی متن به دو طریق طبقه بندی می شود خلاصه سازی انتزاعی و خلاصه سازی استخراجی. روش پردازش زبان طبیعی برای تقسیم، کاهش کلمات و تولید خلاصه متنی از خلاصه های غیر انتزاعی استفاده می شود. در حال حاضر NLP، یک روش کم هزینه و فاقد دقت است. خلاصه استخراجی انعطاف پذیر بوده و مصرف زمان کم تری درمقایسه با خلاصه سازی انتزاعی دارد(پاتیل و برادزیل 2007). در خلاصه سازی استخراجی، این موضوع توالی را به شکل ماتریس و بر اساس برخی بردار های ویژگی در نظر می گیرد که استخراج همه جملات مهم و ضروری در آن ها دیده می شود. یک بردار ویژگی، یک بردار N بعدی از ویژگی های عددی است که نشان دهنده برخی اشیا است. هدف اصلی خلاصه سازی متن بر اساس رویکرد استخراج، انتخاب جمله مناسب به ازای ملزومات هر کاربر است.

به طور کلی، خلاصه سازی متن، فرایند کاهش یک محتوی متن به یک نسخه کوتاه تر با حفظ محتوی ثابت آن و انتقال مفهوم مطلوب و واقعی است(مانی 2001 الف و. ب). خلاصه سازی تک اسنادی فرایندی است که تنها به یک سند رسیدگی می کند. خلاصه سازی چند اسنادی روش کوتاه سازی نه تنها یک سند بلکه مجموعه ای از اسناد مربوطه به یک خلاصه می باشد(او و همکاران 2008). این مفهوم ظاهرا ساده است با این حال پیاده سازی آن کمی سخت است. گاهی اوقات این راهبرد قادر به دست یابی به اهداف مطلوب نیست. بیشتر فنون مشابه مورد استفاده در خلاصه سازی نک اسنادی، در خلاصه سازی چند اسنادی استفاده می شود. برخی تفاوت های مشهود وجود دارند: 1- درجه افزونگی موجود در یک گروهی از مقالات موضوعی به طور قابل ملاحظه ای بیش از درجه افزونگی در یک مقاله است زیرا هر مقاله برای تشریح مهم ترین نکات و نیز سوابق مورد نیاز لازم است. از این روی روش های ضد افزونگی نقش مهمی ایفا می کنند. نسبت فشرده سازی به طور قابل ملاحظه ای برای یک مجموعه گسترده از اسناد نسبت به خلاصه اسناد کم تر است. به منظور ارایه حجم زیادی از اطلاعات معنایی، کار خلاصه سازی توسط کنفرانس تحلیل متن معرفی شده است.هدف ان تولید خلاصه معنایی با استفاده از فهرستی از ابعاد مهم است. فهرست ابعاد معرف کننده مهم ترین اطالعات می باشد با این حال این خلاصه شامل سایر حقایقی است

که به صورت مهم در نظر گرفته می شود. به علاوه، خلاصه سازی آپدیت از مجموعه مقالات نیوزوایر برای موضوع فرضی ایجاد می شود که در آن کاربر قبلا مقالات را خوانده است. خلاصه تولید شده توسط ابعاد از پیش تعریف شده برای بهبود کیفیت و خوانایی خلاصه استفاده می شود(کاگلیوانی و بالاسمرانی 2012).

در این مطالعه، ما یک سیستم تخلیص چند اسنادی با استفاده از الگوریتم یادگیری عمیق را توسعه داده ایم که موسوم به ماشین بولتزمن محدود است. این ماشین، یگ الگوریتم پیشرفته بر اساس شبکه عصبی است که کار های لازم را برای تخلیص متنی انجام می دهد. اولا،مراحل پیش پردازش به کار برده می شود و این مراحل شامل 1- بخشی از تگ کردن 2- متوقف کردن فیلترینگ کلمه 3- استریمینگ است. سپس وارد بخش استخراج ویژگی می شویم. در این بخش از متن، خلاصه سازی شامل ویژگی های جملات استخراج شده است. ویژگی های استخراجی شامل موارد زیر هستند: عنوان شباهت، ویژگی موضعی، اصطلاح وزن و مفهوم ویژه. تقریبا همه مدل های خلاصه سازی متن با دو مسئله روبرو هستند که اولی مسئله رتبه بندی و دومی ایجاد زیر مجموعه ای از رتبه بندی می باشد. انواع روش های مختلف برای مسئله رتبه بندی وجود دارند. در این مطالعه ما اقدام به حل مسئله با یافتن ارتباط بین پرس و جوی کاربر و یک جمله ویژه می کنیم. بر این اساس، امتیاز جمله برای هر جمله تولید شده و به ترتیب نزولی مرتب می شود. از این جملات رتبه بندی شده برخی جملات بر طبق نرخ فشرده سازی انتخاب می شوند. به این ترتیب می توان مسئله رتبه بندی را حل کرد. در پایان، از مجموعه داده 2002DUC برای ارزیابی نتایج خلاصه سازی شده بر اساس شاخص های دقت، یاد اوری و F استفاده می کنیم.

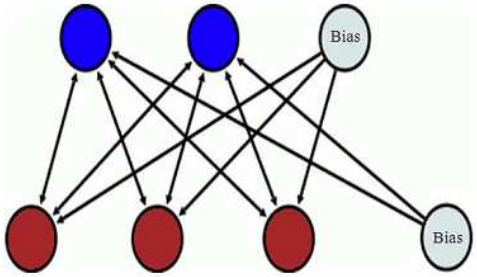
1-1 هدف

امروزه اطلاعات بسیار زیادی از طریق اینترنت و منابع بسیاری قابل دسترس است. برای مدیریت کارامد این داده ها، ما به ابزاری برای استخراج مجموعه مناسبی از جملات از اسناد مربوطه نیاز داریم. خلاصه سازی متن، برای دست یابی به اطلاعات ضمن رسیدگی به مجموعه زیادی از اسناد لازم است. با WWW، اطلاعات به یک بخش لاینفگ از زندگی ما تبدیل شده است. برای یاد اوری جزییات هر اطلاعات، نیاز به ذهن انسان است. از اینر وی خلاصه سازی اسناد متنی نقش مهمی در جمع اوری اطلاعات ایفا می کند. در این مطالعه، ما از الکوریتم یادگیری عمیق برای کار های خلاصه سازی استفاده می کنیم. یادگیری عمیق یک زمینه نوهظور یادگیری ماشینی است که برای حل مسئله تعدادی از زمینه های علوم کامپیوتری استفاده می شود نظیر پردازش تصویر، ربات، حرکت،

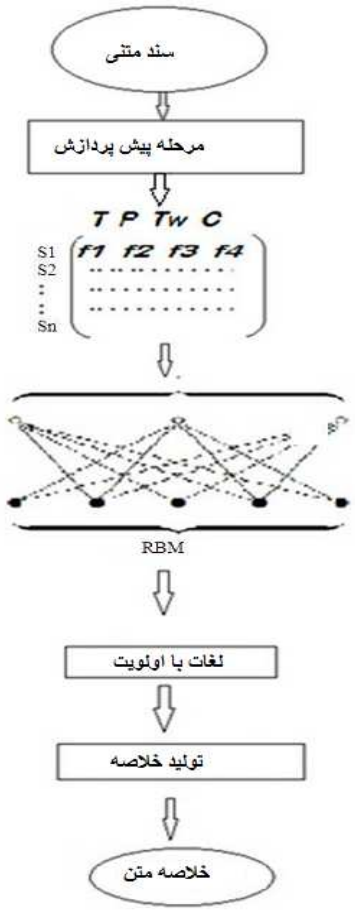
اخیرا، این در زمینه پردازش زبان طبیعی با نتایج مناسب استفاده شده است. یک الگوریتم در صورتی عمیق است که و.رودی آن از چندین ویژگی غیر خطی قبل از خروجی عبور کند. در این جا ما از ماشین بولتزمن محدود برای استخراج برجسته ترین ویژگی کلمه متن استفاده می کنیم.

2-1 ماشین بولتزمن محدود

ماشین بولتزمن محدود، یک شبکه عصبی تصادفی(که شبکه ای از نورون هایی است که در آن هر نورون دارای رفتار تصادفی است)



شکل 1: ماشین بولتزمن محدود



شکل 2: نمودار بلوک خلاصه سازی متنی

این متشکل از یک لایه از نورون ها یا واحد های مرئی و یک لایه از واحد های مخفی است.واحد ها در هر لایه فاقد ارتباط بین خود هستند و به همه واحد های دیگر در آن لایه متصل هستند. اتصالات بین نورون ها دو سویه و متقارن است. این بدین معنی است که اطلاعات در هر دو جهت در طی اموزش و در طی مصرف شبکه جریان دارند و یا اوزان در هر دو جهت یکسان هستند.

3-1 شبکه RBM به طریق زیر کار می کند

نخستین شبکه با استفاده از برخی مجموعه داده ها آموزش داده شده و نورون ها بر روی لایه مرئی مطابق با نقاط داده در این مجموعه داده هستند.

بعد از آموزش شبکه می توان از آن بر روی داده های جدید برای طبقه بندی داده هااستفاده کرد.

4-1 رویکرد یادگیری عمیق پیشنهادی

روش خلاصه سازی متن به دو ریکرد استخراجی و انتزاعی تقسیم می شود. ولی به دلیل محدودیت روش های تولید زبان طبیعی در تولید خلاصه انتزاعی برای خلاصه سازی استفاده می شود. برای خلاصه سازی متن، نیاز به سازمان دهی متن به یک مدل است که به RBM به صورت ورودی نگاه می کند. اول، خلاصه سازی متن، سند متنی با استفاده از روش های پیش پردازش قبلی پیش پردازش می شود و سپس به ماتریس جمله در واژگان تبدیل می شود.بعد از دست یابی به مجموعه ای از لغات با اولویت بالا از RNM، پرس و جوی ورودی، برای تولید خلاصه ای استخراجی از اسناد متنی مقایسه می شود.

5-1 پیش پردازش

برای این که سند سبک تر شود، پیش پردازش اسناد متنی برای ساختار بندی با استفاده از روش های مختلف توسعه یافته توسط یک زبان شناس انجام می شود. هزاران روش برای کاهش تراکم اسناد متنی وجود دارد. در این مطالعه ما از روش های زیر استفاده می کنیم:

6-1 خشی از برچسب گذاری اجزای واژگانی کلام

بخشی از برچسب گذاری اجزای واژگانی کلام، فرایند علامت گذاری یا دسته بندی کلمات متن بر اساس بخش مقوله گفتار مثل اسم، فعل، قید و صفت می باشد. انواع الگوریتم ها برای انجام برچسب زنی POS نظیر مدل های مارکوف مخفی، استفاده از برنامه نویسی پویا وجود دارد.

1-7 فیلترینگ کلمه بازدارندگی

این کلمات کلماتی هستند که قبل یا بعد از پیش پردازش فیلتر می شود که در آن یک قاعده خاص در کلمه خاص برای کلمه بازندارندگی وجود دارد و از این روی بر اساس وضعیت کاملا ذهنی است. در شرایط ما، کلماتی نظیر یک استفاده می شود . این کلمه را از سند اصلی فیلتر می کند. فیلترینگ کلمه باز دارندگی یک فیلتریتک استاندارد می باشد.

1-8 ریشه یابی

یک روش مهم دیگر مورد استفاده ریشه یابی است. ریشه یابی فرایند یافتن ریشه از کلمات مفرد به جای استفاده از اسامی جمع است و از این روی ing را از فعل جدا می کند. تعدادی از الگوریتم های موسوم به ریشه یاب وجود دارد و از این روی برای ریشه یابی استفاده می شوند.

1.9 استخراج ویژگی های بردار

بعد از کاهش تراکم اسناد، سند به صورت ماتریس سازمان دهی می شود. یک ماتریس جمله S با مرتبه $n*v$ حاوی ویژگی هایی برای هر جمله ماتریس است. برای هر خلاصه سازی، ما در حال استخراج چهار ویژگی جمله اسناد متنی ، موقعیت نسبی جمله، کلمات تشکیل جملات، مفهوم استخراج حکم می باشد. بردار ردیفی ماتریس جمله نشان دهنده جمله ای است که ایجاد سند کرده و بردار ستونی حاوی ورودی برای این ویژگی های استخراج شده است.

1-10 محاسبه ویژگی

1-10-1 تشابه عنوان

جمله در صورتی به صورت مهم در نظر گرفته می شود که مشابه با عنوان سند متنی است. از این روی تشابه بر اساس وقوع کلمات رایج در عنوان و جمله در نظر گرفته می شود. جمله در صورتی دارای امتیاز ویژگی خوب است که دارای حداکثر تعداد لغات مشترک برای عنوان باشد. نسبت تعداد کلمات در جمله ک در عنوان تعداد کل کلمات به محاسبه امتیاز جمله این ویژگی کمک می کند و به صورت زیر محاسبه می شود:

$$f1 = \frac{s \cap t}{t}$$

که:

S: مجموعه کلمات جمله

T: مجموعه کلمات یک عنوان

$s \cap t$: لغات رایج در جمله و عنوان اسناد

1-11 ویژگی های مکانی

مقدار موضعی یک جمله استخراج می شود. یک جمله مناسب است و توسط موقعیت خود در متن قضاوت نمی شود. برای محاسبه امتیاز مکانی جمله، شرایط زیر در نظر گرفته می شود

$f2 = 1$: در صورتی که جمله یک جمله اغازین باشد

$f2 = 0$: اگر جمله در بند میانی متن باشد

$f2 = 1$: در صورتی که جمله در پایان متن باشد

1-12: وزن

این دیگر ویژگی مهم برای خلاصه سازی متن است. از این روی منظور از وزن، فراوانی و اهمیت آن است. این مهم ترین ویژگی مورد نظر در پردازش زبان طبیعی است. فراوانی در این جا نشان دهنده اهمیت کلمه در یک سند است و نشان دهنده دفعات ظهور در متن است. این فراوانی کلمه با $tf(f,d)$ نشان داده می شود که در آن F فراوانی کلمه و D متن سند است. عبارت وزن با محاسبه $tf(f,d)$ محاسبه شده و idf نشان دهنده یک سند است. در این جا، idf اشاره به فراوانی سند معکوس است که اطلاعاتی را در مورد اسناد می دهد. این با تقسیم تعداد کل اسناد بر تعداد اسناد حاوی اطلاعات و لگاریتم گیری بدست می اید.idf به صورت زیر بیان می شود

$$idf(t,D) = \log \frac{D}{d \in D : t \in d}$$

که D تعداد کل اسناد $= D: t \in d,$ می باشد که تعداد اسنادی است که در آن عبارت ها ظاهر می شود. وزن کل $tf*idf$ بدست می اید کهبه صورت زیر محاسبه می شود

$$tf * idf(t,d,D) = tf(t,d) * Idf(t,D$$
$$f3 = tf * idf.$$

1-13 مفهوم ویژگی

این مفهوم ویژگی اسناد متنی با استفاده از اطلاعات متقابل و فرایند پنجره بندی استخراج می شود. در فرایند پنجره بندی، یگ پنجره مجازی با اندازه K از سند به سمت راست وارد می شود. هدف ما یافتن وقوع هم زمان کلمات در یگ پنجره می باشد که با فرمول زیر محاسبه می شود

$$MI(w_i, w_j) = \log 2 \frac{P(w_i, w_j)}{P(w_i) * P(w_j)}$$

که $P(w_i, w_j)$ احتمال مشترک این است که هر دو لغات کلیدی با هم در یک پنجره متنی ظاهر شود.

احتمال $P(w_i)$ این که کلمه کلیدی Wi در پنجره متن ظاهر شود با معادله زیر محاسبه می شود

$$P(w_i) = \frac{|sw_t|}{|sw|}$$

که sw_i تعداد پنجره های حاوی کلمات کلیدی و $|sw|$ تعداد کل پنجره های ساخته شده از سند متن است.

ماتریس جمله با مراحل بالا به صورت زیر است:

$$\begin{array}{c} \\ S1 \\ S2 \\ . \\ . \\ Sn \end{array} \begin{pmatrix} T & P & Tw & C \\ f1 & f2 & f3 & f4 \\ .. & ... & .. & .. \\ .. & ... & .. & .. \\ ... & .. & .. & .. \end{pmatrix}$$

1-14 ماتریس جمله

در این جا، ماتریس جمله $S = (s_1, s_2, \ldots\ldots s_n)$ می باشد که در آن $s_i = (f_1, f_2, \ldots\ldots f_4), i <= n$ یک بردار ویژگی است.

1-15 الکوریتم یادگیری عمیق

ماتریس جمله $S = (s_1, s_2,.....s_n)$ که بردار ویژگی است، دارای همه چهار ویژگی استخراج شده برای جمله si می باشد. در این جا این مجموعه از بردار های ویژگی s به صورت ورودی ساختار عمیق RBM تعیین می شود. مقادیر تصادفی به صورت H_i بایایس تعیین می شود که در آن i=1.2 می باشد زیرا RBM دارای خداقل دو لایه مخفی است. کل فرایند را می توان با معادله زیر بدست اورد

$$S=(s_1,s_2.....s_n)$$

که $s_i = (f_1,f_2,.....f_4),\ i<= n$ می باشد و در آن n تعداد جملات در اسناد است. ماشین بولتزمن محدود دارای دو لایه مخفی می باشد و برای آن ها دو مجموعه از مقادیر اریبی انتخاب می شود یعنی H_0H_1 :

$$H_0=\{h_0,h_1,h_2.....h_n\}$$
$$H_1=\{h_0,h_1,h_2.....h_n\}$$

این مجموعه مقادیر اریب، مقادیری هستند که به طور تصادفی انتخاب می شوند. کل عملیات ماتریس جمله با این دو مجموعه از مقادیر تصادفی انتخاب شده انجام می شوند. عملیات کل با RBM با ارایه ماتریس جمله به صورت ورودی شروع می شود. در این جا، $s_1,s_2,.....s_n$ به صورت ورودی RBMداست. به طور کلی RBM دارای دو لایه مخفی است.

دو لایه برای مسئله ما کافی هستند. برای دست یابی به مجموعه ای از ویژگی های جمله، RBM به دو طریق کار می کند. ورودی اولین مرحله، مجمو عه ماتریس جمله $S = (s_1,s_2,.....s_n),$ است که دارای جهار ویژگی به عنوان عنصر مجموعه چمله است. در طی اولین سیکل RBM، یک ماتریس جمله اصلاح شده تعریف می شود.

$$s'=(s'_1,s'_2,.....s'_n)$$

معادله فوق را می توان به صورت زیر انجام داد

$$\sum_{1}^{n} s_i + h_i$$

در طی مرحله 2، همین روش بری دست یابی به مجموعه ماتریس جمله با H1 اعمال می شود:

$$s''=(s''_1,s''_2,.....s''_n)$$

بعد از دست یابی به ماتریس جمله اصلاح شده از RBM، بر روی یک مقدار آستانه تصادفا تولید شده برای هر ویژگی تست می شود و از این روی ما ان را محاسبه کردیم. برای مثال، ما استانه thr_c را به صورت مقدار استانه برای ویژگی استفاده می کنیم. اگر برای هر جمله $f_4<thr$ باشد، آنگاه، فیلتر شده و تبدیل به عضوی از مجموعه جدید بردار ویژکی می شود

مرحله 1: $s_1, s_2 \cdots, s_n$

$$[f1,f2,f3,f4]\ [f1,f2,f3,f4]\ [f1,f2,f3,f4]$$
$$\searrow \quad \downarrow \quad \swarrow$$
$$\sum_1^n s_i + h_i(H_0)$$
$$\downarrow$$
$$s' = (s_1', s_2' \cdots, s_n')$$

مرحله 2: $s_1', s_2' \cdots, s_n'$

$$[f1,f2,f3,f4]\ [f1,f2,f3,f4]\ [f1,f2,f3,f4]$$
$$\searrow \quad \downarrow \quad \swarrow$$
$$\sum_1^n s_i + h_i(H_1)$$
$$\downarrow$$
$$s'' = (s_1'', s_2'' \cdots, s_n'')$$

1.16 تولید مجموعه بردارهای بهینه ویژگی

در بخش اول، ما به یک مجموعه خوبی از بردار های ویژگی با الگوریتم یادگیری عمیق یافتیم. در این مرحله، ابتدا بردار ویژگی را با تعدیل وزن واحد RBM تعدیل می کنیم. برای میزان سازیدقیق بردار ویژگی،باید از الگوریتم انتشار پسین استفاده کرد. الگوریتم انتشار پسین روشی برای میزان سازی ساختار و معماری عمیق با بردار ویژگی بهینه برای خلاصه زمینه ای دقیق است. خطای انتروپی برای تعدیل برای هر ویژگی جمله محاسبه می شودبرای مثال، عبارت ویژگی وزنی جمله با استفاده از فرمول زیر باز سازی می شود

$$[-\sum_v f_v \log \hat{f_v} \wedge -\sum_v (1-f_v)\log(1-\hat{f_v})]$$

که:

f_v: مقدار t_f از Vامین کلمه است

$\hat{f_v}$: مقدار tv باز سازی است.

1-17 تولید خلاصه

در مرحله خلاصه سازی، بردار و.یژگی بهینه برای تولید خلاصه استخراجی اسناد استفاده می شود.برای تولید خلاصه، باید امتیاز جمله برای هر موضوع سند بدست بیاید. امتیاز نمره با یافتن اثر متقابل کاربر با پرس و چو و جمله بدست می اید.بعد از این مرحله رتبه بندی جمله انجام شده و مجموعه جملات نهایی برای تولید خلاصه بدست می اید

1-18 امتیاز جمله

نسبت امتیاز جمله در پرس و چوی کاربر و جمله با تعداد کلکات کل در سند متنی یافت شده است. و با معادله زیر بدست می اید

$$S_c = \frac{s \cap Q}{wc}$$

که

Sc: امتیاز جمله

S: جمله

Q: پرس و جوی کاربر

Wc: تعداد کل کلمات متن

1-19 رتبه بندی جمله

این گام نهایی برای دست یابی به خلاصه متن است. در این جا رتبه بندی جمله بر اساس امتیاز جمله بدست امده از مرحله قبلی انجام می شود. جمله به ترتیبنزولی بر اساس امتیاز جمله بدست امده ارایش می یابد. برای یافتن تعداد جملات برتر برای اتخاب از ماتریس، فرمول زیر مطرح می .شود

$$N = \frac{C \times N_s}{100}$$

که Ns:تعداد جملات در سند

C: نرخ فشرده سازی است

1-120 نتایج و تجزیه تحلیل

رویکرد پیشنهادی به بررسی خلاصه سازی متن بر اساس روش یادگیری عمیق می پردازد این روش از الکوریتمRBM برای دست یابی به کارایی بهتر می پردازد. عملکرد رویکرد پیشنهادی در بخش 12-1 ارزیابی می شود.

1-21 توصیف مجموعه داده

ارزیابی آزمایشی الکوریتم خلاصه متن پیشنهادی بر روی اسناد مختلفی اجرا می شود. اسناد از زمینه های خاص جمع اوری می شوندنظیر داده کاوی و مهندسی نرمافزار.اسناد مختلف ازهریک از حوزه هایوختلف جمع وری شده و پردازش می شود.کلمه داده کاوی در کوگلسرچ شدهوبعد از ان بردار ویگی مطر- می شود

1-22 شاخص های ارزیابی

ارزیابی روش خلاصه سازی متن پیشنهادی بر اساس سه معیار است. معیار های مختلف در زیر نشان داده شده اند

1-23 فراخوانی

فرا خوانی نسبت تعداد جملات بازیابی شده به تعداد جملات است.فراخوانی برای اندازه گیری پایایی روش خلاصه

$$\text{Recall} = \frac{S_{Ret} - S_{Rel}}{S_{Ret}}$$

سازی متنی اسفاده می شود/

که S_{ret} و S_{rel} اعداد بازیابی شده و جملات مربوطه است.

1-24 دقت

نسبت جملات بازیای شده به جملاتمربوطه به صئورت شاخص دقیق است

دقت: $\frac{S_{Ret} - S_{Rel}}{S_{Rel}}$

شاخص 1-25: شاخص F

مقادیر دقت و مقادیر فراواخی برای یافتن مقدار شاخص F برای کل مجموعه داده در نظر گرفته می شود

از این روی شاخص F به صورت زیر بیان می شود

$$F-\text{measure}=\frac{2\times \text{Recall}\times \text{Precision}}{\text{Recall}+\text{Precision}}$$

1-26 استخراج بردار ویژگی

نتایج استخراج ویژگی خلاصه سازی اسناد، در بخش 1-26 مطرح شده است. از این روی ده سند با موضوع مشابه به عنوان ورودی وجود دارد. خلاصه تولید شده با استفاده از خلاصه موجود با پردازش شاخص فرا خوانی و F تعیین می شود.جدول 1 بردار های ویژگی استخراج شده از مجموعه اسناد را نشان می دهد. مقادیر نشان دده شده بر اساس بالاترین مقداردی رارزیابی شده می باشد که در جدول فوق نشان داده شده است.

1-27 ارزیابی عملکرد

ارزیابی عملکرد رویکرد پیشنهادی در بخش 1-27 نشان داده شده است. فرایند ارزیابی در سه مجموعه اسناد متفاوت انچام می شود. رویکرد پیشنهادی در بخش 1-28 نرسیم شده است. فراخوانی، دقت و شاخص f برای همه سه مجموعه محاسبه می شود. مقادیر مختلف استانه برای تایید پاسخ الگوریتک خلاصه سازی تحت شرابط مختلف استفاده می شود. این استانه از RBM انتخاب می شود. سه استانه فیلترینگ برای هر سند استفاده می شود.

در شکل 3، پاسخ مجموعه سند نشان داده شده است. مجموعه اسناد متشکل از اسناد مربوط به حوزه شبگکه است. تعداد اسناد در مجموعه سند، ده سند است.خلاصه با کمک الگوریتم خلاصه سازی متنی تولید می شود. ماکزیمم مقدار فرا خوانی برای حوزه شبکه بندی برابر با 0.429 است. به طور مشابه، ارزش دقت ماکزیمم به صورت 0.6 است/ مقدار شاخص F، براساس مقدار دقت و فراخوانی محاسبه می شود.

شکل 4، پاسخ مهندسی نرم افزار مربوط به داده های اسناد را نشان می دهد. پاسخها در مقایسه با اولین مجموعه اسناد متفاوت هستند. مقدار شاخص بهصورت0.469 است.

پاسخ مجموعه سمد درشکل 5نشان داده شده است. پاسخ حوزه شبگه کاملا متفاوت از حوزههای دیگر است.بر اساس این تحلیل،بدیهی است که الگوریتم خلاصه سازی متن به داده حساس است.

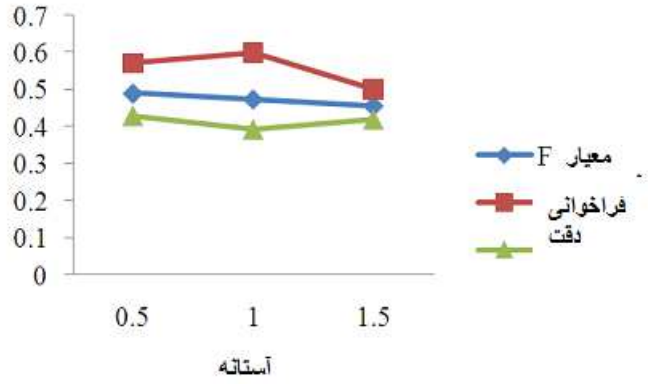
28-1 تحلیل قیاسی

نمودار تحلیل قیاسی عملکرد رویکرد پیشنهادی وروش موجود ترسیم شد.هر دوروش بر اساس الگوریتم یادگیری هستند. این الگوریتم بر مقادیر فرا.خوانی روش پینشهادی متمرکز می باشد.مقادیر فراوخوانی هر دوالگوریتم بر اساس اتخاذ شده است

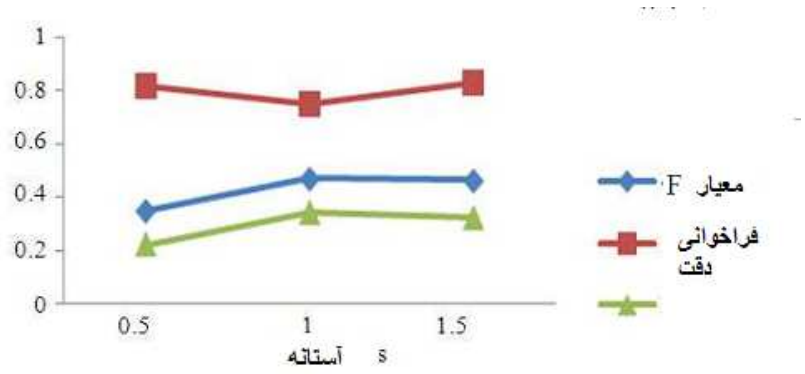
شکل 6 تحلیل تفضیلی رویکرد پیشنهادی و رویکرد موجود را نشان می دهد. مقادیر فراخوانی از مقادیر آستانه از 0.5 تا2 متغیر است. تحلیل گراف نشان می دهد که رویکرد پیشنهادی به رویکرد موجود می چربد.

شکل 1: استخراج بردار ویژگی

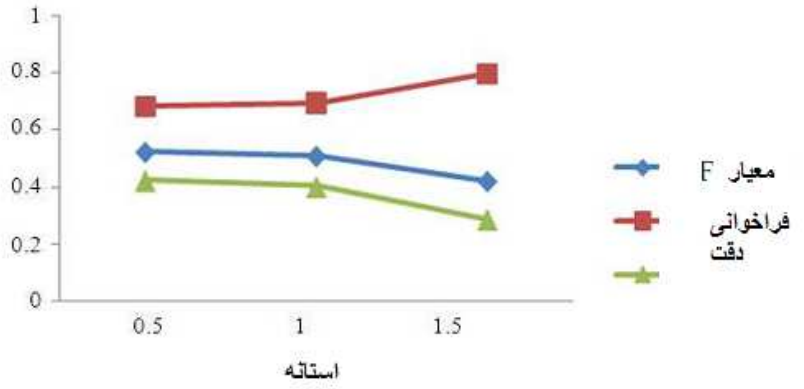
Document no:	Paragraph no:	Line no:	Title value:	Position value:	tf_idf:	Concept:
2	0	1	0.888889	3.0	0.736645722	0.290139693
2	2	1	0.777779	2.8	0.730378687	0.655319108
2	2	2	0.666667	2.6	0.731382288	0.674829335
3	0	1	0.700000	3.0	0.694924858	0.213265682
3	2	3	0.800000	2.4	0.952361002	0.471023052
8	4	3	0.555556	2.4	0.489351427	0.182562528
8	5	1	0.444444	1.0	0.462419924	0.148672137
9	0	1	0.727273	3.0	0.724465870	0.219798458
9	5	2	0.545455	2.6	0.671540289	0.405503813



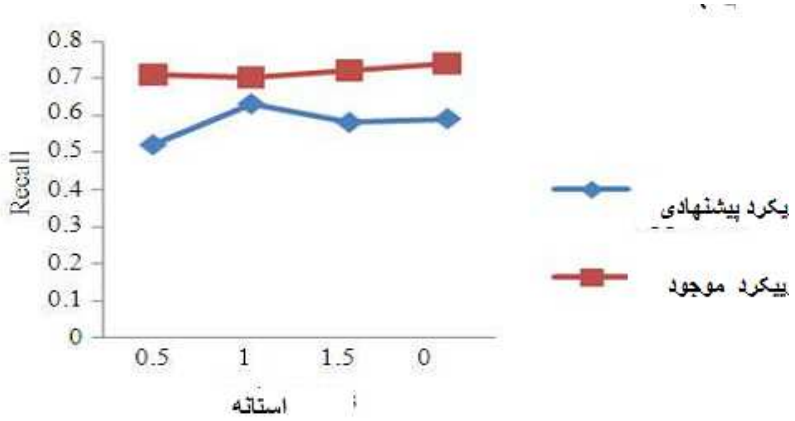
شکل 3: عملکرد دامنه شبکه بندی



شکل 4: عملکرد دامنه مهندسی نرم افزار



شکل 5: عملکرد دامنه شبکه بندی



شکل 6: تحلیل قیاسی

ماکزیمم مقادیر فراخوانی برای روش موجود و پیشنهادی به ترتیب 0.72 و 0.62 است.

2-نتیجه گیری

تحقیقات مختلف برای تولید خلاصه از چند سند در روز های اخیر انجام شده است. ما یک سیستم خلاصه سازی چند اسنادی خودگاررا توسعه دادیم که شامل RBM است. ما از چهار ویژگی های مختلف برای مرحله استخراج ویژگی استفاده کردیم. امتیاز ویژگی جمله به RBM اعمال شده و قواهد RBM با کمک الگوریتم یادگیری عمیق بهینه سازی می شود. ویژکگی ها از طریق سطوح مختلف الکوریتم RBM پردازش شده و خلاصه متن بر این اساس تولید می شود. نتیجه بر اساس ماتریس ارزیابی تست می شود. ماتریس تکامل در متن بهصورت فراخوانی، دقت و معیار F مطرح شده است. ازمایش الگوریتم با در نظر گرفتن سه مجموعه اسناد انجام شد. پارامتر قضاوت عملکرد معیار F دارای مقادیر 0.49، 0.469 و 0520 است. بهبود آینده نگرانه رویکرد پیشنهادی با در نظر گرفتن ویژگی های مختلفو افزودن چند لایه مخفی به الگوریتم RBMحاصل می شود.