علوم رایانش و فناوری اطلاعات

نشریه علمی انجمن کامپیوتر ایران

صاحب امتیاز: انجمن کامپیوتر ایران

مدیر مسئول: دکتر جعفر حبیبی

سردبیر: دکتر احمد خونساری

نشانی

تهران، خیابان آزادی، ضلع غربی دانشگاه صنعتی شریف، کوچه شهید ولی‌ا... صادقی، پلاک ۲۶، طبقه ۴، واحد ۱۶، دفتر انجمن کامپیوتر ایران، نشـریه علوم رایانش و فناوری اطلاعات

تلفن: ۶۶۰۳۲۰۰۰-۶۶۰۸۷۲۲۴

دورنگار: ۶۶۰۲۱۱۴۹

پست‌الکترونیکی: csitjour@gmail.com

نشانی سایت: http://csi.org.ir/fa/publication/archive/name/csit

لیتوگرافی، چاپ و صحافی:

علوم رایانش و فناوری اطلاعات نشریه علمی انجمن کامپیوتر ایران

مجلد ۱۴	شماره ۲	۱۳۹۵

فهرست مقالات

علوم رایانش و فناوری اطلاعات
نشریه علمی انجمن کامپیوتر ایران
مجلد ۱۴، شماره ۲، ۱۳۹۵
صفحات ۱- ۱۱
مقاله عادی

دسته‌بندی و حاشیه‌نویسی همزمان تصویر با استفاده از مدل‌های احتمالاتی موضوع و کدگذاری LLC کلمات بصری

سید نوید محمدی فومنی احمد نیک‌آبادی

دانشکده مهندسی کامپیوتر و فناوری اطلاعات، دانشگاه صنعتی امیرکبیر، تهران، ایران

چکیده

تاکنون تلاش‌های زیادی به منظور استفاده از مدل‌های موضوعی نظیر مدل احتمالاتی LDA جهت دسته‌بندی و حاشیه‌نویسی همزمان تصاویر صورت گرفته است. اخیراً مدل‌های موضوع دیگری بر مبنای شبکه‌های عصبی احتمالاتی نظیر SupDocNADE معرفی شده‌اند که نتایج خوبی در مدل‌کردن داده‌های چندمقداری ارائه داده‌اند. در این مدل‌ها کلمات حاشیه‌نویسی نیز در کنار کلمات بصری تعبیه شده و به عنوان بردار ویژگی برای شبکه در نظر گرفته می‌شود. عدم‌تعادل در تعداد کلمات بصری و حاشیه‌نویسی سبب می‌شود تا سهم کلمات حاشیه‌نویسی برای بازنمایی در لایه پنهان شبکه‌عصبی مورد استفاده در این مدل، بسیار کمتر از کلمات بصری باشد. برای حل این مشکل در این مقاله، کلمات حاشیه‌نویسی در هیستوگرام بردار ویژگی وزن‌دهی می‌شوند. با افزودن قابلیت وزن‌دهی ورودی‌ها می‌توان از کدگذار LLC که چندین کلمه مشابه در فرهنگ لغت را بصورت وزن‌دار در ساخت بردار ویژگی دخیل می‌کند، برای تولید کلمات بصری استفاده نمود. با آزمایش مدل پیشنهادی بر روی پایگاه داده‌های UIUC_Sports و LabelMe، بهبود ۵ درصدی در معیار F در کلمات حاشیه‌نویسی و بهبود ۱ درصدی در دقت دسته‌بندی نسبت به مدل‌های موجود مشاهده می‌شود.

کلمات کلیدی: دسته‌بندی و حاشیه‌نویسی تصویر، مدل‌های موضوع، مدل احتمالاتی، شبکه عصبی، کدگذار LLC.

۱- مقدمه

گسترش روزافزون داده‌ها و ذخیره‌ی آنها بصورت دیجیتال مدیریت و سازماندهی آنها را به یک امر مهم تبدیل کرده است. در این میان، طبقه‌بندی این داده‌ها و اطلاعات به طور خودکار به یک چالش مهم تبدیل شده است. تصاویر و متون حجم عظیمی از این داده‌ها را به خود اختصاص داده‌اند. دسته‌بندی و حاشیه‌نویسی[۱] تصاویر از جمله مسائل پرکاربرد در حوزه پردازش تصاویر هستند. در مسأله دسته‌بندی تصاویر، هدف تعیین دسته یک تصویر ورودی از میان مجموعه‌ای از دسته‌های از قبل تعریف شده است. در مسأله حاشیه‌نویسی سعی می‌شود که مجموعه‌ای از کلمات کلیدی مرتبط با یک تصویر استخراج شود.

مدل‌های احتمالاتی موضوع[۲] یا تخصیص پنهان دیریکله[۳] را می‌توان به عنوان یکی از معروف‌ترین رویکردها در مدل کردن داده‌های چندمقداری[۴] دانست [۱، ۲، ۳] تخصیص پنهان دیریکله که به اختصار LDA نامیده می‌شود، مدل مولدی است که ابتدا در زمینه پردازش زبان‌های طبیعی معرفی شد. در این مدل هر سند بصورت توزیع چندجمله‌ای روی موضوعات تعریف شده است. از طرفی هر موضوع یک توزیع چندجمله‌ای روی کلمات می‌باشد. لازم به ذکر است که موضوعات بین همه سندها مشترک بوده ولی توزیع موضوعات برای هر سند، خاص آن سند می‌باشد [۴].

استفاده از مدل‌های احتمالاتی موضوع در بینایی ماشین با تعریف "کلمات بصری[۵]" آغاز شد [۵]. با استخراج کلمات بصری، هر تصویر تبدیل به یک سند می‌شود و حال می‌توان با آموزش LDA روی کیسه‌ای از کلمات بصری، تصاویر را مدل کرد. LDA یک مدل بدون ناظر بوده و برای کاربردهایی مانند دسته‌بندی مناسب نمی‌باشد. از این‌رو مدل تخصیص پنهان دیریکله با ناظر sLDA معرفی شد که با مدل کردن برچسب هر کلاس به همراه کلمات بصری، سبب استخراج ویژگی تمایزی می‌شود [۶]. اولین مدل جدی که دسته‌بندی و حاشیه‌نویسی را بطور

همزمان انجام داد را می‌توان تخصیص پنهان دیریکله با ناظر چند کلاسه[۶] (MC_sLDA) دانست [۷]. در این مدل و مدل‌های مشابه تغییرات جدی در مدل پایه‌ای LDA انجام شده تا بتواند یک فضای مشترک بین برچسب تصاویر و متن حاشیه برقرار کند [۸، ۹، ۱۰].

یکی از معایب این مدل‌ها عدم‌وجود استنتاج دقیق[۷] و قابل محاسبه[۸] می‌باشد. در این مدل‌ها برای بدست آوردن توزیع پسین[۹] مجبور به استفاده از روش‌های تقریبی هستیم که به مراتب کندتر بوده و بار محاسباتی بالایی دارند. برای حل مشکلات فوق، مدلی بر پایه ماشین بولتزمن محدود با نام سافت‌مکس تکراری[۱۰] معرفی شد [۱۱]. استنتاج برروی اسناد بازنمایی شده توسط این مدل، بسیار کاراتر از مدل‌های موضوع پیشین است. مهمترین مشکل مدل‌هایی که بر پایه ماشین بولتزمن محدود معرفی می‌شوند قابل محاسبه نبودن تابع نرمالیزه‌کننده یا همان تابع تقسیم[۱۱] است [۱۲]. عیب دیگر این روش‌ها ناتوانی در مدل کردن حالتی است که احتمال وقوع یک کلمه در هر موضوع خاص کم اما در ترکیبی از موضوعات زیاد است. به عبارت دیگر این کلمات توسط موضوعات استخراج شده، قابل تخمین نمی‌باشد. در مدل سافت‌مکس تکراری روش نمونه‌برداری مبتنی بر تابکاری (AIS[۱۲]) مورد استفاده قرار گرفته شده است که روشی مناسب جهت تخمین نسبت تابع تقسیم است [۱۳]. مهمترین عیب مدل سافت‌مکس تکراری وجود پیچیدگی زمانی خطی برابر با اندازه فرهنگ لغت، در زمان بروزرسانی (یادگیری) پارامترها می‌باشد.

شبکه عصبی تخمین‌گر توزیع اتورگرسیو[۱۳] (NADE) جایگزین مناسبی برای RBM[۱۴]ها است [۱۴]. این مدل بیشتر شبیه اتوانکدرها[۱۵] است که اندازه واحدهای ورودی و خروجی یکسان می‌باشد. در این مدل‌ها بدون نیاز به تخمین و با استفاده از روش‌های گرادیان، پارامترهای مدل آموزش داده می‌شود که این مزیت اصلی آنها به شمار می‌رود.

در [۱۵] مدل مولدی برای اسناد ارائه شده که از ترکیب NADE و سافت‌مکس تکراری الهام گرفته شده است. این مدل که با نام DocNADE شناخته می‌شود، توزیع توام را با استفاده از قاعده زنجیره‌ای تجزیه کرده و هر احتمال شرطی را مانند NADE توسط یک گره از شبکه مدل می‌کند. از این مدل علاوه بر استفاده در داده‌های متنی می‌توان در داده‌های چند مقداری نیز استفاده نمود. به دلیل بدون ناظر بودن این مدل و نیاز به استخراج ویژگی‌های تمایزی بیشتر، مدل باناظر آن با عنوان SupDocNADE معرفی شد [۱۶]. مدل SupDocNADE به‌دلیل با ناظر بودن یادگیری آن، توانایی استخراج ویژگی‌های تمایزی بیشتری دارد از این‌رو جهت مدل کردن توام برچسب کلاس، کلمات حاشیه‌نویسی و کلمات بصری استفاده شده است.

در تمامی مدل‌های ذکر شده جهت دسته‌بندی و حاشیه‌نویسی تصاویر از روش کیسه ویژگی‌ها [۱۶] استفاده شده است که به اختصار BoF نیز گفته می‌شود [۱۶،۹،۷]. در این روش هر تصویر توسط هیستوگرام تعداد رخداد الگوها بازنمایی می‌شود و مهمترین مزیت این روش‌ها کم بودن بارمحاسباتی آن است [۱۷، ۱۸].

در روش سبد ویژگی‌ها، اطلاعات مکانی مورد توجه قرار نمی‌گیرد و برای این منظور روش تطبیق هرم مکانی[۱۷] مطرح شد که به اختصار SPM گفته می‌شود. در روش SPM چند سطح مختلف از تصویر جهت بازنمایی نهایی استفاده می‌شود [۱۹، ۲۰]. این روش نیازمند یک دسته‌بند غیرخطی می‌باشد که محاسبات لازم جهت دسته‌بندی را افزایش می‌دهد. از طرفی محاسبه توصیف‌گر در سطوح مختلف دارای پیچیدگی زمانی بسیاری است. برای حل مشکل سرعت محاسبات در روش SPM و دقت پایین روش BoF روش کدگذاری خطی با محدودیت محلی بیان شد که به اختصار LLC گفته می‌شود. این روش کدگذاری جای چندی‌سازی[۱۸] برداری را در روش‌های گفته شده می‌گیرد [۲۱]. روش چندی‌سازی برداری در روش‌های ساخت کلمات بصری BoF و SPM استفاده شده است. چندی‌سازی در این روش‌ها یعنی تخصیص تنها نزدیک‌ترین کلمه بصری در فرهنگ لغات جهت بازنمایی هر تکه از تصویر. حال با استفاده از روش LLC چندین کلمه بصری بصورت وزن‌دار در بازنمایی تکه‌های استخراج شده از تصویر شرکت می‌کنند.

در مدل SupDocNADE کلمات حاشیه‌نویسی و کلمات بصری کنار هم تعبیه شده و بردار ورودی را تشکیل می‌دهند. از طرفی کلمات بصری بسیار بیشتر از کلمات حاشیه‌نویسی می‌باشند و این عدم‌تعادل سبب می‌شود تا مدل کارایی مناسبی در حاشیه‌نویسی تصاویر نداشته باشد. برای حل این مشکل، در این مقاله پیشنهاد می‌شود که کلمات حاشیه‌نویسی در هیستوگرام بردار ویژگی وزن‌دهی شود. این وزن‌دهی سبب می‌شود تا بردار کلمات حاشیه‌نویسی تاثیر مناسبی در بازنمایی پنهان مدل داشته باشند. از طرفی با افزودن قابلیت وزن‌دار شدن ورودی‌ها می‌توان از کدگذارهایی مانند LLC به جای روش‌های سنتی کوانتیزاسیون برداری استفاده نمود.

در بخش ۲ ابتدا مدل پایه‌ای DocNADE و نحوه‌ی مدل کردن اسناد متنی شرح داده شده و سپس نحوه‌ی گسترش این مدل جهت کار با داده‌های برچسب‌دار معرفی می‌شود. در ادامه این بخش روش‌های مختلف کدکردن کلمات بصری شرح داده شده است. همانطور که پیش‌تر بیان شد عدم‌تعادل بین تعداد کلمات بصری و کلمات حاشیه‌نویسی سبب مشکلاتی در یادگیری کلمات حاشیه می‌شود که در بخش ۳ روش پیشنهادی جهت حل این مشکلات مطرح شده است. در بخش ۴ آزمایشات مختلفی جهت بررسی کارایی مدل انجام شده و نتایج آن مقایسه شده است. در بخش پایانی جمع‌بندی روی مطالب گفته شده و روش پیشنهادی انجام شده است.

۲- مبانی نظری

در این بخش ابتدا مدل‌های موضوع بر پایه شبکه‌های عصبی شرح داده شده است و سپس دو روش مورد استفاده در ساخت کلمات بصری شرح داده می‌شود و در انتها نحوه‌ی کد کردن با استفاده از روش LLC بیان می‌شود.

۲-۱- مدل‌های DocNADE و SupDocNADE

در این بخش ابتدا به شرح شبکه عصبی تخمین‌گر توزیع اتورگرسیو برای اسناد (DocNADE) پرداخته شده است. سپس مدل باناظر DocNADE که برچسب کلاس را نیز مدل می‌کند شرح داده می‌شود. این مدل با دخیل کردن برچسب داده‌ها در روند آموزش ویژگی‌ها، تمایز بیشتری توسط ویژگی‌های پنهان مدل کسب می‌کند. در ادامه نحوه‌ی بهره‌برداری از اطلاعات مکانی توسط این مدل شرح داده شده و در انتها نحوه‌ی مدل کردن کلمات حاشیه به همراه سایر ویژگی‌ها در مدل SupDocNADE شرح داده شده است.

۲-۱-۱- DocNADE

DocNADE به عنوان رویکردی برای مدل کردن اسناد معرفی شده است. برای استفاده از این مدل زمانی که نوع ورودی تصویر باشد لازم است که هر تصویر به صورت کیسه کلمات بصری بازنمایی شود. ابتدا تصویر به تکه‌های مشخص تقسیم می‌شود سپس برروی هر تکه توصیف‌گر SIFT اعمال می‌شود. حال با اعمال الگوریتم خوشه‌بندی نماینده‌هایی از این توصیف‌گرها به عنوان کلمات بصری در نظر گرفته می‌شوند (برای اطلاعات بیشتر به بخش ۴-۲ مراجعه شود). حال هر تصویر می‌تواند بصورت کیسه‌ای از کلمات بصری $v = [v_1, v_2, \dots, v_D]$ بازنمایی

شود که v_i اندیس نزدیکترین خوشه به iامین توصیف‌گر SIFT می‌باشدو همچنین D تعداد توصیف‌گرهای استخراج شده از تصویر را مشخص می‌کند [۱۵].

توزیع توام کلمات بصری $p(v)$ در DocNADE بصورت احتمالات شرطی $p(v_i|v_{<i})$ بازنویسی می‌شود:

$$p(v) = \prod_{i=1}^{D} p(v_i|v_{<i}) \quad (۱)$$

که در آن بردار $v_{<i}$ شامل همه v_jهای می‌شود که $j < i$ است. لازم به ذکر است که رابطه (۱) برای تمامی توزیع‌ها براساس قانون زنجیره‌ای احتمال صادق است. در واقع مهمترین فرض DocNADE این است که احتمالات شرطی می‌توانند توسط یک شبکه عصبی پیشخور مدل شوند.

یکی از معماری‌هایی که می‌توان برای مدل کردن $p(v_i|v_{<i})$ در نظر گرفت به صورت زیر است:

$$h_i(v_{<i}) = g\left(c + \sum_{k<i} W_{:,v_k}\right) \quad (۲)$$

$$p(v_i = w|v_{<i}) = \frac{\exp\left(b_w + V_{w,:}h_i(v_{<i})\right)}{\sum_{w'} \exp\left(b_{w'} + V_{w',:}h_i(v_{<i})\right)} \quad (۳)$$

که $g(.)$ تابع فعالیت درایه به درایه غیرخطی است، $W \in \mathbb{R}^{H\times K}$ و $V \in \mathbb{R}^{K\times H}$ ماتریس اتصال، $b \in \mathbb{R}^{K}$ و $c \in \mathbb{R}^{N}$ پارامترهای بایاس و H، K به ترتیب تعداد واحدهای پنهان (موضوعات) و اندازه فرهنگ‌لغت می‌باشد.

محاسبه توزیع $p(v_i = w|v_{<i})$ نیازمند محاسبات خطی از مرتبه K بوده و تکرار این کار به تعداد کلمات بصری D بسیار زمان‌بر است. برای حل این مشکل از درخت دودویی استفاده شده است. استفاده از درخت دودویی محاسبات را از مرتبه خطی K به لگاریتمی تبدیل می‌کند. برای این کار هر برگ درخت به صورت تصادفی به یک اندیس کلمه تخصیص پیدا می‌کند و احتمال آن کلمه برابر با ضرب احتمالات موجود در مسیر ریشه تا آن برگ تعریف می‌شود. احتمال انتقال به چپ/راست در درخت توسط تابع رگرسیون لاجستیک دودویی مدل شده که ورودی آن $h_i(v_{<i})$ می‌باشد.

فرض کنید $l(v_i)$ نشان‌دهنده‌ی دنباله گره‌های درخت از ریشه به برگ v_i و $\pi(v_i)$ نشان‌دهنده دنباله دودویی انتخاب چپ/راست گره در مسیر باشد. به عنوان مثال $l(v_i)_1$ همواره گره‌ی ریشه و $\pi(v_i)_1$ برابر ۱ است اگر v_i در زیردرخت چپ باشد و برابر صفر در حالت عکس آن. حال $V \in \mathbb{R}^{T\times H}$ ماتریسی شامل وزن‌های رگرسیون لاجستیک است و $b \in \mathbb{R}^{T}$ بردار بایاس که T تعداد گره‌های داخلی درخت دودویی و H تعداد واحدهای مخفی می‌باشد. احتمال $p(v_i = w|v_{<i})$ بصورت زیر مدل شده است:

$$p(v_i = w|\mathbf{v}_{<i}) = \prod_{k=1}^{|\pi(v_i)|} p(\pi(v_i)_k|\mathbf{v}_{<i}) \quad (۴)$$

که خروجی رگرسیون لاجستیک گره‌های میانی

$$p(\pi(v_i)_k = 1|v_{<i}) = \mathrm{sigm}\left(b_{l(v_i)_m} + V_{l(v_i)_m,:}h_i(v_{<i})\right) \quad (۵)$$

و $\mathrm{sigm}(x) = 1/(1+\exp(-x))$ تابع سیگموئید و $m \in \{1,2,\dots,|\pi(v_i)|\}$ است. تابع رگرسیون لاجستیک رابطه (۴) از مرتبه $O(\log K)$ است.

بنابر آنچه گفته شد، با ترکیب روابط (۲، ۴ و ۵) می‌توان احتمال $p(v) = \prod_{i=1}^{D} p(v_i|v_{<i})$ را برای هر سندی محاسبه کرد. یادگیری پارامترهای $\{W, V, b, c\}$ مدل توسط بهینه کردن درست‌نمایی داده‌های آموزشی با استفاده از رویکرد نزول در امتداد گرادیان حاصل می‌شود.

پس از آموزش مدل، بازنمایی پنهان سند جدید v^* بصورت زیر استخراج می‌شود:

$$h_y(v^*) = g\left(c + \sum_{i}^{D} W_{:,v_i^*}\right) \quad (۶)$$

این بازنمایی می‌تواند به عنوان ورودی یک دسته‌بند جهت اعمال باناظر بینایی ماشین در نظر گرفته شود. در واقع اندیس y نشان‌دهنده استفاده این بازنمایی در تخمین برچسب کلاس تصویر می باشد [۸].

SupDocNADE -۲-۱-۲

مشاهده شده است که استخراج ویژگی‌های بصری از تصویر، با استفاده از مدل‌های احتمالاتی موضوع مانند LDA نمی‌توانند نتایج مناسبی برای کاربردهایی همچون دسته‌بندی داشته باشند. یکی از دلایل این امر را می‌توان در نوع آموزش ویژگی‌ها دانست. ویژگی‌هایی که به صورت بدون ناظر از تصاویر استخراج می‌شوند برای توصیف ساختار آماری تصویر آموزش داده شده‌اند و این ویژگی‌ها نمی‌توانند ساختاری را استخراج کنند تا تمایز بین کلاسی حداکثر شود. این موضوع سبب ابداع انواع مختلف LDA مانند sLDA شده است [۶، ۷]. DocNADE نیز یک مدل موضوع بدون ناظر می‌باشد از این رو SupDocNADE معرفی شد تا ویژگی‌های مناسبی را جهت کاربرد دسته‌بندی استخراج نماید [۱۶].

الف) آموزش مدل

بطور خاص اگر تصویر $v = \{v_1, v_2, \dots, v_D\}$ و برچسب کلاس $y \in \{1, \dots, C\}$ به عنوان ورودی مدل باشد آنگاه توزیع توام SupDocNADE بصورت زیر تعریف می‌شود:

$$p(\mathbf{v}, y) = p(y|\mathbf{v}) \prod_{i=1}^{D} p(v_i|\mathbf{v}_{<i}) \quad (۷)$$

و مانند DocNADE احتمالات شرطی توسط واحدهای شبکه عصبی مدل می‌شود. در این مدل نیز از معماری مشابه DocNADE برای $p(v_i|v_{<i})$ استفاده شده است و تنها نیازمند تعریف مدلی برای $p(y|\mathbf{v})$ می‌باشد. از آنجایی که $h_y(v)$ بازنمایی تصویر است، می‌توان از آن جهت دسته‌بندی استفاده کرد. از اینرو $p(y|v)$ بصورت یک رگرسیون لجستیک چند کلاسه مدل می‌شود که نحوه‌ی محاسبه آن از روی $h_y(v)$ بصورت زیر است:

$$p(y|v) = \mathrm{softmax}\left(d + Uh_y(v)\right)_y \quad (۸)$$

که تابع $\mathrm{softmax}(a)_i = \exp(a_i) / \sum_{j=1}^{C} \exp(a_j)$ و $d \in \mathbb{R}^{C}$ پارامتر بایاس برای لایه با ناظر و $U \in \mathbb{R}^{C\times H}$ ماتریس اتصال بین لایه پنهان h_y و برچسب کلاس است.

به عبارت دیگر $p(y|v)$ به صورت یک شبکه عصبی چند کلاسه مدل می‌شود که ورودی آن کیسه‌ای از کلمات بصری است. تفاوت اساسی این مدل با شبکه‌های عصبی در این است که برخی پارامترهای آن (پارامترهای پنهان W و c) نیز جهت مدل کردن احتمال شرطی کلمات بصری $p(v_i|v_{<i})$ استفاده می‌شوند. برای بیشینه کردن درست‌نمایی مدل باید تابع

$$-\log p(v, y) = -\log p(y|v) + \sum_{i=1}^{D} -\log p(v_i|v_{<i}) \quad (۹)$$

روی تصاویر آموزشی کمینه شود که این به عنوان یادگیری از نوع مولد شناخته می‌شود [۱۱]. عبارت اول در رابطه فوق بطور کامل تمایزی است، در حالی که عبارت دوم بدون ناظر بوده و می‌تواند به عنوان تنظیم کننده در نظر گرفته شود. این عبارت تنظیم کننده بصورت بدون ناظر ساختار آماری بین کلمات بصری را مدل می‌کند. در عمل این تنظیم‌کننده می‌تواند جواب را به سمتی بایاس کند که خاصیت تمایزی زیادی نداشته باشد و سبب تعمیم‌پذیری بهتر مدل شود. همانند کارهای قبلی می‌توان ترکیبی از یادگیری مولد/تمایزی را در نظر گرفت که این امر با اعمال وزن‌دهی صورت می‌گیرد.

$$-\log p(v,y) = -\log p(y|v) + \lambda \sum_{i=1}^{D} -\log p(v_i|v_{<i}) \qquad (۱۰)$$

که λ به عنوان پارامتر تنظیم کننده رفتار می‌کند. جهت بهینه کردن رابطه (۱۰) بر روی مجموعه آموزشی از روش نزول در امتداد گرادیان تصادفی، و از قانون پس انتشار خطا برای محاسبه مشتق پارامترها استفاده شده است.

ب) موقعیت مکانی ویژگی‌های بصری

اطلاعات مکانی همواره نقش مهمی در درک تصاویر داشته است. به عنوان مثال آسمان همواره در قسمت بالای تصویر دیده می‌شود و یا ماشین همواره در نیمه پایینی تصویر به چشم می‌خورد. بسیاری از کارهای پیشین از این اطلاعات بطور موثر در کار خود استفاده کرده‌اند [۱۶].

فرض کنید تصویر به چند ناحیه مجزا $R = \{R_1, R_2, \dots, R_M\}$ تقسیم شودکه M تعداد نواحی مختلف می‌باشد. حال یک تصویر می‌تواند به صورت زیر بازنمایی شود

$$v^R = [v_1^R, v_2^R, \dots, v_D^R] = [(v_1, r_1), (v_2, r_2), \dots, (v_D, r_D)] \qquad (۱۱)$$

که $r_i \in R$ ناحیه‌ای است که کلمه بصری v_i از آن استخراج شده است. برای مدل کردن احتمال توام کلمات بصری ابتدا توزیع بصورت $\prod_i p((v_i, r_i)|v_{<i}^R)$ تجزیه می‌شود و با هر یک از $K \times M$ جفت کلمه/ناحیه بصورت یک کلمه بصری مجزا رفتار می‌شود. این امر به این معنی است که به ازای هر حالت یک برگ به درخت دودویی افزوده می‌شود و از طرفی چون محاسبات بصورت لگاریتمی با اندازه برگ‌های درخت رشد می‌کند مشکلی با افزایش تعداد مناطق نخواهیم داشت.

ج) افزودن کلمات حاشیه‌نویسی

در این بخش به شرح چگونگی مدل کردن کلمات حاشیه‌نویسی پرداخته می‌شود. فرض کنید A فرهنگ‌لغت از پیش تعریف شده برای کلمات حاشیه‌نویسی باشد. برای یک سند خاص کلمات حاشیه‌نویسی بصورت $a = [a_1, a_2, \dots, a_L]$ نمایش داده شود که $a_i \in A$ و L تعداد کلمات حاشیه‌نویسی می‌باشند. حال یک تصویر به همراه کلمات حاشیه‌نویسی آن بصورت ترکیبی از کلمات بصری و کلمات حاشیه‌نویسی می‌توانند بازنمایی شوند:

$$\begin{aligned} v^A &= [v_1^A, v_2^A, \dots, v_D^A, v_{D+1}^A, \dots, v_{D+L}^A] \\ &= [v_1^R, v_2^R, \dots, v_D^R, a_1, a_2, \dots, a_L] \end{aligned} \qquad (۱۲)$$

بطور خاص در این مدل کلمات حاشیه‌نویسی بصورت توام با کلمات بصری اندیس‌دهی می‌شوند و همچنین برگ‌هایی از درخت دودویی برای اندیس کلمات حاشیه‌نویسی تخصیص داده می‌شود.

۲-۲- ساخت کلمات بصری

در این بخش سه روش کد کردن سبد کلمات، روش SPM و LLC جهت ساخت کلمات بصری شرح داده شده است.

۲-۲-۱- سبد کلمات

این روش که از روش‌های مبتنی بر فرهنگ لغت محسوب می‌شود به صورت زیر عمل می‌کند:

۱- استخراج نقاط کلیدی مربوط به همه تصاویر آموزشی (از همه دسته‌ها)

۲- حصول توصیف‌گر حول تمام نقاط کلیدی

۳- خوشه‌بندی تمام توصیف‌گرهای بدست آمده با روش k-means برای حصول فرهنگ‌لغت. اعضای این فرهنگ لغت همان کلمات بصری هستند که فرض می‌شود هر تصویر با استفادهاز این کلمات بصری ایجاد شده است (در پردازش زبان‌ها طبیعی تفسیر به این گونه است که هر جمله از تعدادی کلمه در فرهنگ لغت استفاده می‌کند و در اینجا هر تصویر از تعدادی کلمه بصری).

پس از حصول فرهنگ لغت، زمان آن می‌رسد که بازنمایی جدیدی برای هر یک از تصاویر موجود در مجموعه دادگان مورد استفاده بدست آید. این روند برای هر تصویر در ذیل آورده شده است:

۱- استخراج نقاط کلیدی مربوط به تصویر و حصول توصیف‌گرهای حول تمام این نقاط کلیدی (همانند فاز یادگیری فرهنگ لغت)

۲- مرحله کد کردن: یافتن نزدیک‌ترن کلمه بصری موجود در لغت نامه به هر توصیف‌گر و تخصیص عدد مربوط به آن کلمه بصری (یا همان مرکز خوشه) به آن توصیف‌گر .

۳- ایجاد هیستوگرام مبتنی بر تعداد کلمات بصری موجود در تصویر.

هیستوگرام به دست آمده از روش فوق بازنمایی جدید آن تصویر را نمایش می‌دهد [۱۷]. در ادامه روش‌های دیگری بیان می‌شود که ساختار دیگری را برای ساخت کلمات بصری بیان می‌کنند.

۲-۲-۲- روش تطبیق هرمی مکانی (SPM)

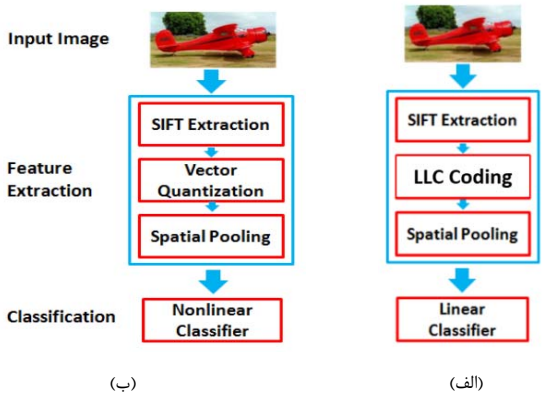
در روش سبد کلمات، اطلاعات مکانی نقاط کلیدی مورد توجه قرار گرفته نمی‌شود (این امر از نام سبد کلمات نیز بر می‌آید). بنابر این، یک منبع اطلاعاتی ارزشمند که می‌توانست در دسته‌بندی تصاویر کارا باشد مورد استفاده قرار نمی‌گرفت. در روش تطبیق هرمی مکانی، فاز یادگیری فرهنگ لغت همچون روش اصلی سبد کلمات انجام می‌گرفت، اما این بار با یک ایده ساده، چند سطح مختلف برای تصویر در نظر گرفته شده است که در هر سطح، تصویر به چند بخش با اندازه مساوی تقسیم می‌شود و برای هر بخش یک هیستوگرام مبتنی بر کلمات بصری ایجاد می‌شود.

در انتها هیستوگرام‌های حاصل با اعمال یک ضریب (ضریبی که برای سطوح بالاتر، بزرگتر و برای سطوح پایین‌تر، کوچکتر است تا اهمیت هیستوگرام‌هایی که اطلاعات مکانی بیشتر دارند افزایش یابد)، به همدیگر الحاق شده تا بازنمایی نهایی را شکل دهند. پس از حصول بازنمایی نهایی در روش تطبیق هرم مکانی، از یک دسته‌بندی کننده قوی همچون ماشین بردار پشتیبان برای دسته‌بندی تصویر استفاده شده است. این ایده ساده توانست دقت را به نسبت روش سبد کلمات تا حد بسیارخوبی افزایش دهد [۲۰].

۲-۲-۳- کد کردن LLC

روش اولیه تطبیق هرم مکانی برای عملکرد خوب نیاز داشت تا از یک دسته‌بندی کننده غیرخطی (برای مثال SVM با کرنل توابع پایه گوسی[۱۹]) استفاده کند [۲۱]. با توجه به اینکه مرتبه زمانی آموزش یک دسته‌بند SVM غیرخطی از درجه $O(n^2)$ تا $O(n^3)$ بود (n برابر تعداد تصاویر آموزشی است)، بنابراین توسعه‌پذیری روش سنتی تطبیق هرم مکانی پایین بود. روش LLC با جزئیاتی که در ادامه شرح داده می‌شود، با ایجاد یک کد جدید برای هر یک از توصیف‌گرهای صحنه، موفق شد تا به یک بازنمایی جدید از تصویر برسد که این بازنمایی با یک دسته‌بندی کننده خطی که با مرتبه زمانی $O(n)$ هم قابل آموزش بود، موفق شد که زمان آموزش را به شکل خوبی کاهش، و دقت حاصل را افزایش دهد.

پیش از پرداختن به نحوه حصول کد توسط روش LLC بهتر است تا دقیقاً جایگاه تغییر ایجاد شده در روش تطبیق هرم مکانی را مشخص کنیم. تفاوت این دو روش در شکل ۱ نمایش داده شده است.

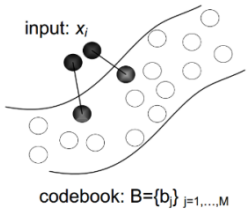


شکل ۱- محل ایجاد تغییر در LLC به نسبت روش تطبیق هرم مکانی (اقتباس از [۲۲]). الف) LLC خطی ب) SPM غیرخطی

همانطور که در شکل مشاهده می‌شود، در روش LLC که در ستون راست تصویر واقع شده، کد کردن LLC جای کوانتیزاسیون برداری[۲۰] را گرفته است. روش کوانتیزاسیون برداری که در سبد کلمات و تطبیق هرم مکانی وجود داشت به این صورت بود که هنگام تخصیص یکی از کلمات بصری فرهنگ لغت به یکی از توصیف‌گرها، تنها نزدیک‌ترین کلمه بصری در نظر گرفته شده و مابقی کلمات بصری نادیده گرفته می‌شدند. به زبان ریاضی، اگر فرض کنیم هر تصویر دارای N نقطه کلیدی است که توصیف‌گرهایش را بدست آورده‌ایم، و هر توصیف‌گر قرار است با یک کد M بعدی جایگزین شود (M تعداد کلمات بصری فرهنگ لغت است)، می بایست رابطه بهینه‌سازی زیر حل شود:

$$arg\ \min_{c} \sum_{i=1}^{N} \|x_i - Bc_i\|^2$$
$$s.t.\ \|c_i\|_{l^0} = 1\ .\ \|c_i\|_{l^1} = 1\ .\ c_i \geq 0.\forall i \qquad (۱۳)$$

که در رابطه فوق، $B = \{b_j\}_{j=1,\dots,M}$ نمایش دهنده فرهنگ لغت و c_i کد جدیدی بوده که قرار است جایگزین توصیف‌گر x_i شود. در صورت حل رابطه بهینه‌سازی فوق که شرایط نرم l_0 و نرم l_1 آن آورده شده است، درواقع گویی برای توصیف‌گر ورودی x_i به دنبال نزدیک‌ترین عضو فرهنگ لغت بوده‌ایم که مفهوم آن در شکل ۲ به تصویر کشیده شده است.



شکل ۲- تخصیص نزدیکترین عضو فرهنگ لغات به توصیف‌گر ورودی در روش کوانتیزاسیون برداری [۲۱]

اما روش LLC برای یافتن کد مربوط به هر توصیف‌گر سعی در بهینه‌سازی رابطه دیگری دارد. این رابطه بهینه‌سازی در ذیل آورده شده است:

$$\arg\min_{c} \sum_{i=1}^{N} \|x_i - Bc_i\|^2 + \lambda \|d_i \odot c_i\|^2$$
$$s.t.\ 1^T c_i = 1.\forall i \qquad (۱۴)$$

که در رابطه فوق:

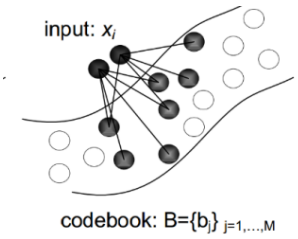
$$d_i = \exp\left(\frac{dist(x_i.B)}{\sigma}\right)$$
$$dist(x_i.B) = [dist(x_i.b_1).\dots.dist(x_i.b_M)]^T \qquad (۱۵)$$

ترم اول این رابطه بهینه‌سازی با قیدی که ذکر شده است، در صدد است تا به جای اینکه همچون روش کوانتیزاسیون برداری که یک کد M بعدی به صورت $[0,0,\dots,1,0,\dots,0]$ که عنصر مربوط به نزدیک‌ترین کلمه بصری آن ۱ است را تولید می‌کند، یک کد M بعدی تولید کند که خطای کوانتیزاسیون را با دخیل کردن تمام کلمات بصری کمینه می‌کند. اما ترم دوم رابطه بهینه‌سازی سعی دارد تا معیار مجاورت[۲۱] را ارضا کند و این امر باعث می‌شود تا ضریب مربوط به کلمات بصری نزدیک‌تر در کد حاصل بزرگتر شود. پارامتر σ نحوه محاسبه فاصله بین توصیف‌گر و اعضای فرهنگ لغت را کنترل می‌کند و به ازای یک σ ثابت نیز، λ میزان وزن‌دهی به مجاورت را مشخص می‌کند. هر چه λ بزرگتر شود وزن مجاورت نیز بزرگتر می‌شود. به این معنا که تنها به همسایه‌های نزدیک توصیف‌گر x_i در فرهنگ لغت ضرایب غیر صفر می‌دهیم. در شکل ۳ مثالی از نحوه کد کردن روش LLC آورده شده است. یال‌های متصل به اعضای فرهنگ لغت نشان دهنده این هستند که این اعضا به طور خطی و با وزن‌هایی که برایشان طی حل مسأله بهینه‌سازی بدست می‌آید، توصیف‌گر x_i را بازسازی می‌کنند.

یکی دیگر از نکات جذاب در مورد روش LLC این است که رابطه بهینه‌سازی ارائه شده یک راه حل تخمینی از مرتبه $O(M+K^2)$ به جای $O(M^2)$ دارد که M تعداد کلمات بصری در فرهنگ لغت و K تعداد همسایه‌های در نظر گرفته شده برای وزن دهی در LLC است که $K \ll M$ است. برای این کار بجای بهینه کردن رابطه (۱۴) می‌توان از روش k نزدیکترین همسایه جهت پیدا کردن توصیف‌گرهای مجاور استفاده نمود و رابطه زیر که بسیار ساده تر است را حل نمود.

$$\min_{\tilde{c}} \sum_{i=1}^{N} \|x_i - \tilde{c}_i B_i\|^2$$
$$s.t.\ 1^T \tilde{c}_i = 1,\ \forall i \qquad (۱۶)$$

این موضوع، استفاده از این روش در کاربردهای بلادرنگ را نیز امکان‌پذیر کرده است.



شکل ۳- وزن‌دهی بیشتر به همسایه‌های نزدیک توصیف‌گر در لغت نامه در روش LLC [۲۱]

آخرین مطلبی که باید ذکر شود، این است که پس از حصول کدهای مربوط به روش LLC برای هر توصیف‌گر، بازنمایی نهایی براساس انباشت این کدها صورت می‌گیرد که در زیر نحوه‌ی انباشت حداکثری که بهترین روش انباشت در LLC است نشان داده شده است.

انباشت حداکثر: اگر c_{ij} نشان‌دهنده عنصر jام مربوط به کد توصیف‌گر iام باشد، انباشت حداکثر به صورت ذیل است:

$$\text{finalCode}_j = \max(c_{1j}, c_{2j}, \dots, c_{Nj}) \qquad (۱۷)$$

۳-روش پیشنهادی

در شکل ۴ بلاک دیاگرام مربوط به مدل مورد نظر جهت دسته‌بندی و حاشیه‌نویسی تصویر نشان داده شده است. در قسمت پیش‌پردازش ابتدا هر تصویر تکه‌تکه شده و به توصیف‌گر داده می‌شود. پس از حصول خروجی توصیف‌گرهای مربوط به تصویر توسط کدگذار بازنمایی مربوط به هر تصویر تولید می‌شود. در بخش آموزش کلمات حاشیه‌نویسی در کنار کلمات بصری تعبیه شده‌اند و به عنوان بردار ویژگی به مدل SupDocNADE داده شده است. در قسمت آزمایش مدل فقط خروجی حاصل از توصیف‌گر تکه‌های تصویر به مدل داده شده و برچسب کلاس به همراه کلمات حاشه نویسی توسط مدل تولید می‌شود.

در عمل تعداد کلمات بصری استخراج شده از تصویر بسیار بزرگتر از کلمات حاشیه‌نویسی است. به عنوان مثال از یک تصویر با اندازه ۳۰۰×۴۰۰ حدود ۲۰۰۰ کلمه بصری استخراج می‌شود (شکل ۴ قسمت پیش‌پردازش) این در حالی است که کلمات حاشیه‌نویسی بین ۵ الی ۲۰ عدد متغیر می‌باشند.

عدم‌تعادل بین کلمات بصری و کلمات حاشیه‌نویسی ممکن است سبب برخی مشکلات شود. به عنوان مثال سهم کلمات حاشیه‌نویسی برای بازنمایی در لایه پنهان بسیار کمتر از کلمات بصری است. از طرفی هر کلمه به نسب تعداد تکرار خود به کل کلمات استخراج شده بر روی گرادیان خطا تاثیر می‌گذارد. پس گرادیانی که از کلمات حاشیه تولید می‌شود بسیار کوچک بوده تا بتواند تاثیر با معنایی در افزایش احتمال شرطی حاصل از کلمات حاشیه داشته باشد.

برای حل این مشکل، پیشنهاد می‌شود که کلمات حاشیه‌نویسی در هیستوگرام بردار ویژگی $v_{<i}$ وزن‌دهی شوند. برای این منظور یک بردار ω به همراه v در نظر گرفته می‌شود. حال بردار ورودی به صورت زیر تعریف می‌شود:

$$\tilde{\mathbf{v}} = \left[\left(v_1, \omega(v_1)\right), \left(v_2, \omega(v_2)\right), \dots, \left(v_{D+L}, \omega(v_{D+L})\right)\right] \qquad (۱۸)$$

بردار ω برای مولفه‌های متناظر با کلمات بصری برابر ۱ و برای مولفه‌های متناظر با کلمات حاشیه برابر ρ می‌باشد. حال رابطه (۱۰) بصورت زیر بازنویسی می‌شود:

$$-\log p(v,y) = -\log p(y|v) - \lambda \sum_{i=1}^{D} \omega(v_i) \log p(v_i|v_{<i})) \qquad (۱۹)$$

و در رابطه (۲) نیز بجای v از $\tilde{v}$ استفاده می‌شود.

با وزن‌دهی کلمات حاشیه، مدل با توجه بیشتر به این کلمات سبب کاهش مشکلات عدم‌تعادل بین کلمات بصری و کلمات حاشیه‌نویسی خواهد شد. در عمل مقدار وزن ρ توسط اعتبارسنجی متقابل[۲۲] مشخص می‌شود. در الگوریتم ۱ شبه کد بروزرسانی گرادیان پارامترهای رابطه (۱۹) براساس روش پیشنهادی نشان داده شده است.

الگوریتم ۱- نحوه محاسبه گردایان مدل برروی داده‌های آموزشی

Input: training vector $\mathbf{v}$, training weight vector ω_D
unsupervised learning weight λ
Output: gradient of Equation 14 w. r. t parameters
$f(\mathbf{v}) \leftarrow \text{softmax}(\mathbf{d} + \mathbf{U}\mathbf{h}^{c}(\mathbf{v}))$
$\delta\mathbf{d} \leftarrow (f(\mathbf{v}) - 1_y)$
$\delta\mathbf{act} \leftarrow (\mathbf{U}^T\delta\mathbf{d}) o 1_{h_y>0}$
$\delta\mathbf{c} \leftarrow 0, \delta\mathbf{b} \leftarrow 0, \delta\mathbf{V} \leftarrow 0, \delta\mathbf{W} \leftarrow 0$
for i from D to 1 do
  $\delta\mathbf{h}_i \leftarrow 0$
for m from 1 to $|\pi(v_i)|$ do
  $\delta t \leftarrow \frac{\lambda\omega(v_i)(p(\pi(v_i)_m|\mathbf{v}_{<i}) - \pi(v_i)_m)}{D}$
  $\delta b_{l(v_i)_m} \leftarrow \delta b_{l(v_i)_m} + \delta t$
  $\delta\mathbf{V}_{l(v_i)_m,:} \leftarrow \delta\mathbf{V}_{l(v_i)_m,:} + \delta t\mathbf{h}_i^T$
  $\delta\mathbf{h}_i \leftarrow \delta\mathbf{h}_i + \delta t\mathbf{V}_{l(v_i)_m,:}^T$
end for
  $\delta\mathbf{act} \leftarrow \delta\mathbf{act} + \delta\mathbf{h}_i o 1_{\mathbf{h}_i>0}$
  $\delta\mathbf{c} \leftarrow \delta\mathbf{c} + \delta\mathbf{h}_i o 1_{\mathbf{h}_i>0}$
  $\delta\mathbf{W}_{:,v_i} = \omega(v_i)\delta\mathbf{W}_{:,v_i} + \delta\mathbf{act}$
end for

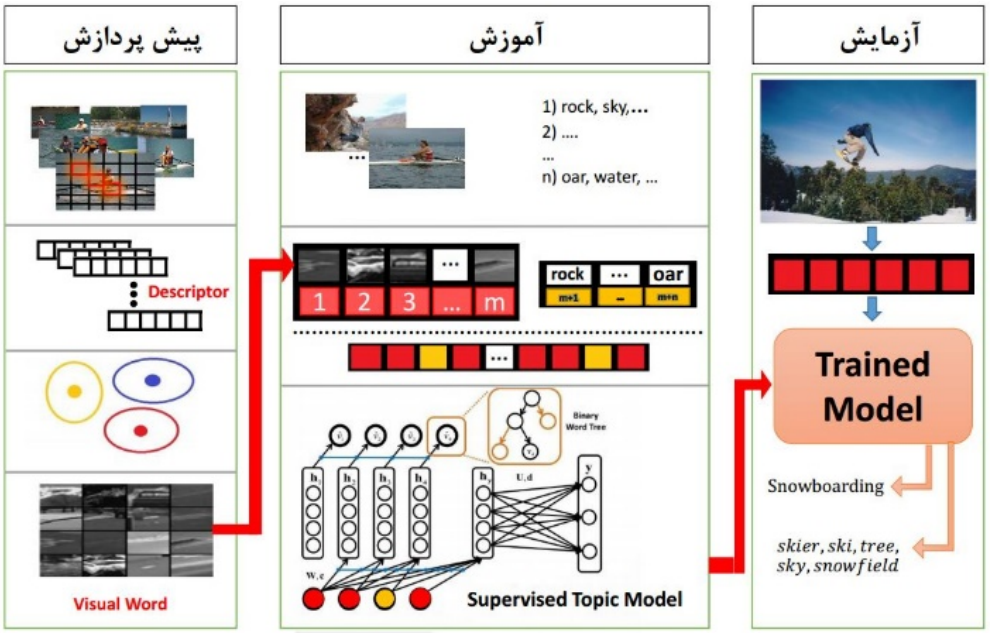
با افزودن قابلیت وزن‌دار شدن ورودی‌های مدل، می‌توان از LLC استفاده نمود. همانطور که در بخش ۲-۲-۳ بیان شد روش LLC برخلاف روش‌های سنتی که تنها از یک کلمه جهت بازنمایی استفاده می‌نمایند، چندین کلمه مشابه در فرهنگ لغت را بصورت وزن‌دار در ساخت بردار ویژگی دخیل می‌کند. حال هر تکه از تصویر توسط یک بردار وزن بازنمایی شده است. پس از اعمال توصیف‌گر و حصول بازنمایی، انباشت حداکثری روی بردار نهایی طبق رابطه (۱۷) اعمال می‌شود. بردار حاصل از انباشت به عنوان کلمات بصری بوده و مقادیر درایه‌های آن وزن ورودی می‌باشد. این کلمات بصری با کلمات حاشه‌نویسی و وزن‌های آن به عنوان ورودی به مدل داده می‌شود.

۴- نتایج و آزمایشات

برای ارزیابی عملکرد مدل پیشنهادی در دسته‌بندی و حاشیه‌نویسی تصاویر از دو پایگاه‌دادهی LabelMe [۲۳] و UIUC-Sports [۱] استفاده شده است. این پایگاه داده‌ها به عنوان مجموعه داده‌های پایه جهت حاشیه‌نویسی و دسته‌بندی می‌باشند که هر تصویر به همراه کلمات حاشیه‌نویسی موجود است. مدل پیشنهادی با مدل‌های SupDocNADE [۱۶]، DocNADE [۱۵] و Mc_sLDA [۷] مقایسه شده است.

۴-۱- معرفی مجموعه داده‌ها

در این مقاله از دو مجموعه داده استفاده شده است. مجموعه داده اول UIUC-Sports می‌باشد که شامل ۱۷۹۲ تصویر در ۸ دسته ورزشی تقسیم شده است.



شکل ۴- بلوک دیاگرام مربوط به مدل جهت دسته‌بندی و حاشیه‌نویسی تصویر

برچسب هر کلاس به علاوه تعداد تصاویر بصورت ۱- بدمینتون (۳۱۳ تصویر) ۲- بوچی (۱۳۷ تصویر) ۳- کراکت (۳۳۰ تصویر) ۴- چوگان بازی (۱۸۳ تصویر) ۵- صخره نوردی (۱۹۴ تصویر) ۶- قایقرانی (۲۵۵ تصویر) ۷- موج‌سواری (۱۹۰ تصویر) ۸- اسکی (۱۹۰ تصویر) است. که تصاویر مطابق [۱۶] به عرض ۴۰۰ پیکسل تغییر اندازه داده شده است.

پایگاه داده LabelMe بطور آنلاین و با استفاده از جعبه ابزارهای موجود آن قابل دسترس است. در این مقاله از مجموعه یکسان با [۱۶] استفاده شده است. در این مجموعه داده ۸ کلاس: بزرگراه، درون‌شهری، ساحل، جنگل، ساختمان بلند، خیابان‌ها، برون‌شهری با تعداد ۲۰۰ تصویر به ازای هر کلاس موجود می‌باشد. در هر دو این مجموعه داده‌ها کلماتی که کمتر از ۳ بار در متن حاشیه‌نویسی ظاهر شده‌اند از پایگاه داده حذف شده است.

۴-۲- پارامترهای مدل و معیارهای ارزیابی

با توجه به [۷] از توصیف‌گر سیفت ۱۲۸ بعدی جهت استخراج کلمات بصری استفاده شده است. اندازه هر تکه‌ی[۲۳] تصویر که عملگر سیفت روی آن اعمال می‌شود برابر ۱۶×۱۶ پیکسل می‌باشد و اندازه قدم برای در نظر گرفتن تکه بعدی ۸ پیکسل است. از هر تصویر که در داده‌های آموزشی قرار دارد ۱۰۰ خروجی به تصادف انتخاب شده و به خوشه‌بند k-means داده می‌شود. در این مقاله ۲۴۰ مرکز خوشه به عنوان فرهنگ‌لغت کلمات بصری در نظر گرفته شده است. هر تصویر به شبکه‌های ۲×۲ تقسیم شده و تعداد ۲۴۰×۲×۲=۹۶۰ کلمه بصری مختلف برای فرهنگ لغات را تشکیل می‌دهند.

از دقت دسته‌بندی جهت ارزیابی دسته‌بند مدل و از میانگین معیار F- پنج کلمه محتمل برای ارزیابی حاشیه‌نویسی مدل استفاده شده است. معیار F- بصورت زیر تعریف می‌شود.

$$\text{F} - \text{measure} = \frac{2\times\text{Precision}\times\text{Recall}}{\text{Precision}+\text{Recall}} \quad (۱۵)$$

در این مقاله همانند [۱۶] از ماشین بردار پشتیبان جهت دسته‌بندی تصاویر استفاده شده است. لایه مخفی h_y که توسط کلمات بصری و کلمات وزن‌دار حاشیه‌نویسی آموزش داده شده به عنوان ورودی به دسته‌بند ماشین بردار پشتیبان داده می‌شود. پارامترهای این دسته‌بند توسط اعتبارسنجی متقابل محاسبه می‌شود.

۴-۳- نتایج آزمایشات

در این بخش به بررسی تاثیر پارامترهای مختلف روش پیشنهادی بر دقت آن پرداخته شده و راهکارهایی برای تنظیم این پارامترها ارائه می‌گردد. در ادامه نتایج حاصل از آزمایشات مختلف شرح داده شده است.

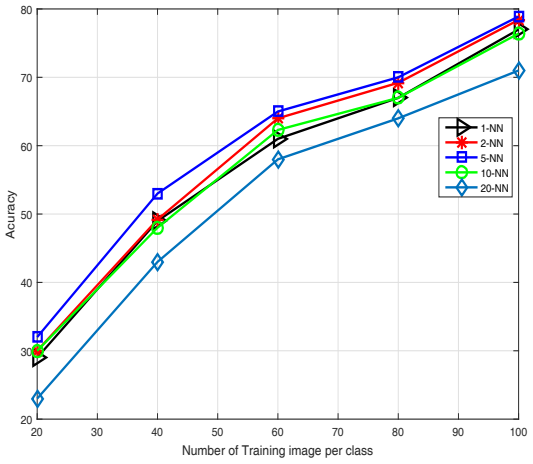
۴-۳-۱- تنظیم پارامترها

ابتدا جهت بدست آوردن مقدار k همسایه مناسب جهت بازنمایی هر تکه از تصویر در LLC تخمینی، دقت دسته‌بندی به ازای تعداد K مختلف مورد ارزیابی قرار گرفته است. مقادیر ۱، ۲، ۵، ۱۰ و ۲۰ همسایه مختلف برای K در نظر گرفته شده که در شکل ۵ و ۶ به ترتیب برای پایگاه داده‌ی UIUC_Sports و LabelMe دقت دسته‌بندی به ازای تعداد تصاویر در نظر گرفته شده به ازای هر کلاس جهت آموزش مورد ارزیابی قرار گرفته است. لازم به ذکر است زمانی که تعداد همسایه‌ها برابر یک باشد روش LLC جواب یکسانی با روش کیسه کلمات دارد.

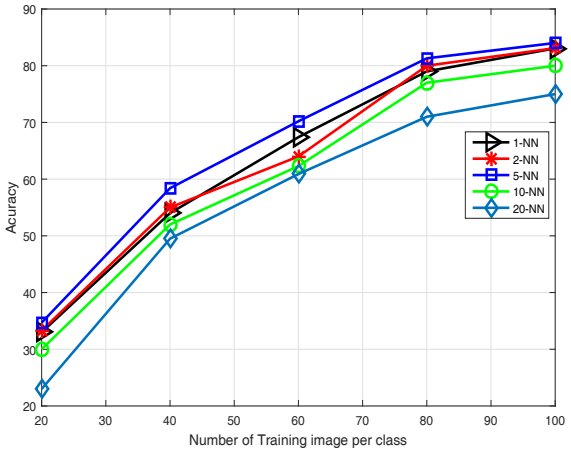
همانطور که مشاهده می شود افزایش تعداد همسایه از یک تا ۵ باعث افزایش دقت روش پیشنهادی می‌شود اما افزایش بیشتر آن باعث کاهش کارآیی شده است. از سویی افزایش تعداد نمونه‌های آموزشی تا حدود ۸۰ نمونه به ازاء هر کلاس باعث افزایش دقت شده است. از آنجا که افزایش تعداد همسایه‌ها باعث افزایش بار محاسباتی و کاهش سرعت اجرای برنامه می‌شود، مقادیر کمتر این پارامتر که دقت مناسبی را فراهم سازند توصیه می‌شود.

برای بدست آوردن وزن مناسب کلمات حاشیه‌نویسی، از روش اعتبارسنجی متقابل استفاده شده است. الگوریتم به ازای وزن‌های مختلف با در نظر گرفتن

روش LLC با ۵ و ۱ همسایه اجرا شده که در شکل ۷ و شکل ۸ نتایج معیار F برروی داده‌های تست نشان داده شده است. همانطور که در این نمودارها مشخص است برای پایگاه داده‌ی UIUC_Sports وزن ۵۰۰ و در حالتی که از کدگذار LLC استفاده می‌شود وزن ۲۰۰ مناسب است. برای پایگاه داده LabelMe وزن ۳۰۰ و در حالتی که از کدگذار LLC استفاده شده است وزن ۱۰۰ مناسب می‌باشد. با استفاده از وزن‌های بدست آمده در مرحله قبل یادگیری مدل برروی داده‌های آموزشی صورت می‌گیرد.



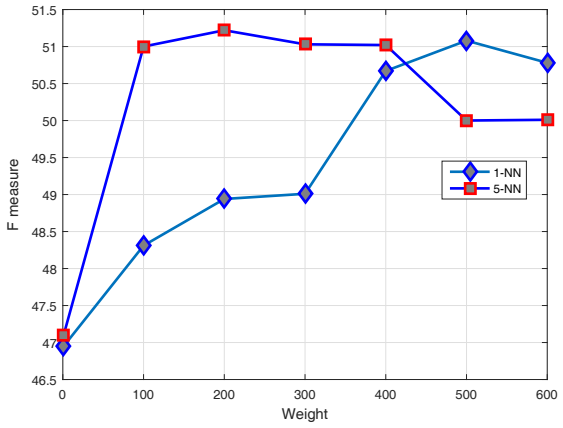
شکل ۵- نمودار دقت دسته‌بندی پایگاه داده UIUC_Sports به ازای تعداد همسایه‌های در نظر گرفته شده در روش LLC



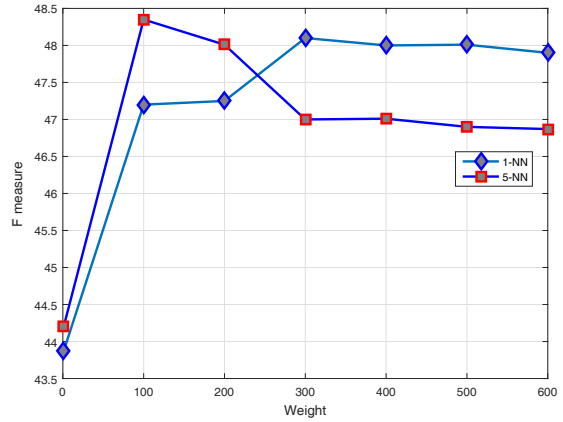
شکل ۶- نمودار دقت دسته‌بندی پایگاه داده LabelMe به ازای تعداد همسایه‌های در نظر گرفته شده در روش LLC

۴-۳-۲- ارزیابی نتایج

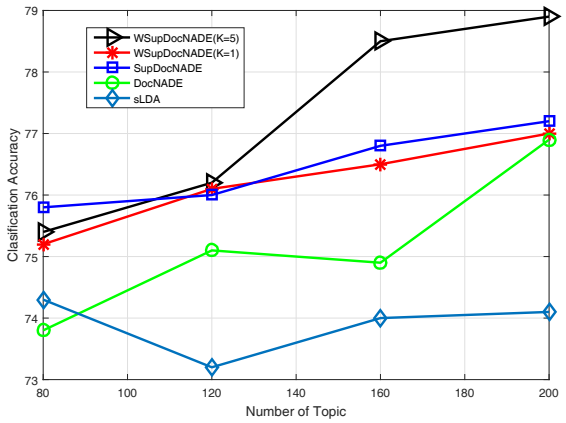
نتایج حاصل از دسته‌بندی برروی پایگاه داده‌های موردنظر در شکل ۹ و شکل ۱۰ نشان داده شده است. همانطور که مشاهده می‌شود با افزایش تعداد موضوعات مختلف برای روش‌های مختلف دسته‌بندی و حاشیه‌نویسی دقت دسته‌بندی به مراتب افزایش یافته است. زمانی که از روش کیسه کلمات بدون LLC استفاده شود دقت دسته‌بندی بهبودی نسبت به مدل SupDocNADE ندارد ولی زمانی که LLC با در نظر گرفتن ۵ همسایه جهت کدگذاری استفاده شده است شاهد بهبود حداقل ۱ درصدی در دقت دسته‌بندی نسبت به قبل هستیم.



شکل ۷- نمودار معیار F پایگاه داده UIUC_Sports به ازای وزن‌دهی‌های مختلف

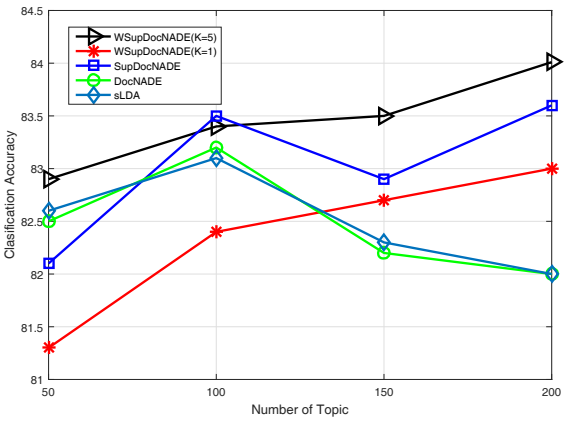


شکل ۸- نمودار معیار F برای پایگاه داده LabelMe به ازای وزن‌دهی‌های مختلف

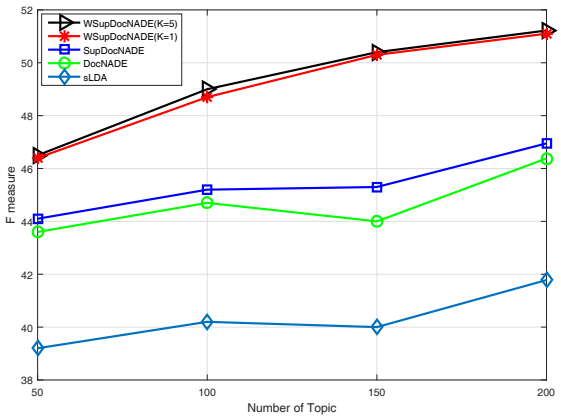


شکل ۹- نمودار دقت دسته‌بندی پایگاه داده UIUC_Sports به ازای تعداد واحدهای پنهان (تعداد عناوین)

شکل ۱۱ و شکل ۱۲ نتایج حاصل از حاشیه‌نویسی مدل را نشان داده است. همانطور که مشاهده می‌شود مدل پیشنهادی معیار F ۵۱.۲۲ درصد برای داده‌های UIUC_Sports و برای دادهای LabelMe مقدار ۴۸.۳۵ را کسب کرده است. با توجه به این نمودارها بهبود ۵ درصدی در معیار F کلمات حاشیه‌نویسی مشاهده می‌شود. جدول ۱ بهترین نتایج حاصل از دسته‌بندی و حاشیه‌نویسی مدل‌های مختلف را نشان می‌دهد.



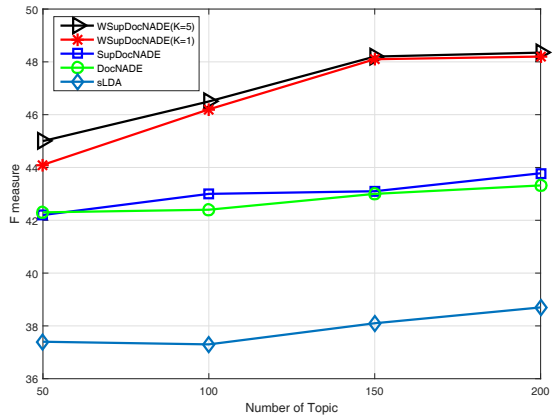
شکل ۱۰- نمودار دقت دسته‌بندی پایگاه داده LabelMe به ازای تعداد واحدهای پنهان (تعداد عناوین)



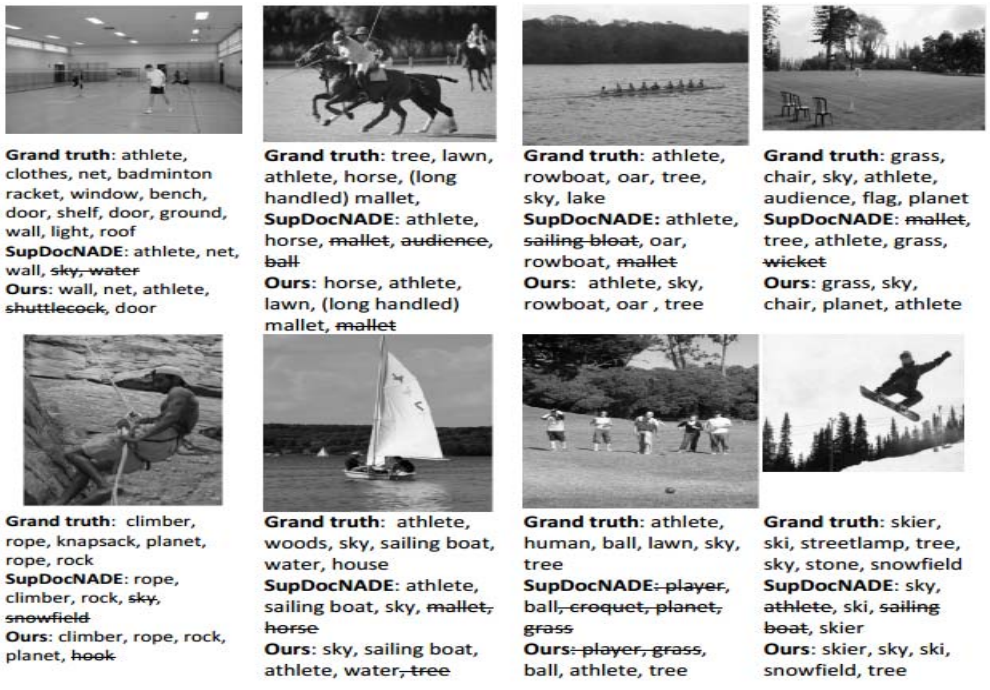
شکل ۱۱- نمودار معیار F برای پایگاه داده UIUC_Sports به ازای تعداد واحدهای پنهان (تعداد عناوین)

جدول ۱- مقایسه کارایی مدل پیشنهادی با روش‌های دیگر

مدل	UIUC-Sport		LabelMe	
	F-mesure	ACC	F-mesure	ACC
sLDA [۷]	٪ ۳۸/۰	٪ ۷۶/۸۷	٪ ۳۸/۷	٪ ۸۱/۸۷
DocNADE [۱۵]	٪ ۴۶/۳۸	٪ ۷۴/۲۳	٪ ۴۳/۳۲	٪ ۸۱/۹۷
SupDocNADE [۱۶]	٪ ۴۶/۹۵	٪ ۷۷/۲۹	٪ ۴۳/۸۷	٪ ۸۳/۶
روش پیشنهادی (k=1)	٪ ۵۱/۰۸	٪ ۷۷/۰	٪ ۴۸/۲	٪ ۸۳/۰۱
روش پیشنهادی (k=5)	**٪ ۵۲/۲۲**	**٪ ۷۸/۹**	**٪ ۴۸/۳۵**	**٪ ۸۴/۰۱**



شکل ۱۲- نمودار معیار F برای پایگاه داده LabelMe به ازای تعداد واحدهای پنهان (تعداد عناوین)



شکل ۱۳- برای هر تصویر کلمات حاشیه درست (Grand truth)، کلمات حاصل از روش SupDocNADE و کلمات حاصل از روش پیشنهادی این مقاله بیان شده است. کلمات خط خورده کلمات پیشنهادی اشتباه هستند

در شکل ۱۳ نمونه‌ای از حاشیه‌نویسی کلمات توسط مدل پیشنهادی و مدل [۱۶] نشان داده شده است. همانطور که مشاهده می‌شود کیفیت کلمات حاشیه‌نویسی در مدل پیشنهادی بهبود یافته است. بر خلاف روش SupDocNADE روش پیشنهادی از تکرار کلمات حاشیه‌ای نظیر mallet و athlete پرهیز کرده و در کلیه تصاویر بررسی شده تعداد کلمات حاشیه‌ای پیشنهاد شده‌ی اشتباه کمتری داشته است.

۵- نتیجه‌گیری

در این مقاله ابتدا مدل موضوعی SupDocNADE معرفی شد که نتایج خوبی در مدل‌کردن داده‌های چند مقداری مانند دسته‌بندی و حاشیه‌نویسی تصاویر ارائه داده است. در این مدل کلمات حاشیه‌نویسی در کنار کلمات بصری تعبیه شده و به عنوان بردار ویژگی برای شبکه در نظر گرفته می‌شود. در عمل تعداد ویژگی‌های استخراج شده از تصویر بسیار بزرگتر از ویژگی‌هایی است که از کلمات حاشیه‌نویسی بدست می‌آیند. عدم‌تعادل بین کلمات بصری و حاشیه‌نویسی سبب می‌شود تا سهم کلمات حاشیه‌نویسی برای بازنمایی در لایه پنهان شبکه عصبی مورد استفاده در این مدل، بسیار کمتر از کلمات بصری باشد. در این مقاله برای حل مشکل فوق، از وزن‌دهی ویژگی‌های ورودی شبکه استفاده شده است به این صورت که به کلمات حاشیه‌نویسی شده وزن بیشتری نسبت به کلمات بصری استخراج شده از تصویر داده شود تا فراوانی کم این کلمات نسبت به کلمات بصری جبران شود. از طرفی با افزودن قابلیت وزن‌دار کردن ورودی از روش کدگذاری LLC به جای روش سنتی کوانتیزاسیون برداری استفاده شد. نتایج نشان‌دهنده بهبود ۵ درصدی در معیار F مدل در حاشیه‌نویسی تصاویر و بهبود حداقل ۱ درصدی در دقت دسته‌بندی تصاویر است. در ادامه می‌توان از روش‌های پیشرفته‌تری به منظور استخراج ویژگی‌های تصویر استفاده کرد و همچنین می‌توان مفهوم توجه[۲۴] را در تولید کلمات حاشیه‌نویسی مد نظر قرار داد.

مراجع

[1] L.-J. Li, and L. Fei-Fei, "What, where and who? classifying events by scene and object recognition," in *Computer Vision IEEE 11th International Conference on*, pp. 1–8, 2007.

[2] D. M. Blei, and M. I. Jordan, "Modeling annotated data," in *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pp. 127–134, 2003.

[3] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," in *Journal of machine Learning research*, vol. 3, no. Jan, pp. 993–1022, 2003.

[4] D. M. Blei, "Probabilistic topic models," in *Communications of the ACM*, vol. 55, no. 4, pp. 77–84, 2012.

[5] L. Fei-Fei, and P. Perona, "A bayesian hierarchical model for learning natural scene categories," in *Computer Vision and Pattern Recognition IEEE Computer Society Conference on*, vol. 2, pp. 524–531, 2005.

[6] J. D. Mcauliffe, and D. M. Blei, "Supervised topic models," in *Advances in neural information processing systems*, pp. 121–128, 2008.

[7] W. Chong, D. Blei, and F.-F. Li, "Simultaneous image classification and annotation," in *Computer Vision and Pattern Recognition IEEE Conference on*, pp. 1903–1910, 2009.

[8] L.-J. Li, R. Socher, and L. Fei-Fei, "Towards total scene understanding: Classification, annotation and segmentation in an automatic framework," in *Computer Vision and Pattern Recognition IEEE Conference on*, pp. 2036–2043, 2009.

[9] X. LI, C. SUN, L. U. Peng, X. WANG, and Y. ZHONG, "Simultaneous image classification and annotation based on probabilistic model," in *Journal of China Universities of Posts and Telecommunications,* vol. 19, no. 2, pp. 107–115, 2012.

[10] Y. Wang, and G. Mori, "Max-margin Latent Dirichlet Allocation for Image Classification and Annotation," in *BMVC*, vol. 2, no. 6, pp. 7, 2011.

[11] G. E. Hinton, and R. R. Salakhutdinov, "Replicated softmax: an undirected topic model," in *Advances in neural information processing systems*, pp. 1607–1614, 2009.

[12] R. Salakhutdinov, and I. Murray, "On the quantitative analysis of deep belief networks," in *Proceedings of the 25th international conference on Machine learning*, pp. 872–879, 2008.

[13] R. M. Neal, "Annealed importance sampling," *Statistics and computing*, vol. 11, no. 2, pp. 125–139, 2001.

[14] H. Larochelle, and I. Murray, "The neural autoregressive distribution estimator," in *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pp. 29–37, 2011.

[15] H. Larochelle, and S. Lauly, "A neural autoregressive topic model," in *Advances in Neural Information Processing Systems*, pp. 2708–2716, 2012.

[16] Y. Zheng, Y.-J. Zhang, and H. Larochelle, "Topic modeling of multimodal data: an autoregressive approach," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1370–1377, 2014.

[17] G. Csurka, C. Dance, L. Fan, J. Willamowski, and C. Bray, "Visual categorization with bags of keypoints," in *Workshop on statistical learning in computer vision, ECCV*, vol. 1, no. 1–22, pp. 1–2, 2004.

[18] J. Sivic, and A. Zisserman, "Video Google: A text retrieval approach to object matching in videos," in *Computer Vision Ninth IEEE International Conference on*, pp. 1470, 2003.

[19] K. Grauman, and T. Darrell, "The pyramid match kernel: Discriminative classification with sets of image

features," in *Computer Vision Tenth IEEE International Conference on*, vol. 2, pp. 1458–1465, 2005.

[20] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *Computer vision and pattern recognition IEEE computer society conference on*, vol. 2, pp. 2169–2178, 2006.

[21] J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, and Y. Gong, "Locality-constrained linear coding for image classification," in *Computer Vision and Pattern Recognition IEEE Conference on*, pp. 3360–3367, 2010.

[22] J. Yang, K. Yu, Y. Gong, and T. Huang, "Linear spatial pyramid matching using sparse coding for image classification," in *Computer Vision and Pattern Recognition IEEE Conference on*, pp. 1794–1801, 2009.

[23] B. C. Russell, A. Torralba, K. P. Murphy, and W. T. Freeman, "LabelMe: a database and web-based tool for image annotation," *International journal of computer vision*, vol. 77, no. 1, pp. 157–173, 2008.

سید نوید محمدی فومنی مدرک کارشناسی و کارشناسی‌ارشد خود را به ترتیب از دانشگاه زنجان در سال ۹۳ و دانشگاه صنعتی امیرکبیر در سال ۹۵ دریافت نموده است. پایان‌نامه ایشان در مورد استفاده از مدل‌های عنوان جهت دسته‌بندی و حاشیه‌نویسی تصاویر می‌باشد که از رویکرد شبکه‌های عصبی کانولوشنی جهت استخراج ویژگی استفاده شده است.

آدرس پست‌الکترونیکی ایشان عبارت است از:

navid_foumani@aut.ac.ir

احمد نیک‌آبادی عضو هیئت علمی دانشکده مهندسی کامپیوتر و فناوری اطلاعات دانشگاه صنعتی امیرکبیر است. وی مدرک کارشناسی‌ارشد و دکترای خود را در زمینه هوش مصنوعی از دانشگاه صنعتی امیرکبیر دریافت کرده است. زمینه‌های تحقیقاتی مورد علاقه ایشان مدل‌های احتمالاتی گرافی، پردازش تصویر و هوش محاسباتی است.

آدرس پست‌الکترونیکی ایشان عبارت است از:

nickabadi@aut.ac.ir

اطلاعات بررسی مقاله:

تاریخ ارسال: ۱۳۹۵/۱۰/۲۸

تاریخ اصلاح: ۱۳۹۵/۱۱/۱۵

تاریخ قبول شدن: ۱۳۹۵/۱۱/۳۰

نویسنده مرتبط: دکتر احمد نیک‌آبادی، دانشکده مهندسی کامپیوتر و فناوری اطلاعات، دانشگاه صنعتی امیرکبیر، تهران، ایران.

[1]Annotation
[2]Probabilistic Topic Models
[3]Latent Dirichlet Allocation
[4]Multimodal
[5]Visual Word
[6]Multi-Class Supervised Latent Dirichlet Allocation
[7]Exact Inference
[8]Intractable
[9]Posterior
[10]Replicated Softmax
[11]Partition Function
[12]Annealed Importance Sampling
[13]Neural Autoregressive Distribution Estimator
[14]Restricted Boltzmann Machine
[15]Autoencoder
[16]Bag-of-Feature
[17]Spatial Pyramid Matching
[18]Quantization
[19]Radial Basis Function
[20]Vector Quantization
[21]Locality
[22]Cross Validation
[23]Patch
[24]Attention

علوم رایانش و فناوری اطلاعات
نشریه علمی انجمن کامپیوتر ایران
مجلد ۱۴، شماره ۲، ۱۳۹۵
صفحات ۱۲- ۲۲
مقاله عادی

نشریه علوم رایانش و فناوری اطلاعات

بهبود ترجمه ماشینی مبتنی بر قاعده با استفاده از قواعد نحوی آماری

حکیمه فدائی هشام فیلی فرناز قاسمی تودشکی

دانشکده مهندسی برق و کامپیوتر، دانشگاه تهران، تهران، ایران

چکیده

ترجمه ماشینی مبتنی بر قاعده[1] از مجموعه‌ای از قواعد که دربردارنده اطلاعات زبانی هستند در فرایند ترجمه استفاده می‌کند. نتایج تولید شده توسط این مترجم‌ها معمولاً از نظر دستور زبان و ترتیب کلمات بهتر از نتایج مترجم‌های آماری هستند. ولی تحقیقات نشان داده است که این ترجمه‌ها از نظر روانی و انتخاب کلمات مناسب، ضعیف‌تر از مترجم‌های آماری هستند. در این مقاله هدف، بهبود انتخاب لغات در مترجم مبتنی بر قاعده است. این کار با استفاده از مجموعه‌ای از قواعد نحوی- لغوی مبتنی بر گرامر درخت- پیوندی[2] (TAG) انجام می‌شود. این قواعد احتمالاتی به‌صورت آماری از یک پیکره موازی با اندازه بزرگ استخراج شده‌اند. در سیستم ارائه شده، کلمات با ترتیب پیشنهادی مترجم مبتنی بر قاعده در زبان مقصد قرار می‌گیرند و به همین دلیل در ترجمه جملات از یک رمزگشای یکنواخت مبتنی بر برنامه‌ریزی پویا[3] استفاده شده است. در این سیستم بهترین ترجمه با استناد به احتمال قواعد استفاده شده و امتیاز مدل زبانی انتخاب می‌شود. آزمایش‌ها روی ترجمه انگلیسی به فارسی نشان داد که کیفیت نتایج به‌دست آمده از روش پیشنهادی حدود ۱/۳+ واحد بلو از کیفیت ترجمه به‌دست آمده توسط مبتنی بر قاعده پایه بالاتر است.

کلمات کلیدی: ترجمه ماشینی ترکیبی، ترجمه ماشینی مبتنی بر قاعده، قواعد آماری، قواعد نحوی-لغوی، گرامر درخت-پیوندی.

۱- مقدمه

ترجمه ماشینی یکی از شاخه‌های پرکاربرد و پیچیده در پردازش زبان طبیعی است. با توجه به گسترش روزافزون اسناد تحت وب، درخواست استفاده از سرویس‌های ترجمه ماشینی بسیار بالاست و شرکت‌های بزرگ فعال در این حوزه به‌طور مداوم در حال تلاش برای بهبود محصولات خود هستند. در سال‌های گذشته رویکردهای متفاوتی برای این مسئله ارائه شده است که هر کدام نقاط قوت و ضعف خاص خود را دارند. از میان این رویکردها می‌توان به ترجمه مبتنی بر قاعده، ترجمه آماری، ترجمه مبتنی بر مثال و ترجمه بر پایه شبکه‌های عصبی اشاره کرد. همچنین با ترکیب رویکردهای مختلف روش‌های ترکیبی[4] ارائه داد. این روش‌ها سعی در هم‌افزایی نقاط مثبت رویکردهای ترکیب شده دارند. در روش‌های ترکیبی معمولاً یکی از رویکردهای عنوان شده به‌عنوان رویکرد اصلی انتخاب می‌شود و فرایند ترجمه با تکیه بیشتر بر آن روش انجام می‌شود و روش‌های استفاده شده دیگر برای بهبود و کم کردن خطاهای روش پایه به کار می‌آیند. از روش‌های ترکیبی متداول در ترجمه ماشینی می‌توان به اضافه کردن اطلاعات زبانی و نحوی به مترجم‌های آماری و یا استفاده از عبارات استخراج شده در مترجم آماری برای غنی‌سازی مترجم مبتنی بر قاعده اشاره کرد.

رویکرد مبتنی بر قاعده به‌عنوان قدیمی‌ترین رویکرد در حوزه ترجمه ماشینی شناخته می‌شود و بر پایه مجموعه‌ای از قواعد که معمولاً توسط انسان ایجاد شده است استوار است. این مجموعه قواعد نحوه انتقال نحوی و لغوی از یک زبان به زبان دیگر را مدل‌سازی می‌کنند. از آنجایی‌که این مجموعه قواعد با نظارت انسان تولید می‌شوند استفاده از این رویکرد هزینه بسیار زیادی به همراه دارد. در ازای این هزینه، نتایج تولید شده توسط مترجم‌های مبتنی بر قاعده از نظر دستور زبان صحیح‌تر هستند و ترتیب کلمات در آن‌ها بهتر رعایت شده است. ترجمه ارائه شده توسط این مترجم‌ها، به دلیل در نظر گرفتن اطلاعات زبان‌شناسی، معمولاً از نظر تطابق فعل و فاعل، زمان، شخص و شمار افعال و خصوصیات ساخت‌واژی دیگر، بهتر از نتایج مترجم‌های آماری است. این در حالی است که مترجم‌های آماری از نظر انتخاب معادل مناسب برای کلمات در زبان مقصد، معمولاً بهتر از مترجم‌های مبتنی بر قاعده عمل می‌کنند. مترجم‌های آماری در انجام جابجایی‌های نزدیک

قوی‌تر هستند و به دلیل استفاده از مدل زبانی معمولاً ترجمه‌های روان‌تری نسبت به ترجمه مبتنی بر قاعده ارائه می‌دهند [۱]. بدین‌ترتیب اگر بتوان مترجم مبتنی بر قاعده را با استفاده از اطلاعات مترجم آماری غنی کرد، امید است که نتایج بهتری نسبت به مترجم مبتنی بر قاعده پایه به دست آید.

در این مقاله هدف ما استفاده از اطلاعات آماری در بهبود نتایج مترجم مبتنی بر قاعده است. بدین‌ترتیب که می‌خواهیم برای ترتیب قرارگیری کلمات در زبان مقصد به مترجم مبتنی بر قاعده استناد کنیم. از آنجایی که این مترجم‌ها از قواعد نحوی استفاده می‌کنند معمولاً در تشخیص جابجایی‌های دور بین کلمات قوی‌تر از مترجم‌های آماری عمل می‌کنند و از این نظر برای ترجمه بین زوج زبان‌های دور مانند انگلیسی و فارسی مناسب‌تر هستند. از آنجایی که بافت نحوی کلمه در زبان مبدأ می‌تواند در انتخاب معادل مناسب در زبان مقصد بسیار مفید باشد، ما مجموعه‌ای از قواعد همگام نحوی-لغوی را به‌صورت آماری استخراج کرده‌ایم و برای هر کلمه یا مجموعه از کلمات در جمله مبدأ باتوجه به بافت نحوی این کلمات و با استناد به این قواعد معادل مناسب را پیشنهاد می‌دهیم.

قواعد استخراج شده مبتنی بر گرامر درخت- پیوندی [۲] هستند و با توجه به میزان رخدادشان در پیکره آموزشی به هریک از آن‌ها احتمالی نسبت داده شده است. سیستم ما در نهایت ترجمه کلمات را از بین ترجمه‌های پیشنهاد شده توسط مترجم مبتنی بر قاعده و ترجمه‌های ارائه شده توسط مدل آماری انتخاب می‌کند. همچنین از آنجایی که ترتیب کلمات توسط مترجم مبتنی بر قاعده تعیین می‌شود، ترجمه به‌صورت یکنواخت[۵] انجام می‌شود و می‌توان برای انتخاب ترجمه بهینه از برنامه‌ریزی پویا استفاده کرد که پیچیدگی زمانی بسیار کمتری نسبت به الگوریتم‌های جستجو در ترجمه آماری دارد. در این مقاله از مترجم مبتنی بر قاعده فرازین [۳] به‌عنوان مترجم پایه استفاده شده است.

ادامه مقاله به این ترتیب سامان‌دهی شده است: در بخش ۲ به مرور کارهای پیشین و مرتبط با این مقاله می‌پردازیم. در بخش ۳ گرامر درخت - پیوندی را به اختصار معرفی می‌کنیم. بخش ۴ به توضیح جزئیات سیستم پیشنهادی اختصاص داده شده است. در بخش ۵ نتایج به دست آمده از این سیستم را مورد بررسی قرار می‌دهیم و درنهایت در بخش ۶ به نتیجه‌گیری می‌پردازیم.

۲- کارهای پیشین

روش‌های ترکیبی در ترجمه ماشینی مورد توجه بسیاری از محققان بوده‌اند. این روش‌ها معمولاً یک رویکرد را به‌عنوان رویکرد پایه انتخاب کرده و با ترکیب رویکردهای دیگر سعی در بهبود نتایج رویکرد پایه را دارند. در [۴] روش‌های ترکیبی به دو دسته تقسیم شده‌اند؛ روش‌هایی که بر پایه مترجم‌های آماری هستند و سعی دارند نتایج این مترجم‌ها را با تزریق اطلاعات زبان‌شناسی تقویت کنند و دسته دیگر روش‌هایی که بر پایه مترجم‌های مبتنی بر قاعده هستند و با استفاده از داده‌های به دست آمده از روش‌های آماری سعی در بهبود کیفیت مترجم مبتنی بر قاعده دارند.

۲-۱- روش‌های ترکیبی بر پایه مترجم آماری

ترجمه مبتنی بر قاعده می‌تواند در مرحله رمزگشایی و یا پیش و پس پردازش با ترجمه آماری ترکیب شود [۵]. معمولاً هدف سیستم‌هایی که از روش مبتنی بر قاعده در فاز پیش‌پردازش استفاده می‌کنند، نزدیک کردن ساختار جمله ورودی به ساختار زبان مقصد است [۶] [۷]. در این روش‌ها ترتیب کلمات در جمله ورودی باتوجه‌به قواعد نحوی تغییر پیدا می‌کند و جمله تغییریافته توسط مدل مبتنی بر عبارت ترجمه می‌شود. این تغییر در ترتیب کلمات هم در زمان آموزش مترجم و هم در زمان ترجمه انجام می‌شود. یعنی هر جمله سمت مبدأ پیکره موازی آموزش، توسط این قواعد بازآرایی می‌شود و پیکره جدید در فرایند آموزش مورد استفاده قرار می‌گیرد. این تغییرات روی مجموعه داده‌های توسعه و آزمون نیز انجام می‌شود. در برخی روش‌ها به جای قواعد تولید شده توسط خبره، از قواعد تولید شده با روش‌های آماری استفاده می‌شود [۸] [۹].

اطلاعات به‌دست آمده از ترجمه مبتنی بر قاعده ممکن است در فاز رمزگشایی مورد استفاده قرار بگیرد. در [۱۰] از ترجمه‌های به‌دست آمده توسط چند سیستم مبتنی بر قاعده، مجموعه‌ای از عبارات دوزبانه استخراج شده است. این عبارات به جدول عبارات یک مترجم مبتنی بر عبارت اضافه شده‌اند و در فاز رمزگشایی مورد استفاده قرار گرفته‌اند. احسن و همکارانش [۱۱] به روش‌های مختلف از ترجمه مبتنی بر قاعده برای بهبود ترجمه مبتنی بر عبارت استفاده کرده‌اند. از جمله این روش‌ها می‌توان به جابجایی کلمات در زبان مبدأ توسط مترجم مبتنی بر قاعده و همچنین غنی‌سازی جدول عبارات اشاره کرد. در [۱۲] نیز، عبارات دوزبانه منطبق با قواعد و همچنین واژه‌نامه استفاده شده در ترجمه مبتنی بر قاعده را به جدول عبارات روش مبتنی بر عبارت اضافه کرده‌اند.

از قواعد می‌توان در فاز پس‌پردازش نیز استفاده کرد. در [۱۳] روشی ارائه شده است که برخی خطاهای دستور زبان را در خروجی مترجم آماری با استفاده از مجموعه‌ای از قواعد تشخیص می‌دهد و تصحیح می‌کند. این قواعد بر پایه گرامر درخت - پیوندی هستند که ویژگی‌هایی به درختان آنها نسبت داده شده است. با تطبیق این ویژگی‌ها بر خروجی تولید شده، خطاها تشخیص داده می‌شوند.

۲-۲- روش‌های ترکیبی بر پایه مترجم مبتنی بر قاعده

در [۱] مقایسه‌ای بین رویکرد مبتنی بر قاعده و رویکرد آماری ارائه شده است. براساس این تحقیق مترجم‌های مبتنی بر قاعده در رعایت کردن ترتیب صحیح کلمات و تولید صورت صحیح ساخت‌واژی کلمات زبان مقصد مشکلات کمتری نسبت به مترجم‌های آماری دارند. از این‌رو این رویکرد برای ترجمه بین زوج زبان‌هایی که از نظر ساختاری از هم فاصله زیادی دارند مناسب به نظر می‌رسد. همچنین در ترجمه به زبان‌هایی که از نظر ساخت‌واژی غنی هستند، مترجم‌های مبتنی بر قاعده می‌توانند بهتر عمل کنند. به همین دلیل ما در سیستم ترکیبی پیشنهادی خود، یک مترجم مبتنی بر قاعده را به‌عنوان مترجم پایه در نظر گرفتیم و با افزودن اطلاعات آماری در جهت بهبود کیفیت نتایج این مترجم قدم برداشتیم.

مترجم مبتنی بر قاعده نیز می‌تواند در مراحل مختلف ترجمه از ترکیب شدن روش آماری و یا اطلاعات آن بهره ببرد. به عنوان مثال در روش ارائه شده در [۱۴] نتایج تولید شده توسط مترجم مبتنی بر قاعده با استفاده از یک مدل آماری پس‌ویرایش می‌شوند. این مدل آماری با گرفتن ترجمه مترجم مبتنی بر قاعده به‌عنوان متن مبدأ و ترجمه انسانی به‌عنوان متن مقصد، آموزش داده شده است. پس از آموزش مدل، نتایج مترجم مبتنی بر قاعده بر روی مجموعه داده تست به مترجم آماری مبتنی بر عبارت جدید داده می‌شوند تا پس‌ویرایش‌های لازم روی آن‌ها اعمال شود.

برخی از سیستم‌های ترکیبی سعی بر غنی کردن دادگان مترجم‌های مبتنی بر قاعده دارند. در [۱۵] واژه‌نامه مترجم مبتنی بر قاعده با استفاده از واژه‌های استخراج شده از منابع تحت وب تقویت شده است.همچنین در [۱۶] از عبارات استخراج شده به روش آماری از پیکره موازی برای تقویت واژه‌نامه مترجم مبتنی بر قاعده استفاده شده است. ما نیز به نحوی در روش پیشنهادی خود این کار را انجام می‌دهیم، با این تفاوت که ما ترجمه کلمات و عبارات را به همراه بافت نحوی‌شان به مجموعه اضافه می‌کنیم.

۲-۳- گرامرهای مورد استفاده در مدل کردن اطلاعات نحوی

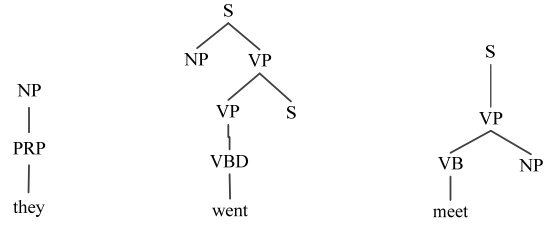
استفاده از اطلاعات نحوی در ترجمه ماشینی را از منظر نوع گرامر مورد استفاده نیز، می‌توان بررسی کرد. برخی مترجم‌ها از ساختارهای وابستگی در ترجمه استفاده می‌کنند [۱۷] در حالی که دیگر مترجم‌ها از گرامرهای مبتنی بر سازه استفاده می‌کنند. قدرت گرامرهای مختلف استفاده شده متفاوت است. یک فرمالیسم ممکن است قادر به مدل کردن پدیده‌هایی در زبان باشد که گرامر دیگر توانایی نمایش آن‌ها را ندارد. در بین گرامرهای مبتنی بر سازه، ابتدا گرامرهای مستقل از متن (CFG) در ترجمه ماشینی مورد استفاده قرار گرفتند [۸]. ناتوانی این گرامرها در مدل کردن برخی جابجایی‌ها باعث شد محققین این حوزه به سراغ گرامرهایی با دامنه محلی بزرگ‌تر[۶] مانند TSG[۷] بروند [۱۸] [۱۹]. در TSG تنها یک عملیات "جایگزینی[۸]" وجود دارد و در هر مرحله از اشتقاق یک غیرپایانه با یک درخت جایگزین می‌شود. در CFG قواعد به‌صورت درخت‌هایی دو سطحی هستند و به همین دلیل تنها می‌توانند جابجایی‌های در سطح همزادها[۹] را مدل کنند ولی در TSG این محدودیت وجود ندارد و قواعد می‌توانند از درختانی با بیش از دو سطح تشکیل شده باشند.

TSG نیز دارای محدودیت‌هایی است و نمی‌تواند ساختارهای حساس به متن[۱۰] را مدل کند. همچنین برخی از قواعد تولید شده توسط مدل مبتنی بر TSG ساختارهای مسطح و بزرگی را مورد پوشش قرار می‌دهند که باعث می‌شوند یک کلمه با تعداد زیادی از وابسته‌هایش در قالب یک قاعده دیده شود. بدین ترتیب با دیدن ترکیب جدیدی از وابسته‌ها در زمان ترجمه، سیستم قاعده‌ای برای پوشش آن‌ها پیدا نمی‌کند. برای رفع این مشکل، و استخراج قواعد عمومی‌تر از دودویی‌سازی درختان استفاده می‌شود. در این روش هر سازه با بیش از دو فرزند به تعدادی زیرسازه شکسته می‌شود و درخت مربوطه به شکل دودویی در می‌آید. این کار باعث شکسته شدن ساختارهای بزرگ و مسطحی که راجع به آن‌ها صحبت شد، به ساختارهای کوچک‌تر می‌شود. در [۲۰] عنوان می‌شود که دودویی‌سازی درختان که از آن به‌عنوان روشی برای کم کردن محدودیت‌های نحوی و رسیدن به قواعدی عمومی‌تر مطرح شد، ممکن است به آزادی بیش‌ازحد در مدل بینجامد که باعث تولید نتایج نادرست از نظر نحوی شود. همچنین در دودویی‌سازی تفاوتی بین وابسته‌های اجباری و اختیاری هسته[۱۱] گذاشته نمی‌شود و به همین دلیل در صورتی که بین یک هسته و یکی از وابسته‌های اجباری آن یک وابسته اختیاری قرار بگیرد، وابسته اجباری با هسته در قالب یک قاعده استخراج نمی‌شوند. برای کم کردن این مشکل و همچنین بالا بردن قدرت مدل در پذیرفتن ساختارهایی که TSG قادر به پذیرفتن آن‌ها نیست، مترجم‌هایی به سمت استفاده از گرامر درخت-پیوندی رفتند که در دسته‌بندی چامسکی جزء زبان‌های حساس به متن ملایم[۱۲] دسته‌بندی می‌شود. در این گرامرها علاوه بر عملیات جایگزینی که در TSG نیز وجود دارد، دارای عملیات الحاق[۱۳] نیز هست. این عملیات قادر است زیر درخت‌هایی را به جای یک گره در یک درخت موجود اضافه کند و برای مدل کردن اتصال سازه‌های اختیاری[۱۴] و ساختارهای بازگشتی استفاده می‌شود. درختان TAG اطلاعات ساختار وابستگی را نیز در خود دارند و به همین دلیل قدرتمندتر از گرامرهای وابستگی هستند. در سیستم پیشنهادی در این مقاله از گرامر درخت-پیوندی در مدل کردن اطلاعات نحوی استفاده شده است. در بخش بعد در مورد گرامر درخت- پیوندی توضیحاتی ارائه خواهد شد.

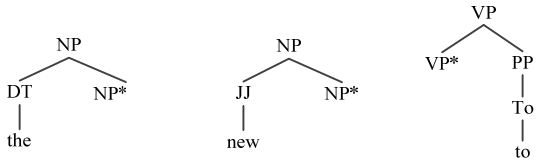
۳- گرامر درخت- پیوندی

گرامر درخت - پیوندی یکی از فرمالیسم‌های قوی در مدل کردن پدیده‌های زبانی است. در این گرامر واحدهای سازنده، درختانی هستند که به آن‌ها درختان اولیه[۱۵] گفته می‌شود. درختان اولیه خود به دو دسته درختان ابتدایی و درختان کمکی تقسیم می‌شوند. درختان ابتدایی ساختارهای نحوی اصلی در جمله را تشکیل می‌دهند و در آن‌ها تمام گره‌های داخلی، غیرپایانه و برگ‌ها پایانه یا غیرپایانه هستند. در گرامر درخت- پیوندی لغوی[۱۶] هر درخت اولیه حتماً دارای یک برگ لغوی است که به آن لنگر[۱۷] گفته می‌شود. یک اشتقاق TAG حتماً با یک درخت ابتدایی شروع می‌شود. نمونه‌هایی از درختان ابتدایی در شکل ۱ نمایش داده شده‌اند.

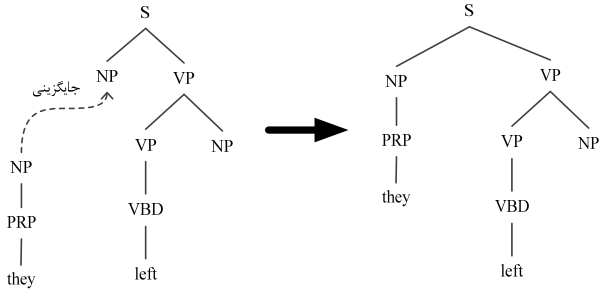
دسته دوم از درختان، که وجه تمایز TAG با TSG به‌حساب می‌آیند، درختان کمکی هستند. این درختان برای وارد کردن سازه‌ها در یک درخت دیگر استفاده می‌شوند و معمولاً نشان‌دهنده سازه‌های اختیاری و ساختارهای بازگشتی هستند. در هر درخت کمکی یک برگ با علامت ستاره (*) مشخص می‌شود که به آن گره انتهایی[۱۸] گفته می‌شود. غیرپایانه مربوط به این گره حتماً با غیرپایانه‌ای که در ریشه درخت قرار دارد یکسان است. چند نمونه از این درختان در شکل ۲ نمایش داده شده‌اند. قاعده سمت چپ نشان می‌دهد که حرف تعریف "the" می‌تواند قبل از یک عبارت اسمی اضافه شود و یک عبارت اسمی جدید تولید کند. قاعده وسط نشان می‌دهد که صفت "new" می‌تواند پیش از یک گروه اسمی اضافه شود و یک گروه اسمی جدید بسازد. و در نهایت قاعده سمت راست نشان دهنده اضافه شدن حرف اضافه"to" به انتهای یک عبارت فعلی و تولید یک عبارت فعلی جدید است.



شکل ۱- نمونه‌هایی از درختان ابتدایی TAG



شکل ۲- نمونه‌هایی از درختان کمکی TAG



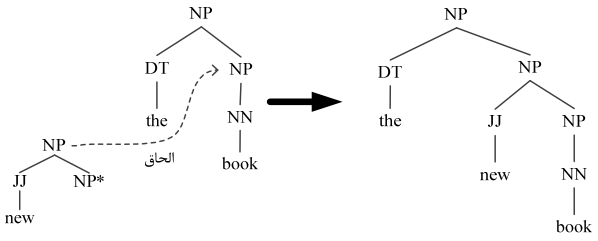
شکل ۳- نمونه‌ای از عملیات جایگزینی

در TAG دو عملیات روی درختان اولیه قابل اعمال است که باعث اتصال درختان اولیه به یکدیگر می‌شوند: عملیات جایگزینی و الحاق. جایگزینی نوعی اتصال محسوب می‌شود و روی درختان ابتدایی قابل اعمال است. در این عملیات یک درخت ابتدایی با ریشه X به یک درخت دیگر با برگ X متصل می‌شود. این

عملیات در TSG نیز وجود دارد و به‌طور مشابه انجام می‌شود. نمونه‌ای از این عملیات که اضافه شدن فاعل به جمله را نشان می‌دهد در شکل ۳ نمایش داده شده است.

الحاق عملیاتی است که روی درختان کمکی انجام می‌شود. در این عملیات یک درخت کمکی در میان یک درخت اولیه اضافه می‌شود. گره‌ای در درخت پدر که درخت کمکی در آن اضافه می‌شود دارای یک "جایگاه الحاق[۱۹]" است. در این عملیات فرزندان گره محل الحاق جدا شده و به زیر گره انتهایی متصل می‌شوند. سپس درخت کمکی تغییریافته در گره محل الحاق جایگزین می‌شود. نمونه‌ای از این عملیات که اضافه شدن یک صفت به یک عبارت اسمی را نشان می‌دهد در شکل ۴ نمایش داده شده است. قدرت گرامر درخت - پیوندی در عملگر الحاق نهفته است. این عملگر به ما این امکان را می‌دهد که یک درخت اولیه را در میان یک درخت اولیه دیگر وارد کنیم و این کار می‌تواند به دفعات انجام شود. در درخت اشتقاق درخت - پیوندی یک جمله، سازه‌های اختیاری معمولاً در قالب عملیات الحاق از هسته خود جدا می‌شوند. این جداسازی باعث می‌شود در استخراج قواعد به قواعد کلی‌تری دست پیدا کنیم که کمتر با خطر تنکی[۲۰] روبه‌رو هستند [۲۱].

درخت تجزیه هر جمله در زبان، از ترکیب درختان ابتدایی و کمکی با عملیات جایگزینی و الحاق ساخته می‌شود و به آن درخت استنتاج[۲۱] گفته می‌شود و از نظر ظاهری تفاوتی با درخت تجزیه CFG ندارد. سابقه درختان اولیه سازنده و عملیات انجام شده روی آن‌ها برای ساخت جمله، در درخت اشتقاق[۲۲] ذخیره می‌شود.



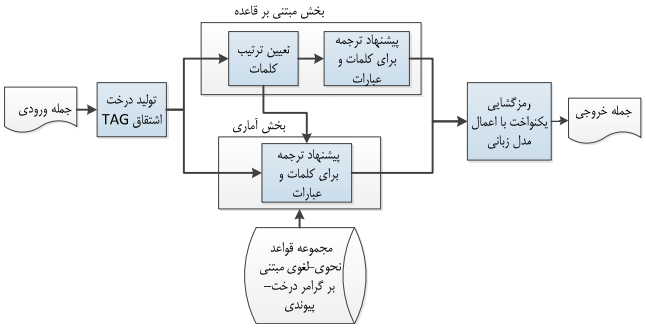
شکل ۴- نمونه‌ای از عملیات الحاق

۴- سیستم پیشنهادی

هدف کلی سیستم پیشنهادی، بهبود ترجمه ارائه شده توسط یک سیستم مبتنی بر قاعده با استفاده از مجموعه از قواعدی نحوی است که به‌صورت آماری از یک پیکره موازی استخراج‌شده‌اند. این قواعد مبتنی بر گرامر درخت - پیوندی هستند و نحوه استخراج آن‌ها در بخش ۴-۱ شرح داده شده است.

نمودار بلوکی نمایش داده شده در شکل ۵ روند کار سیستم ترکیبی پیشنهاد شده در این مقاله را نمایش می‌دهد. مطابق این شکل برای هر جمله ورودی ابتدا درخت اشتقاق در فرمالیسم درخت - پیوندی تولید می‌شود. برای این کار ابتدا جمله توسط تجزیه‌گر استنفورد [۲۲] تجزیه نحوی شده و طی فرایندی مبتنی بر قاعده [۲۳] درخت اشتقاق آن تولید می‌شود. در مرحله بعد مترجم انگلیسی- فارسی فرازین با استناد به این درخت اشتقاق و قواعد داخلی خود، ترتیب صحیح ترجمه کلمات در زبان فارسی را تعیین می‌کند. همچنین برای هر کلمه یا هر دنباله از کلمات که معادلی در واژه‌نامه خود داشته باشد، واژه یا عبارت ترجمه را پیشنهاد می‌دهد. لازم به ذکر است که در روش ارائه شده استفاده از درخت اشتقاق در مترجم مبتنی بر قاعده ضروری نیست و هر مترجم مبتنی بر قاعده دیگری با ساختار دلخواه می‌تواند جایگزین فرازین شود.

بخش آماری سیستم با توجه به ترتیب کلمات ارائه شده توسط فرازین و با استفاده از قواعد خود برای هر کلمه یا هر دنباله متوالی از کلمات، ترجمه‌های پیشنهادی را از مجموعه قواعد آماری استخراج می‌کند. این ترجمه‌های پیشنهادی به همراه ترجمه‌های ارائه شده توسط فرازین در اختیار رمزگشای[۲۳] یکنواخت قرار داده می‌شود تا این رمزگشا با توجه به امتیاز هر ترجمه و با در نظر گرفتن مدل زبانی بهترین ترجمه را از بین ترجمه‌های ممکن انتخاب و به‌عنوان خروجی سیستم اعلام کند. جزئیات مربوط به رمزگشایی در بخش ۴-۲ شرح داده شده است.



شکل ۵- روال کار سیستم پیشنهادی

۴-۱- قواعد نحوی- لغوی

همان‌طور که گفته شد قواعد استفاده شده در این سیستم بر پایه گرامر درخت - پیوندی هستند. روال استخراج قواعد نحوی، که یک پیکره دوزبانه را به‌عنوان ورودی می‌گیرد، شامل مراحل زیر است [۲۰]:

۱. به دست آوردن هم‌ترازی در سطح کلمه بین جملات مبدأ و مقصد پیکره موازی
۲. تجزیه نحوی جملات سمت مبدأ
۳. تبدیل درختان تجزیه به درختان اشتقاق TAG
۴. استخراج قواعد کمینه از درختان اشتقاق
۵. استخراج قواعد مرکب
۶. تخمین احتمالات مربوط به قواعد

برای تولید هم‌ترازی در سطح کلمات از ابزاری مانند GIZA++ [۲۴] استفاده می‌شود. برای تجزیه نحوی جملات سمت مبدأ نیز از یکی از تجزیه‌گرهای موجود مانند تجزیه‌گر استنفورد [۲۲] می‌توان استفاده کرد. پس از تولید درخت تجزیه باید درخت اشتقاق TAG را از روی آن ساخت. برای این کار از روش‌های متفاوتی می‌توان استفاده کرد. روش ارائه شده در [۲۰] و [۲۳] تقریباً مشابه یکدیگر است. هر دو این روش‌ها از اطلاعات زبان‌شناسی برای تشخیص هسته سازه و وابسته‌های مربوط به آن استفاده می‌کنند و با توجه به این اطلاعات درخت اشتقاق TAG را تولید می‌کنند. این در حالی است که الگوریتم ارائه شده در [۲۵] تنها با توجه به ظاهر درخت و بدون استفاده از اطلاعات زبان‌شناسی دست به تولید درخت اشتقاق می‌زند. از این‌رو این روش برای یک درخت تجزیه واحد چند درخت اشتقاق محتمل تولید می‌کند در حالی که الگوریتم ارائه شده توسط چن [۲۳] تنها یک درخت اشتقاق به‌عنوان خروجی تولید می‌کند. به همین دلیل و همچنین به دلیل استفاده از اطلاعات زبانی در روش ارائه شده توسط چن، ما از این الگوریتم استفاده می‌کنیم.

۴-۱-۱- استخراج قواعد کمینه

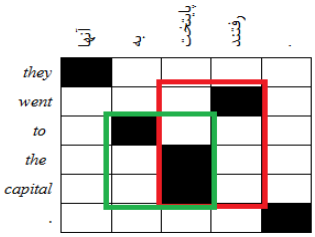
پس از تولید درخت اشتقاق برای جمله ورودی، با توجه به آن درخت و همچنین هم‌ترازی کلمات جمله با کلمات جمله زبان مقصد، قواعد کمینه را استخراج

می‌کنیم. برای این کار تعدادی تعریف را مرور کرده و همزمان روش استخراج قواعد را توضیح می‌دهیم.

تعریف ۱: زوج دنباله‌ای از کلمات در زبان مبدأ و مقصد را سازگار[۲۴] با هم‌ترازی (به‌اختصار "سازگار") گویند که هر کلمه در سمت مبدأ تنها به کلمات داخل دنباله سمت مقصد و یا Null هم‌تراز شده باشد و برعکس [۲۶].

این تعریف اساس استخراج عبارات در مترجم مبتنی بر عبارت است. شکل ۶ هم‌ترازی‌های بین کلمات دو جمله فارسی و انگلیسی را نمایش می‌دهد. با توجه به این شکل زوج عبارت <به پایتخت، to the capital> سازگار است. ولی زوج عبارت <پایتخت رفتند، went to the capital> به دلیل این‌که کلمه "to" به کلمه "به" که بیرون از عبارت فارسی است، هم‌تراز شده است، غیرسازگار است.

تعریف ۲: درخت اولیه‌ای را "کاندیدای استخراج قاعده[۲۵]" گویند که لنگر آن و لنگرهای تحت پوشش هر یک از درختان اولیه فرزندان آن در درخت تجزیه، تشکیل یک عبارت سازگار با سمت مقصد دهند [۲۰].



شکل ۶- مثال‌هایی از عبارات سازگار و غیرسازگار

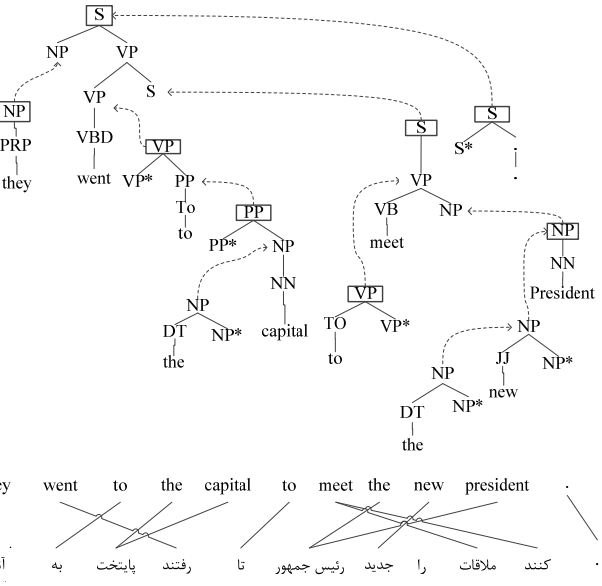
برای استخراج قواعد در مدل نحوی مبتنی بر TAG، درخت اشتقاق به‌صورت پایین به بالا مورد بررسی قرار می‌گیرد و هر درخت اولیه در صورت کاندیدای استخراج بودن، می‌تواند سمت چپ یک قاعده را تشکیل دهد. در غیر این صورت باید آن‌قدر با پدران خود ترکیب شود تا به یک پدر کاندیدای استخراج برسد. لازم به ذکر است که پدر مربوطه دیگر خود به‌تنهائی قادر به تشکیل یک قاعده جدید نخواهد بود. شکل ۷ درخت اشتقاق مربوط به جمله "They went to the capital to meet the new president." را نمایش می‌دهد که درختان اولیه کاندیدای استخراج در آن مشخص شده‌اند.

بدین ترتیب تعریف ۲ را می‌توان به شکل زیر بازنویسی کرد:

تعریف ۳: درخت اولیه و یا ترکیبی از درختان اولیه می‌تواند سمت چپ یک قاعده را تشکیل دهد که در آن درخت اولیه‌ی ریشه، "کاندیدای استخراج" باشد و تمامی فرزندان آن، خود ریشه یک درخت "کاندیدای استخراج" باشند.

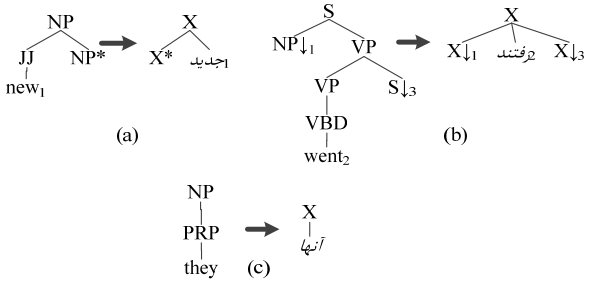
شکل ۸ نمونه‌هایی از قواعد کمینه استخراج شده برای جمله شکل ۷ را نمایش می‌دهد. در سمت راست قواعد تنها متغیر X و پایانه‌ها دیده می‌شوند که تشکیل یک درخت را می‌دهند. این درخت ماهیت نحوی ندارد و از تحلیل نحوی به دست نیامده است و معادل یک رشته در نظر گرفته می‌شود. در این قواعد، جایگاه‌های جایگزینی با ↓ مشخص شده‌اند. اعداد متصل به هر برگ برای مشخص کردن زوج گره‌های همگام در سمت مبدأ و مقصد هستند.

برای به دست آوردن سمت راست قواعد به هم‌ترازی بین عبارت زبان مبدأ و مقصد توجه می‌شود. به ازای هر جایگاه جایگزینی در سمت راست قاعده، یک X در سمت چپ قاعده قرار داده می‌شود تا جایگاه درختی را که در آینده در آن گره جایگزین می‌شود را معین کند. ترتیب قرارگیری این متغیرها به ترتیب قرارگیری لنگرهای وابسته به آن‌ها در زبان مقصد مرتبط است. این ترتیب جابجایی‌های لازم در ترجمه را ایجاد می‌کند و این مسئله قدرت اصلی سیستم‌های ترجمه مبتنی بر نحو است.



شکل ۷- درختان کاندیدای استخراج در جمله "They went to the capital to meet the new president."

در شکل ۸ قاعده اوّل نشان می‌دهد که اگر صفت "new" به یک گروه اسمی بپیوندد، در ترجمه به فارسی جای ترجمه صفت (جدید) با ترجمه گروه اسمی جابجا می‌شود. قاعده دوم ترجمه یک جمله متعدی با فعل "went" را نشان می‌دهد. در انگلیسی فعل بین فاعل و مفعول قرار دارد، در حالی که در ترجمه به فارسی فعل به انتهای جمله و بعد از مفعول انتقال پیدا می‌کند. قاعده سوم یک قاعده لغوی است که هیچ غیرپایانه‌ای در آن حضور ندارد و نشان می‌دهد که ضمیر "they" به "آنها" ترجمه می‌شود.



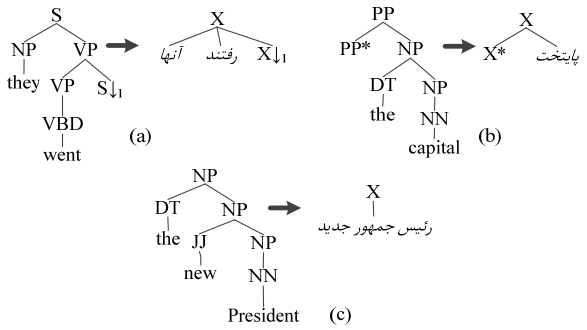
شکل ۸- نمونه‌هایی از قواعد کمینه استخراج شده از جمله مثال

۴-۱-۲- استخراج قواعد مرکب

قواعد کمینه مطابق آنچه در بخش ۴-۱-۱- توضیح داده شد، استخراج می‌شوند. این قواعد کوچک‌ترین قواعد قابل استخراج از داده‌های ورودی و غالباً شامل یک لنگر هستند و به همین دلیل بافت کمی را پوشش می‌دهند. به‌منظور نزدیک شدن به قدرت مترجم‌های مبتنی بر عبارت در ترجمه عبارات چند کلمه‌ای، محققین به سراغ قواعد پیچیده‌تر و بزرگ‌تر رفتند تا بتوانند بافت بزرگ‌تری را مورد پوشش قرار دهند [۲۷]. این قواعد که از ترکیب چندقاعده کمینه به دست می‌آیند قواعد مرکب[۲۶] نامیده می‌شوند. سمت چپ این قواعد ترکیبی از چند درخت اولیه‌ی همبند و سمت راست آنها رشته‌ای در زبان مقصد است که به‌طور مستقل با توجه به هم‌ترازی‌ها استخراج می‌شود. معمولاً برای تولید قواعد مرکب محدودیت‌هایی قائل می‌شوند که اندازه و تعداد قواعد از حدی بزرگ‌تر نشود. این محدودیت می‌تواند بر روی اندازه درخت تولید شده از نظر ارتفاع و پهنا و یا تعداد

غیرپایانه‌های میانی در آن باشد.

بنابراین در مدل پیشنهادی ما نیز از قواعد مرکب استفاده خواهد شد به طوری که هر زیر درخت از درخت اشتقاق که محدودیت‌های تعیین شده را ارضا کند، می‌تواند به‌عنوان سمت چپ یک قاعده مرکب در نظر گرفته شود. در انتخاب این زیردرخت نوع یال بین درختان اولیه و یا همان گره‌های درخت اشتقاق دارای اهمیت نیست. تعدادی از قواعد مرکب استخراج شده از روی این جمله در شکل ۹ نشان داده شده‌اند.



شکل ۹- نمونه‌هایی از قواعد مرکب استخراج شده برای جمله مثال

۴-۱-۳- تخمین احتمالات

پس از استخراج قواعد برای تمام جملات موجود در پیکره آموزش، احتمالات مربوطه به هر قاعده باید محاسبه شوند. برای این کار از روش تخمین بیشینه درست‌نمایی[۲۷] استفاده خواهد شد. برای هر قاعده احتمال P(f|e) محاسبه خواهد شدکه در آن‌ها e درخت نحوی سمت مبدأ قاعده به همراه لنگرهایش و f درخت ساختاری و یا رشته سمت مقصد است. این احتمال را می‌توان با استفاده از رابطه (۱) محاسبه کرد. در این رابطه تعداد رخداد درخت‌های سمت مبدأ و مقصد قاعده به همراه لنگرهای‌شان در کل پیکره آموزشی در صورت کسر قرار می‌گیرد. در مخرج کسر لنگرها در شمارش نادیده گرفته می‌شوند.

$$p(f|e) = \frac{\text{Count}(e,f)}{\sum_{f_i \in \text{Set of all target strings without anchor}} \text{Count}(e,f_i)} \qquad (۱)$$

از آنجایی که قواعد استفاده شده در این سیستم قواعدی نحوی - لغوی هستند، استفاده از آنها به ما کمک می‌کند که ترجمه کلمات را با توجه به ساختار نحوی جمله ورودی انتخاب کنیم. نمونه‌هایی از این قواعد در جدول ۱ نمایش داده شده‌اند.

در جدول ۱ قاعده اوّل نشان می‌دهد که کلمه "break" وقتی به‌عنوان اسم در جمله ظاهر شود به احتمال ۰/۱۴ به "استراحت" ترجمه می‌شود. قاعده دوم نشان می‌دهد عبارت "very different"هنگامی که در نقش صفت برای یک گروه اسمی باشد، در ۲۰٪ موارد به "بسیار متفاوتی" ترجمه شده است. قاعده آخر نشان می‌دهد اگر عبارت "I know" به یک جمله وابسته پیوند داده شود، به احتمال ۰/۱۴ به "من می‌دانم که" ترجمه می‌شود.

۴-۲- رمزگشایی

همان‌طور که گفته شد از آنجایی‌که ترتیب کلمات در زبان مقصد در ابتدای کار توسط مترجم مبتنی بر قاعده تعیین می‌شود، فضای جستجو در زمان رمزگشایی بسیار کوچک‌تر از مترجم‌های آماری است و همچنین می‌توان از برنامه‌ریزی پویا برای پیدا کردن بهترین ترجمه استفاده کرد.

جدول ۱- نمونه‌هایی از قواعد نحوی- لغوی مبتنی بر گرامر درخت- پیوندی

احتمال	ترجمه	درخت سمت مبدأ
۰/۱۴	استراحت	NP NN ***break***
۰/۲	بسیار متفاوتی	NP ADJP NP* RB ADJP ***very*** JJ ***different***
۰/۲۳	شکستن	S VP NP↓ VB ***break***
۰/۳۳	برآورد شده است	S NP↓ VP VP VP VBZ VBN S↓ ***is*** ***estimated***
۰/۱۴	من می‌دانم که	S NP VP PRP VBD SBAR↓ ***I*** ***know***

روش کار به این ترتیب است که یک جدول برنامه‌ریزی پویا با سایز $n \times n$ خواهیم داشت که در آن n تعداد کلمات جمله مبدأ است. و کلمات جمله مبدأ را با ترتیب تعیین شده توسط مترجم مبتنی بر قاعده در این جدول قرار می‌دهیم. بدین‌ترتیب خانه [i, j] در این جدول در بردارنده ترجمه‌های پیشنهادی برای کلمه iام تا jام (در ترتیب جدید) خواهد بود. و در نهایت ترجمه جمله در خانه [1, n] تولید خواهد شد. نمونه‌ای از این جدول برای جمله "I finally went to school yesterday." پس از جابجایی کلمات توسط فرازین در شکل ۱۰ نمایش داده شده است. همان‌طور که توضیح داده شد در این مرحله کلمات هنوز ترجمه

نشده‌اند و تنها ترتیب قرارگیری آنها مطابق ترتیب در زبان فارسی شده است.

برای هر خانه [i, j]، m بهترین ترجمه آن نگه داشته می‌شود و بقیه ترجمه‌ها نادیده گرفته می‌شوند. ترجمه‌های کاندیدا برای خانه [i, j] از دو طریق ایجاد می‌شوند: ۱- مستقیماً توسط فرازین یا مجموعه قواعد ما ارائه می‌شوند. ۲- از ترکیب ترجمه عبارات تشکیل دهنده آن تولید می‌شوند.

برای هر دنباله i تا j از کلمات جمله مبدأ، ترجمه p ممکن است مستقیماً توسط فرازین و یا قواعد ما ارائه شود. امتیاز اختصاص داده شده به این ترجمه که با score(p) نمایش داده می‌شود مطابق رابطه (۲) محاسبه می‌شود. در این رابطه $lm(p)$ امتیاز مدل زبانی عبارت p به شرط کلمات قبل از آن است و $tm(p|w_{i-j})$ احتمال ترجمه کلمات i تا j به عبارت فارسی p است. بسته به این‌که P توسط کدام مرجع ترجمه ارائه شده است، $tm(p|w_{i-j})$ از فرازین یا مجموعه قواعد همگام نحوی گرفته می‌شود. w_1 و w_2 وزن‌هایی هستند که برای امتیاز مدل زبانی و امتیاز ترجمه در نظر گرفته می‌شوند.

$$Score(p) = w_1 \log(lm(p)) + w_2 \log\left(tm(p|w_{i-j})\right) \quad (۲)$$

در روش دوم برای پر کردن خانه [i, j] می‌توان از ترجمه‌های خانه‌های [i, k] و [k+1, j] استفاده کرد. بدین‌ترتیب که ترجمه‌های خانه [i, k] و خانه [k+1, j] را دوبه‌دو باهم در نظر می‌گیریم و به هم متصل می‌کنیم. بدین‌ترتیب برای هر k که بین i و j انتخاب شود حداکثر m^2 ترجمه تولید شده و جزء کاندیداهای خانه [i, j] محسوب می‌شود. امتیاز اختصاص داده شده به هر کاندیدا که از ترکیب ترجمه p از خانه [i, k] و ترجمه q از خانه [k+1, j] تولید شده است مطابق رابطه (۳) محاسبه می‌شود. Score(p) از خانه [i, k] و Score(q) از خانه [k+1, j] خوانده می‌شود. از آنجایی که امتیاز مدل زبانی نیز در محاسبه امتیاز هر خانه دخالت داده شده است لازم است که احتمال n-gram کلمات اولیه q به نسبت کلمات انتهایی p نیز در محاسبه امتیاز Score(pq) دخالت داده شوند.

$$Score(pq) = Score(p) + Score(q) + w_1 \log\left(lm(\text{کلمات مرزی } q_9 p)\right) \quad (۳)$$

در جدول برنامه‌ریزی پویا، قطر اصلی جدول شامل کلمات تکی است. این خانه‌ها با ترجمه‌های فرازین برای هر کلمه و همچنین با استفاده از ترجمه‌های ارائه شده توسط قواعد نحوی-لغوی ما پر می‌شوند. سپس خانه‌های دیگر جدول با ترتیب نمایش داده شده در شکل ۱۰ و با توجه به توضیحات ارائه شده، پر می‌شوند.

یکی از چالش‌هایی که در استفاده از این روش با آن روبه‌رو هستیم این است که خروجی‌های تولید شده توسط سیستم فرازین احتمالاتی نیستند و تنها ترجمه کلمات به‌عنوان خروجی داده می‌شود، یعنی مقداری برای $tm(p|w_{i-j})$ مستقیماً در اختیار ما قرار داده نمی‌شود. راه‌های متعددی برای برطرف کردن این مشکل وجود دارد که ما ساده‌ترین آن‌ها را انتخاب کرده‌ایم و به تمام ترجمه‌های ارائه شده توسط مترجم فرازین یک احتمال ثابت اختصاص دادیم. این احتمال با توجه به مجموعه داده توسعه انتخاب شده است.

۵- آزمایش‌ها و ارزیابی

برای انجام آزمایش‌ها نیاز به یک پیکره دوزبانه برای استخراج قواعد نحوی - لغوی داریم. آزمایش‌ها روی ترجمه انگلیسی- فارسی انجام شده است و ما از پیکره دوزبانه AFEC [۲۸] به‌عنوان پیکره آموزشی خود استفاده کردیم که اطلاعات مربوط به آن در جدول ۲ ارائه شده است.

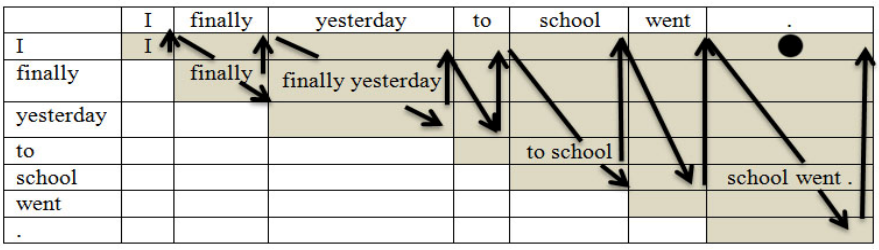
با استفاده از پیکره آموزشی و روش ارائه شده در بخش ۴-۱ مجموعه‌ای از قواعد نحوی- لغوی را استخراج کردیم. همان‌طور که گفته شد هر مجموعه از درختان اولیه که تشکیل یک زیردرخت متصل را بدهند و شرایط گفته شده در بخش ۴-۱ را داشته باشند می‌توانند یک قاعده نحوی را تشکیل دهد. ولی از آنجایی که به‌طور متوسط هر چه اندازه درخت بزرگ‌تر شود، تعداد رخداد آن در پیکره کمتر می‌شود، ما اندازه قواعد خود را محدود کردیم. در آزمایش‌های انجام شده از قواعدی استفاده شده است که سمت راست آن‌ها حداکثر از ترکیب سه درخت اولیه ایجاد شده باشد. همچنین برای بالا بردن کیفیت، قواعدی که تنها یک بار در کل پیکره آموزش دیده شده بودند را حذف کردیم. جدول ۳ اطلاعاتی در مورد مجموعه قواعد استخراج شده ارائه می‌دهد. نمونه‌های از مجموعه قواعد استخراج شده در جدول ۱ نمایش داده شده بود.

جدول ۲- اطلاعات مربوط به پیکره آموزش و مجموعه داده‌های تست و توسعه

		پیکره آموزشی	مجموعه توسعه	مجموعه تست
	تعداد جملات	۶۸۳ هزار	۴۶۰	۴۲۷
انگلیسی	تعداد کلمات	۱۴/۵ میلیون	۱۱،۶۰۰	۱۰،۸۰۰
	تعداد کلمات یکتا	۱۵۸ هزار	۳۰۰۰	۲،۸۰۰
	متوسط طول جمله	۲۱	۲۵	۲۵
فارسی	تعداد کلمات	۱۵/۴ میلیون	-	-
	تعداد کلمات یکتا	۲۰۲ هزار	-	-
	متوسط طول جمله	۲۲	-	-

جدول ۳- آمار مربوط به قواعد استخراج شده

تعداد قواعد کمینه (با ۱ لنگر)	۳۰۶ هزار
تعداد قواعد مرکب (با ۲-۳ لنگر)	۴۰۵ هزار
متوسط تعداد ترجمه‌های مختلف برای هر کلمه و بافت نحوی آن در قواعد کمینه	۳
متوسط تعداد ترجمه‌های مختلف برای هر کلمه و بافت نحوی آن در قواعد مرکب	۲/۶



شکل ۱۰- نمونه‌ای از جدول برنامه‌ریزی پویا

پس از اتمام مرحله استخراج قواعد می‌توانیم روش خود را روی مجموعه داده تست ارزیابی کنیم. اطلاعات مربوط به مجموعه داده‌های توسعه و تست در جدول ۲ ارائه شده است. این مجموعه‌ها از دامنه خبر هستند و در آن‌ها هر جمله انگلیسی چهار ترجمه مرجع دارد.

در روش ارائه شده چند پارامتر وجود دارد که مقادیر آن‌ها باید با توجه به مجموعه داده توسعه، تنظیم شود:

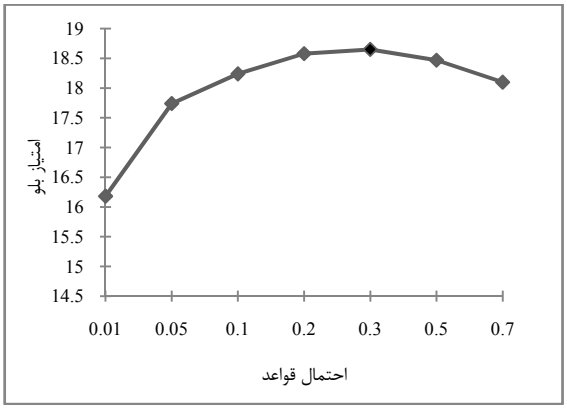
۱- احتمال ثابت اختصاص داده شده به ترجمه‌های فرازین

۲- وزن امتیاز مدل زبانی (w_1)

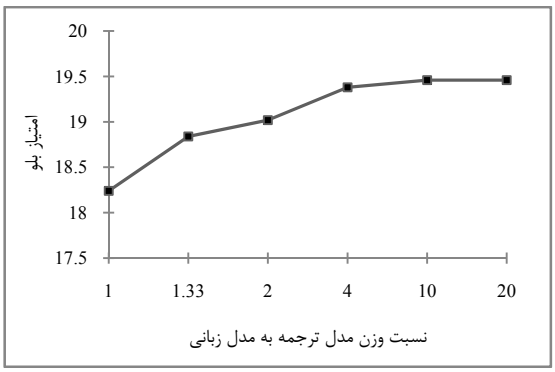
۳- وزن امتیاز ترجمه (w_2)

برای تنظیم مقادیر این پارامترها، در فرایندی تکراری مقادیر این پارامترها را تغییر دادیم و مجموعه داده توسعه را با توجه به آن مقادیر ترجمه کردیم و کیفیت ترجمه را با توجه به ترجمه‌های مرجع سنجیدیم. درنهایت مقادیری که بهترین نتیجه را روی مجموعه داده‌های توسعه داشتند به‌عنوان مقادیر نهایی انتخاب و در ارزیابی سیستم از آن‌ها استفاده کردیم. در ارزیابی‌ها از معیار بلو[۲۸] [۲۹] به‌عنوان معیار ارزیابی کیفیت ترجمه استفاده شده است. همچنین در آزمایش‌ها از مدل زبانی ۴-گرام استفاده شده است و در هر خانه از جدول برنامه‌ریزی پویا ۲۰ کاندیدای ترجمه بهتر را نگه داشتیم.

شکل ۱۱، امتیاز بلوی به‌دست آمده با تغییر احتمال اختصاص داده شده به ترجمه‌های فرازین نمایش داده شده است. در این آزمایش بقیه پارامترها ثابت نگه داشته شده‌اند. با توجه به این شکل اختصاص داده احتمال ۰/۳ به ترجمه‌های ارائه شده توسط فرازین، سیستم را به بهترین نتایج رسانده است.



شکل ۱۱- نمودار امتیاز بلو بر حسب احتمال ثابت نسبت داده شده به ترجمه‌های مترجم مبتنی بر قاعده



شکل ۱۲- نمودار بلو برحسب نسبت وزن مدل ترجمه به وزن مدل زبانی

همان‌طور که توضیح داده شد امتیاز نهایی هر ترجمه از ترکیب خطی لگاریتم امتیاز مدل ترجمه و لگاریتم امتیاز مدل زبانی محاسبه شده برای آن تعیین می‌شود. هریک از مدل‌های ترجمه و زبانی در این ترکیب خطی دارای وزنی است. شکل ۱۲ تأثیر تغییر این وزن‌ها روی کیفیت ترجمه نهایی برای مجموعه داده توسعه بررسی شده است. با توجه به این شکل سیستم پیشنهادی وقتی بهترین عملکرد را داشته است که وزن مدل ترجمه ۱۰ برابر وزن مدل زبانی در نظر گرفته شده است.

جدول ۴ امتیاز های بلوی دو سیستم مورد آزمایش را روی مجموعه داده‌های توسعه و آزمون را پس از پیدا کردن و اعمال وزن‌های بهینه نمایش می‌دهد. همان‌طور که در این جدول نمایش داده شده است، ترجمه‌های ارائه شده توسط سیستم پیشنهادی در این مقاله برای مجموعه داده تست، نسبت به ترجمه‌های مترجم فرازین حدود ۱/۳+ واحد بلو افزایش کیفیت داشته‌اند. برای بررسی معناداری نتایج به‌دست‌آمده از آزمون معناداری ارائه شده در [۳۰] ($P \leq 0.01$; 1000 iterations) استفاده شده است و اختلاف کیفیت‌های به‌دست آمده توسط روش پیشنهادی از نظر آماری معنادار هستند.

در جدول ۵ نیز دو نمونه از جملات مجموعه داده تست به همراه ترجمه‌های ارائه شده توسط فرازین و سیستم پیشنهادی، ارائه شده است. با توجه به این مثال‌ها می‌توان دید که انتخاب لغات در سیستم پیشنهادی بهتر از سیستم پایه است. اختلاف‌ها در این جدول با رنگ متفاوت مشخص شده‌اند.

جدول ۴- نتایج به دست آمده در آزمایش‌ها با استفاده از معیار بلو

مجموعه داده	فرازین	سیستم پیشنهادی
توسعه	۱۷/۸۹	۱۹/۵۹ (+۱/۷)
آزمون	۱۹/۶۵	۲۰/۹۳ (+۱/۲۸)

جدول ۵- نمونه‌هایی از نتایج ترجمه

جمله انگلیسی	mcguinness is expected to return to his role as northern ireland 's deputy first minister.
ترجمه فرازین	مکگاینس **احتمال داده می‌شوند** به نقش او به‌عنوان **جانشین اولین وزیر** ایرلند شمالی بر بگردند.
ترجمه سیستم پیشنهادی	مکگینس**انتظار می‌رود** به نقش او به‌عنوان **معاون نخست وزیر** ایرلند شمالی بر بگردند.
جمله انگلیسی	four afghans , including two students , were also killed , said hashmatstanikzai , spokesman for kabul 's police chief .
ترجمه فرازین	حشمت stanikzai، سخنگو برای رییس پلیس کابل گفت چهار افغان‌ها، از جمله **دو تن از دانشجویانی**، **همچنین** کشته شدند.
ترجمه سیستم پیشنهادی	حشمت stanikzai، سخنگوی رییس پلیس کابل گفت چهار افغان‌ها، از جمله **دو دانشجو**، **نیز** کشته شدند.

۶- نتیجه‌گیری و کارهای آتی

در این مقاله روشی برای غنی‌سازی ترجمه مبتنی بر قاعده پیشنهاد شد. این روش بر پایه مجموعه‌ای از قواعد نحوی- لغوی مبتنی بر گرامر درخت- پیوندی است که

با استفاده از روش‌های آماری از یک پیکره موازی استخراج شده‌اند. این قواعد احتمالاتی برای هر کلمه یا عبارت با توجه به بافت نحوی کلمه، ترجمه‌ای پیشنهاد می‌کنند. ترتیب قرارگیری کلمات در زبان مقصد در ابتدای کار توسط مترجم مبتنی بر قاعده تعیین می‌شود و ادامه کار با ثابت در نظر گرفتن این ترتیب دنبال می‌شود. این امر باعث می‌شود که بتوان عمل رمزگشایی را با استفاده از برنامه‌ریزی پویا انجام داد. نتایج به‌دست آمده از این روش به نسبت مترجم مبتنی بر قاعده پایه از کیفیت بالاتری برخوردار هستند و بهبود ۱/۳+ در واحد بلو در آزمایش‌ها ملاحظه شد. روش پیشنهادی مستقل از زبان است و در صورت وجود پیکره موازی و تجزیه‌گر نحوی مناسب قابل استفاده برای زوج زبان‌های دیگر نیز هست.

در روش ارائه شده، به ترجمه‌های پیشنهادی توسط مترجم مبتنی بر قاعده احتمال یکسانی داده شد. انتخاب روشی مناسب‌تر برای تعیین احتمال برای این ترجمه‌ها می‌تواند باعث بهبود نتایج شود. یکی از راهکارها در این زمینه می‌تواند استفاده از جدول عبارات مدل مبتنی بر عبارت برای اختصاص دادن احتمال به ترجمه‌های مترجم مبتنی بر قاعده باشد. مشکل دیگری که وجود دارد تنکی مجموعه قواعد استخراج شده است. برای کم کردن تأثیر این مشکل می‌توان در مواردی که قاعده‌ای برای کلمه مبدأ در بافت نحوی موردنظر پیدا نمی‌شود به مدلی بدون در نظر گرفتن بافت نحوی عقب‌گرد[۲۹] کرد.

قدردانی

پژوهشی که نتایج آن در این مقاله ارائه شده است در قالب یک طرح تحقیقاتی مصوب و با حمایت مالی صندوق حمایت از پژوهشگران و فناوران کشور انجام شده است.

مراجع

[1] M. R. Costa-jussàa, M. Farr´us, J. B. Mari˜no, and J. A.R. Fonollosa, "*Study And Comparison of Rule-Based And Statistical Catalan-Spanish Machine Translation Systems,"* Computing and Informatics, vol. 31, no. 2, pp. 245-270, 2012.

[2] A. K. Joshi, L. S. Levy, and M. Takahashi, *"Tree Adjunct Grammars,"* Journal of Computer and System Sciences, vol. 10, no. 1, pp. 136–163, 1975.

[3] فرازین: مترجم خودکار متون انگلیسی به فارسی, [Online]. Available: www.faraazin.ir.

[4] M. R. Costa-jussàa, and J. A. R. Fonollosa, *"Latest trends in hybrid machine translation and its applications,"* Computer Speech & Language, vol. 32, no. 1, pp. 3-10, July 2015.

[5] A. Bisazza, and M. Federico, *"A Survey of Word Reordering in Statistical Machine Translation: Computational Models and Language Phenomena,"* Computational Linguistics, vol. 42, no. 2, pp. 163–205, 2016.

[6] M. Collins, P. Koehn, and I. Kucerova, *"Clause Restructuring for Statistical Machine Translation,"* In Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics, pp. 531–540, 2005.

[7] R. N. Patel, R. Gupta, P. B. Pimpale, and M. Sasikumar, *"Reordering rules for English-Hindi SMT,"* In Proceedings of the 2nd Workshop on Hybrid Approaches to Translation (HyTra), pp. 34-41, 2013.

[8] F. Xia, and M. McCord, *"Improving a Statistical MT System with Automatically Learned Rewrite Patterns,"* In Proceedings of the 20th international conference on Computational Linguistics, pp. 508, 2004.

[9] A. Mansouri, H. Fadaei, H. Faili, and M. Arabsorkhi, "*Using Synchronous TAG for Source-Side Reordering in SMT,"* International Journal of Information & Communication Technology Research, vol. 5, no. 4, pp. 47-58, Autumn 2013.

[10] A. Eisele, C. Federmann, H. Saint-Amand, M. Jellinghaus, T. Herrmann, and Y. Chen. "*Using Moses to integrate multiple rule-based machine translation engines into a hybrid system,"* In Proceedings of the 3rd Workshop on Statistical Machine Translation (WMT), pp. 179–182, 2008.

[11] A. Ahsan, P. Kolachina, S. Kolachina, D. Misra Sharma, and R. Sangal, *"Coupling statistical machine translation with rule-based transfer and generation,"* In Proceedings of the 9th Conference of the Association for Machine Translation in the Americas. 2010.

[12] V. M. S´anchez-Cartagena, J. A. P´erez-Ortiz, and F. S´anchez-Mart´ınez, *"Integrating Rules and Dictionaries from Shallow-Transfer Machine Translation into Phrase-Based Statistical Machine Translation,"* Journal of Artificial Intelligence Research, vol. 55, pp. 17-61, 2016.

[13] W. Ma, and K. McKeown, *"Detecting and Correcting Syntactic Errors in Machine Translation Using Feature-based Lexicalized Tree Adjoining Grammars,"* Computational Linguistics and Chinese Language Processing, vol. 17, no. 4, pp. 1-14, December 2012.

[14] A. L. Lagarda, V. Alabau, F. Casacuberta, R. Silva, and E. Diaz-de Liano, *"Statistical post-editing of a rule-based machine translation system,"* In Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers, pp. 217-220, January 2009.

[15] A. Göhring, *"Building a Spanish-German dictionary for hybrid MT,"* The 3rd Workshop on Hybrid Approaches to MachineTranslation (HyTra), pp. 30–35, April 2014.

[16] A. Antonova, and A. Misyurev, *"Improving the precision of automatically constructed human-oriented translation dictionaries,"* In Proceedings of the 3rd Workshop on Hybrid Approaches to Machine Translation (HyTra), pp. 58–66, April 2014.

[17] L. Shen, J. Xu, and R. Weischedel, *"A New String-to-dependency Machine Translation Algorithm with a Target Dependency Language Model,"* In Proceedings of The

Association for Computational Linguistics, pp. 577-585, 2008.

[18] M. Galley, M. Hopkins, K. Knight, and D. Marcu, *"What's in a Translation Rule,"* In Proceedings of The Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL- HLT), Boston, Massachusetts, USA, pp. 273-280, 2004.

[19] L. Huang, K. Knight, and A. Joshi, *"Statistical Syntax-directed Translation with Extended Domain of Locality,"* In Proceedings of AMTA, pp. 66-73, 2006.

[20] S. DeNeefe, *"Tree-adjoining Machine Translation,"* PhD Thesis, Faculty of the USC graduate school University of Southern California, 2011.

[21] S. DeNeefe, K. Knight, W. Wang, and D. Marcu, "*What Can Syntax-based MT Learn from Phrase-based MT?,"* In Proceedings of EMNLP-CoNLL, pp. 755-763, 2007.

[22] D. Klein, and Ch. D. Manning, *"Accurate Unlexicalized Parsing,"* In Proceeding of the 40th Annual meeting of the Association for Computational Linguistics, vol.1, pp. 423-430, 2003.

[23] J. Chen, and K. Vijay-Shanker, *"Automated Extraction of TAGs from the Penn Treebank,"* In Proceedings of the Sixth International Workshop on Parsing Technologies, pp. 73-89, 2000.

[24] F. J. Och, and H. Ney, *"Improved Statistical Alignment Models,"* In Proceedings ofthe 38th Annual Meeting of the Association for Computational Linguistics, pp. 440-447, 2000.

[25] Y. Liu, Q. Liu, and Y. Lu, *"Adjoining Tree-to-String Translation,"* In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics, Portland, Oregon, pp. 1278–1287, 2011.

[26] P. Koehn, *"Statistical Machine Translation,"* Cambridge University Press, 2010.

[27] M. Galley, J. Graehl, K. Knight, D. Marcu, S. DeNeefe, W. Wang and I. Thayer, *"Scalable Inference and Training of Context-Rich Syntactic Translation Models,"* In Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics, pp. 961-968, 2006.

[28] F. Jabbari, S. Bakhshaei, S. M. MohammadzadehZiabary, and S. Khadivi, *"Developing an Open-domain English-Farsi Translation System Using AFEC: Amirkabir Bilingual Farsi-English Corpus,"* In Proceedings of the fourth Workshop on Computational Approaches to Arabic Script-based Languages, pp. 17, 2012.

[29] K. Papineni, S. Roukos, T. Ward, and W. J. Zhu, "*BLEU: a method for automatic evaluation of machine translation,"* In Proceedings of the 40th Annual meeting of the Association for Computational Linguistics, pp.311–318, 2002.

[30] P. Koehn, *"Statistical Significance Tests for Machine Translation Evaluation,"* In Proceedings of EMNLP. pp. 388–395, 2004.

حکیمه فدایی دانشجوی مقطع دکتری رشته مهندسی نرم‌افزار در دانشگاه تهران است. همچنین وی مدرک کارشناسی و کارشناسی‌ارشد خود را در رشته مهندسی نرم‌افزار به ترتیب در سال‌های ۱۳۸۵ و ۱۳۸۸ از دانشگاه شهید بهشتی کسب کرده است. زمینه پژوهشی وی پردازش زبان طبیعی و به‌طور خاص ترجمه ماشینی است.
آدرس پست‌الکترونیکی ایشان عبارت است از:

h.fadaei@ut.ac.ir

هشام فیلی تحصیلات خود را در مقطع کارشناسی مهندسی نرم‌افزار در دانشکدهٔ مهندسی کامپیوتر دانشگاه صنعتی شریف با رتبهٔ یک در سال ۱۳۷۶ به پایان رساند؛ سپس مقاطع کارشناسی‌ارشد نرم‌افزار و دکترای هوش مصنوعی را به ترتیب در سال‌های ۱۳۷۸ و ۱۳۸۵ در همان دانشکده تکمیل کرد. از سال ۱۳۸۷ تاکنون عضو هیأت علمی دانشکدهٔ مهندسی برق و کامپیوتر دانشکدهٔ فنی دانشگاه تهران است. زمینه‌های پژوهشی مورد علاقهٔ ایشان پردازش هوشمند متن و زبان طبیعی، ترجمهٔ ماشینی، داده‌کاوی، بازیابی اطلاعات و شبکه‌های اجتماعی هستند.
آدرس پست‌الکترونیکی ایشان عبارت است از:

hfaili@ut.ac.ir

فرناز قاسمی دوره کارشناسی خود را در سال ۱۳۹۵ در رشته مهندسی فناوری اطلاعات در دانشگاه تهران به پایان رساند. او در حال حاضر دانشجوی کارشناسی‌ارشد دانشگاه تهران در گرایش سامانه‌های شبکه‌ای است. زمینه‌ی پژوهشی مورد علاقه وی پردازش زبان طبیعی است.
آدرس پست‌الکترونیکی ایشان عبارت است از:

f.ghasemi.91@ut.ac.ir

اطلاعات بررسی مقاله:

تاریخ ارسال: ۱۳۹۵/۱۰/۲۷
تاریخ اصلاح: ۱۳۹۵/۱۱/۱۵
تاریخ قبول شدن: ۱۳۹۵/۱۱/۲۵
نویسنده مرتبط: دکتر هشام فیلی، دانشکده مهندسی برق و کامپیوتر، دانشگاه تهران، تهران، ایران.

[1] Rule-Based Machine Translation
[2] Tree Adjoining Grammar
[3] Dynamic Programing
[4] Hybrid Approaches
[5] Monotone
[6] Extended Domain of Locality
[7] Tree Substitution Grammar
[8] Substitution
[9] Siblings

[10]Context Sensitive
[11]Head
[12]Mildly Context Sensitive
[13]Adjunction
[14]Optional
[15]Elementary Tree
[16]Lexicalized Tree Adjoining Grammar
[17]Anchor
[18]Foot Node
[19]Adjunction Site
[20]Sparseness
[21]Derived Tree
[22]Derivation Tree
[23]Decoder
[24]Consistent
[25]Extractable
[26]Composed Rules
[27]Maximum Likelihood Estimation
[28]BLEU
[29]Back-Off

علوم رایانش و فناوری اطلاعات
نشریه علمی انجمن کامپیوتر ایران
مجلد ۱۴، شماره ۲، ۱۳۹۵
صفحات ۲۳- ۲۹
مقاله عادی

نشریه علوم رایانش و فناوری اطلاعات

پیش‌بینی رفتار مشتریان بیمه از طریق ترکیب تکنیک‌های داده کاوی

احسان مختاری سید ابوالقاسم میرروشندل

دانشکده فنی، دانشگاه گیلان، رشت، ایران

چکیده

امروزه مهمترین اقدام شرکت‌های بیمه در بحث بازاریابی و تبلیغات، بخش‌بندی و تفکیک مشتریان براساس رفتار و نیاز آن‌ها است. از این‌رو، این شرکت‌ها برای شناسایی و تحریک کردن مخاطبان خود، بازاریابی و تبلیغات را به‌طور گسترده و هدفمند در تمام محیط‌های ارتباطی به انجام می‌رسانند. برای اثربخشی هرچه بهتر این رویکرد، مشتریان براساس معیارها و اهداف خاصی تفکیک و بخش‌بندی می‌شوند. خوشه‌بندی روشی تحلیلی برای کشف عملکرد و رفتار مخاطبان از طریق اطلاعات آن‌ها است. این امر باعث می‌شود تا شرکت‌ها بتوانند از طریق همین عملکرد مخاطبان، دست به اتخاذ تصمیم و تبلیغات هدفمند نسبت به آن‌ها بزنند. هدف اصلی این پژوهش، ارائه راهکاری برای شناخت و پیش‌بینی عملکرد و رفتار مشتریان جدید در انتخاب نوع بیمه برای حفاظت مسکن خود در برابر مخاطرات، از طریق ترکیب روش K-medoids با شبکه‌های عصبی در جهت تعیین خوشه مشتریان جدید برای ارائه تبلیغ محصولات بیمه‌ای است. در این راستا، بدلیل زیاد بودن مشخصه‌ها در اکثر مجموعه داده‌ها و پراکندگی آن‌ها، ابتدا از طریق تکنیک‌های K-means و K-medoids به کشف الگوهای مفهومی رسیده و با استفاده از همین الگوها بعد از مشخص شدن خوشه مشتریان، فقط با داشتن اطلاعات جمعیت شناختی از سوی مشتریان جدید، خوشه آن‌ها پیش‌بینی و اقدامات لازم صورت می‌گیرد. ویژگی متمایز این پژوهش، ترکیب روش‌های خوشه‌بندی با روش‌های دسته‌بندی در کشف الگو است. آزمایش‌های انجام شده، موفقیت روش پیشنهادی در شناخت و کشف نیازها، همچنین رفتار و عملکرد مشتریان را نشان می‌دهد که براساس آن تبلیغات صورت می‌گیرد.

کلمات کلیدی: بازاریابی و تبلیغات، خوشه‌بندی، شبکه‌های عصبی، K-medoids، K-means.

۱- مقدمه

بسیاری از صاحبنظران عرصه بازار، تقسیم بازار را نوشداروی بازاریابی مدرن دانسته، علت آن را کمبود منابع ذکر می‌کنند [۱]. منطق نهفته در تقسیم بازار، ناهمگنی ترجیح محصولات و رفتار خرید مشتریان بوده، و این تفاوت‌ها معمولاً با اختلافات در محصولات یا مصرف‌کنندگان توضیح داده می‌شود [۲]. افزایش تعداد سازمان‌ها و در نتیجه، تشدید هر چه بیشتر رقابت بین آن‌ها، سازمان‌ها را مجبور کرده است که برای بهبود عملیات تبلیغاتی خود، سرمایه‌گذاری بیشتری انجام دهند. یکی از روش‌هایی که سازمان‌ها برای بهبود عملیات تبلیغاتی خود به کار بسته‌اند، بازاریابی هدفدار[۱] است [۳، ۴]. در بازاریابی هدفدار که در مقابل بازاریابی انبوه[۲] قرار می‌گیرد، سازمان‌ها فعالیت تبلیغاتی خود را (به جای تمامی مشتریان) بر روی گروه‌های خاصی از آن‌ها متمرکز می‌نمایند. بخش‌بندی بازار[۳] به سازمان‌ها این امکان را می‌دهد که بتوانند استراتژی‌های بازاریابی خود را براساس این بخش‌ها تنظیم کنند که تبلیغات یکی از مهمترین ابزار بر روی این بخش‌ها است.

همچنین با برتری رقابتی در این بخش‌ها، بقای خود را در محیط کنونی تضمین کنند [۵]. هدف از خوشه‌بندی[۴] داده‌های مشتریان این است که داده‌ها را به خوشه‌هایی تقسیم کنیم تا داده‌های درون یک خوشه دارای بیشترین شباهت و داده‌های خوشه‌های مختلف دارای کمترین شباهت باشند. مشتریان براساس ویژگی‌های رفتاری[۵]، دموگرافیک[۶]، جغرافیایی[۷] و روانشناختی[۸] در دسته‌های مجزا بخش‌بندی می‌شوند [۵]. فاکتورهای روان شناختی تأثیرگذار بر رفتار خرید مشتری به عنوان مدلی محاسباتی برای قصد خرید مشتری، در بخش‌بندی بازار مورد بررسی قرار گرفته‌اند [۶]. لذا در دنیای امروز استفاده از سیستم‌هایی همچون مدیریت ارتباط با مشتری تنها یک مزیت رقابتی نیست بلکه یک ضرورت برای سازمان محسوب می‌شود. با کندوکاو داده‌های مربوط به مشتریان، به رکوردهای

اطلاعاتی مشتریان ساختار داده می‌شود. جریان تشخیص مشتریان با اهمیت به صورت خودکار صورت می‌گیرد که باعث تغییر در شیوه تشخیص مشتریان خاص و با ارزش از لیست کلیه مشتریان و در نهایت کشف مشتریان وفادار خواهد شد [۶].

۲- پیشینهٔ پژوهش

تخمین درست از رفتار خرید مشتری یکی از مهمترین چالش‌ها در بستر بازاریابی است به گونه‌ای که با ترکیب روش‌های k-means و SOM[۹]، مشتریان به چند بخش مختلف بخش‌بندی شده و با تخصیص مشتریان جدید به یکی از این بخش‌ها با استفاده از روش K-NN[۱۰] با دقت[۱۱] نزدیک به ۹۰ درصد، قصد خرید مشتریان را به درستی تخمین زده است [۶].

همچنین با استفاده از CBR[۱۲] روشی برای بخش‌بندی مشتریان ارائه و از الگوریتم ژنتیک برای بهبود دقت تخصیص یک مشتری جدید به بخش مربوط به خودش استفاده کرده‌اند [۷]. در این روش، ترکیبی از اطلاعات دموگرافیک مشتری و اطلاعات محصول خریداری شده، به عنوان ویژگی‌های توصیف‌کننده یک نمونه استفاده شده است. با استفاده از الگوریتم ژنتیک، نمونه‌ها و ویژگی‌هایی که کمتر معرف ویژگی‌های کلی مجموعه داده هستند، از پایگاه نمونه‌ها حذف شده‌اند. بنابراین الگوریتم با دقت بیشتری عمل بخش‌بندی مشتریان را انجام می‌دهد. با توجه به اطلاعاتی که در خصوص رفتار مشتریان از حیث میزان هزینه کردن بدست می‌آید، مشتریان در سه دسته مشتریان پرخرج، مشتریان کم خرج و مشتریان متوسط قرار می‌گیرند. هنگام ورود یک مشتری جدید، با استفاده از الگوریتم نزدیک‌ترین همسایه و پایگاه حاوی نمونه‌ها، رفتار این مشتری از نظر هزینه‌ای که احتمالا در این شرکت خواهد کرد، تخمین زده می‌شود.

سازمان با شناخت این مشتریان، می‌تواند تمرکز خود را بر روی مشتریان با ارزش (مشتریانی که بیشتر خرج می‌کنند) قرار دهد و یا برای هر دسته، استراتژی بازاریابی متفاوتی را به کار گیرد [۸].

برخلاف بیشتر کارهای صورت گرفته در زمینه بخش‌بندی مشتریان، که از اطلاعات دموگرافیک مشتریان به عنوان متغیرهای بخش‌بندی استفاده می‌کنند، اطلاعات مربوط به اقلام خریداری شده و قیمت هر قلم که به صورت رکوردهای تراکنشی برای هر مشتری در دسترس را به عنوان متغیرهای بخش‌بندی مورد استفاده قرار می‌دهند و با ترکیب الگوریتم ژنتیک و الگوریتم خوشه‌بندی K-means، مشتریان را از لحاظ رفتار خرید، خوشه‌بندی و در نهایت برای تحلیل میزان سودبخشی مشتریان هر خوشه از RFM[۱۳] استفاده می‌کنند. با داشتن میزان سودبخشی هر خوشه، سازمان می‌تواند مشتریان هدف را به سرعت شناسایی کرده و محصولات و خدمات، و منابع مناسبی را به خوشه‌های هدف تخصیص دهد [۹]. در اطلاعات مربوط به بانکداری برای بازاریابی بانکی، نحوه به‌کارگیری روش‌های داده‌کاوی را به منظور انتخاب مجموعه مناسبی از مشتریان بررسی کرده‌اند که یک بانک با در اختیار داشتن اطلاعات مشتریانش، می‌خواهد تعیین کند که در صورت تماس گرفتن با آن‌ها و پیشنهاد افتتاح سپرده مدت‌دار با نرخ سود بالا در آن بانک، چه مشتریانی اقدام به افتتاح حساب خواهند نمود. به عبارت دیگر، بانک برای پیش‌بینی نتیجه تماس با مشتریانش مدلی می‌سازد. این مدل می‌تواند از طریق مدیریت بهتر منابع سازمان (نظیر منابع انسانی، تماس‌های تلفنی، و زمان) کارایی عملیات تبلیغاتی را بهبود ببخشد [۱۰]. همچنین یکی از روش‌های قابل استفاده در تقسیم بازار، استفاده از شبکه‌های عصبی مصنوعی[۱۴] است. علت استفاده از آن، انعطاف‌پذیری در ساختن مدل و توانایی‌اش در استفاده از اطلاعات جدید است [۱۱]. از نمونه‌های مطالعات تقسیم بازار که از شبکه‌های عصبی با نقشه‌های خودسازمانده استفاده کرده‌اند می‌توان تقسیم بازار گردشگران در اتریش و تقسیم گردشگران برتر در استرالیا را نام برد [۱۲]. البته از سایر روشهای شبکه عصبی در پیش‌بینی انتخاب گردشگران، پیش‌بینی رفتار گردشگران و پیش‌بینی تقاضا و تحلیل وفاداری میهمان‌های خارجی استفاده گردیده است [۱۳].

۲-۱- روش شناسی پژوهش

در این پژوهش نیز یک مدل بازاریابی مناسب برای انجام تبلیغات هدفمند بر روی مشتریان بیمه از طریق ترکیب روش K-medoids با شبکه‌های عصبی ارائه می‌شود که برای شرکت بیمه سودآور باشد. مجموعه داده شرکت بیمه سامان مربوط به شهر تهران، توسط ۲۵ نفر از نیروهای بازاریابی این شرکت طی ۵ ماه جمع‌آوری شده است. این مجموعه داده حاوی اطلاعاتی در رابطه با بیمه کردن واحدهای مسکونی و تجاری در برابر حوادث طبیعی و غیرطبیعی بکار رفته است. این مجموعه داده شامل ۳۰ ستون (مشخصه) و ۵۰۰ سطر (رکورد) که از سه بخش اصلی در قسمت مشخصه‌ها تشکیل یافته است. براساس این مشخصه‌ها و ارتباط آنها با یکدیگر خوشه‌بندی انجام گرفته است. این سه بخش شامل:

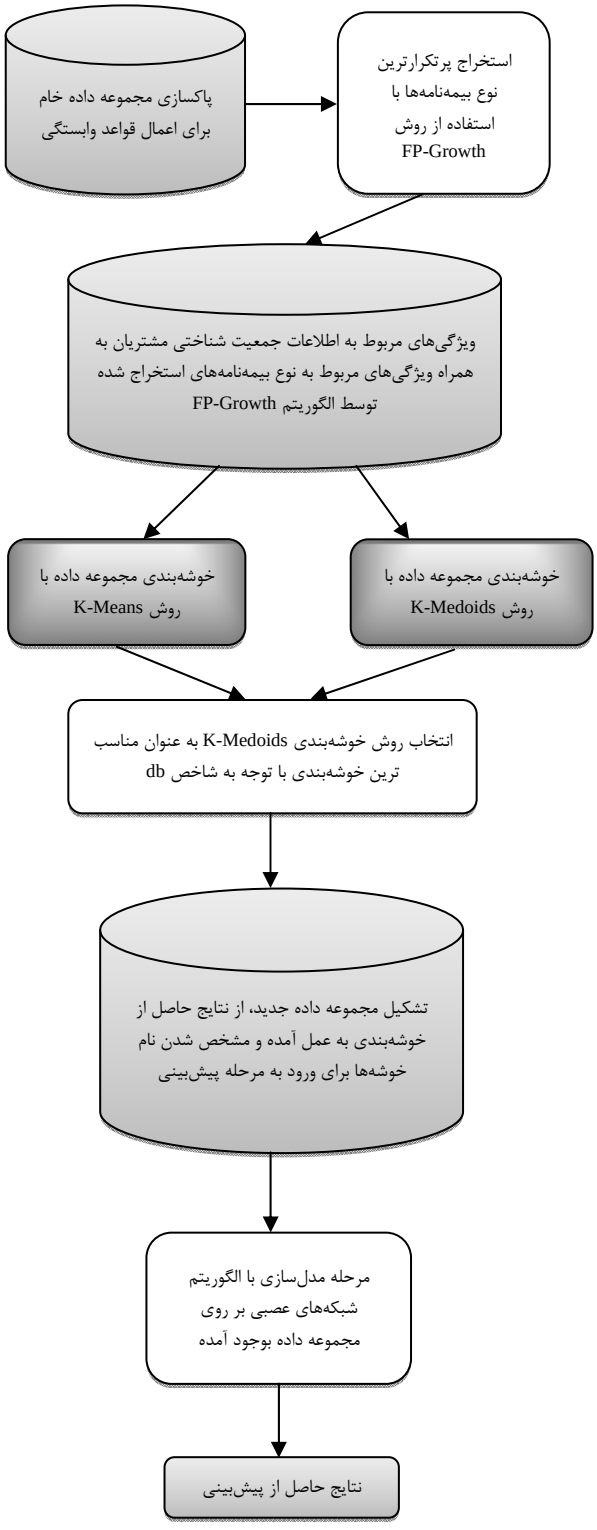
- مشخصه‌های مربوط به اطلاعات جمعیت شناختی مشتریان (سن، میزان در آمد، تحصیلات، نوع اشتغال، تعداد خانوار، میزان استفاده از اینترنت و میزان مطالعه روزنامه)
- مشخصه‌های مربوط به نوع بیمه‌نامه‌های خریداری شده در قبال بیمه مسکن موردنظر (سرقت، ترکیدگی، سیل، زلزله، آتش‌سوزی، رعد و برق)
- مشخصه‌های مربوط به اطلاعات محل شعبات و جایگاه‌های اختصاصی شرکت بیمه سامان در کل مراکز خرید شهر تهران (گلدیس، علاءالدین، اتکا، تیراژه، ایران زمین)

۳- تعیین پرتکرارترین نوع بیمه‌نامه‌ها از طریق روش FP-Growth

برای مشخص کردن پرتکرارترین نوع بیمه‌نامه‌های خریداری شده از روش‌های قواعد وابستگی[۱۵] و تکنیک FP-Growth استفاده شده است. اطلاعات حاصل شده به‌عنوان کشف الگویی در رفتار خرید مشتریان است که همراه با اطلاعات جمعیت شناختی مشتریان وارد مرحله خوشه‌بندی می‌شوند. الگوریتم FP-Growth بر پایه یک ساختار جدید درخت الگوی تکرار شونده است که ساختار توسعه‌یافته‌ای از prefix-tree برای ذخیره اطلاعات فشرده و حساس درباره الگوهای تکرار شونده است. این الگوریتم، روش موثری برای یافتن مجموعه کاملی از الگوهای تکرارشونده با رشد الگو است [۱۴]. سپس از روش رشد الگو استفاده می‌شود تا از تولید پرهزینه تعداد زیاد مجموعه‌های کاندید جلوگیری شود. بعد از خوشه‌بندی، خوشه مشتریان براساس میزان ارزشی که برای شرکت بیمه دارند، نام‌گذاری شده‌اند.

بدین معنی که مشتریانی که بیشترین نوع بیمه را خریده‌اند مشتریان وفادار، و مشتریان دیگر براساس ارزش‌شان نسبت به مشتریان وفادار مشخص شده و خوشه آن‌ها نام‌گذاری شده‌اند. مانند مشتریان فصلی، مشتریان حساس و مشتریان ضعیف. سپس با مشخص شدن وضعیت هر خوشه، همچنین با تعیین نام برای هر خوشه، نام خوشه‌ها این بار خود به عنوان ستون مشخصه هدف در مجموعه داده جدید اضافه شده و برچسب‌گذاری می‌گردند. سپس این مجموعه داده برای مدل‌سازی و پیش‌بینی نام خوشه مشتریان جدیدالورود وارد الگوریتم شبکه‌های عصبی می‌شود. به این صورت که با ورود مشتری جدید و تنها با داشتن اطلاعات مربوط به مشخصه جمعیت شناختی از سوی مشتریان، از طریق شبکه‌های عصبی نام خوشه برای مشتری جدید پیش‌بینی خواهد شد. نهایتاً تبلیغات و نوع بیمه‌های جدید براساس معیارها و متناسب با وضعیت هر خوشه به مشتریان جدید ارائه خواهد

شد. نمودار کلی استخراج ویژگی و مراحل روش پیشنهادی در شکل ۱ نشـان داده شده است.



شکل ۱- نمودار کلی استخراج ویژگی‌ها و مراحل روش پیشنهادی

۳-۱- فرضیه‌ها و مدل مفهومی پژوهش

در راهکار پیشنهادی ترکیبی، پس از پاکسازی داده‌ها برای یـافتن و پـی‌بـردن بـه پرتکرارترین نوع بیمه‌نامه‌های خریداری شده ابتدا از طریق قواعـد وابسـتگی، نـوع بیمه نامه‌هایی که افراد برای خانه‌های خود در برابـر حـوادث و مخـاطرات انتخـاب کرده‌اند مشخص شده است. از سه بخش اصلی در مجموعه داده، تنها مشخصه‌های بخش دوم که مربوط به نوع بیمه‌نامه واحدهای مسکونی و تجاری در برابر حـوادث است و مشتریان آن‌ها را برای خانه‌هـای خـود خریـداری کـرده‌انـد وارد الگـوریتم FP-Growth می‌شود. با این کار آن دسته از بیمه‌نامه‌هـایی کـه کمتـر خریـداری شده‌اند و از دید مشتریان کمتر حائز اهمیت هستند مشخص شده و از بین مـابقی بیمه‌نامه‌ها جدا می‌شوند.

قواعد استخراج شده با درجه پشتیبان و درجه اطمینان بالا و به صورت تجربی انتخاب شده است. که به درجه پشتیبان ۶۰ درصـد و درجـه اطمینـان ۷۰ درصـد حاصل شده‌اند. همچنین مقدار لیفت برای هر کدام از قواعد بیان شده است قواعـد استخراج شده عبارتند از:

Association Rules

[سرقت]--> [ترکیدگی] (confidence:0.787) (lift 100%)
[آتش سوزی]--> [سرقت] (confidence:0.819) (lift 94%)
[زلزله]--> [سرقت] (confidence:0.853) (lift 98%)
[ترکیدگی]--> [سرقت] (confidence:0.904) (lift 100%)

در قواعد استخراج شده چهار قانون استخراج شده است:

قانون اول: کسانی که خانه خود را در برابر سرقت بیمه کرده‌اند، آنگاه با درجه اطمینان ۷۸ درصد خانه خود را در برابر خطر ترکیدگی لوله نیز بیمه کرده‌اند (نوع بیمه ترکیدگی را نیز خریداری کرده‌اند).

قانون دوم: کسانی که خانه خود را در برابر آتش‌سوزی بیمه کرده‌اند، آنگـاه بـا درجه اطمینان ۸۱ درصد خانه خود را در برابر خطر سرقت نیز بیمه کرده‌اند (نـوع بیمه سرقت را نیز خریداری کرده‌اند).

قانون سوم: کسانی که خانه خود را در برابر زلزله بیمه کرده‌اند، آنگاه با درجـه اطمینان ۸۵ درصد خانه خود را در برابر خطر سرقت نیز بیمه کرده‌اند (نـوع بیمـه سرقت را نیز خریداری کرده‌اند).

قانون چهارم: کسانی که خانه خود را در برابر ترکیدگی لولـه بیمـه کـرده‌انـد، آنگاه با درجه اطمینان ۹۰ درصد خانـه خـود را در برابـر خطـر سـرقت نیـز بیمـه کرده‌اند (نوع بیمه سرقت را نیز خریداری کرده‌اند).

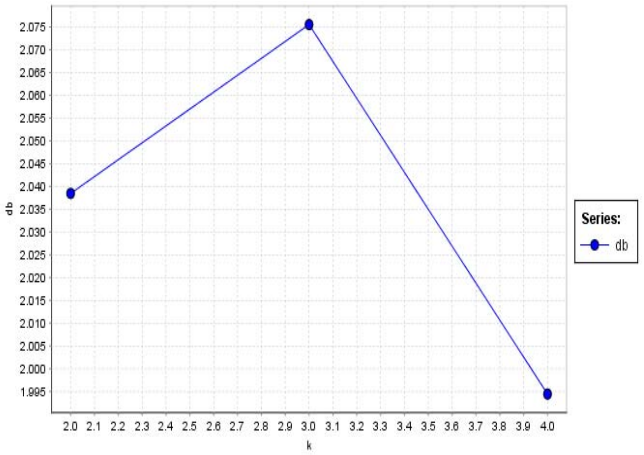
با مشخص شدن پرتکرارترین نوع بیمه‌هـای خریـداری شـده، بـرای آشـکار و تشریح دقیـق‌تـر خوشـه‌هـا نسـبت بـه رفتـار مشـتریان در مرحلـه خوشـه‌بنـدی، مشخصه‌های نوع بیمه‌نامه‌ها با توجه به نتایج حاصل از قواعد وابسـتگی در مرحلـه قبل، همراه با مشخصه‌های جمعیت شناختی وارد مرحله خوشه‌بندی می‌شـوند. در این مرحله، عمل خوشه‌بنـدی بـا اسـتفاده از الگـوریتم K-means و K-medoids انجام می‌شود تا با توجه به معیارهای دقت و صحت خوشه، الگوریتمی که بهتـرین عملکرد و بالاترین دقت را در خوشه‌بندی انجام مـی‌دهـد انتخـاب شـود. براسـاس نتایج حاصل از دقت و صحت خوشه‌بندی‌های صـورت گرفتـه، روش K-medoids انتخاب می‌شود. این انتخاب براساس شاخص دیویس بولدین انجام شده است. ایـن معیار از شباهت بین دو خوشه (Rij) استفاده می‌کند که براساس پراکنـدگی یـک خوشه (si) و عدم‌شباهت بین دو خوشه (dij) تعریف می‌شود. این شاخص در واقع میانگین شباهت بین هر خوشه با شبیه‌تـرین خوشـه بـه آن را محاسـبه مـی‌کنـد. می‌توان دریافت که هرچه مقدار این شاخص بیشتر باشد، خوشه‌های بهتری تولیـد شده است [۱۵]. با مقایسه الگوریتم K-means با K-medoids با توجه به شاخص دیویس بولدین دقت و صحت هر کدام از الگوریتم‌ها مبنی بر تعیین تعـداد خوشـه (K#) برای خوشه‌بندی مجموعه داده موردنظر مشخص شده است. بنابراین با توجه به معیار و شاخص دیویس بولدین الگوریتم K-medoids با عملکـردی دقیـق‌تـر و مناسب‌تر نسبت به الگوریتم K-means، به عنوان الگوریتم خوشه‌بندی با تعیین ۴ خوشه انتخاب شده است. زیرا با توجه به شاخص دیویس بولدین هر چقدر شاخص به عدد ۱ نزدیکتر باشد، خوشه‌بندی با کیفیت‌تر است. جدول ۱ نتایج بررسی‌هـای

حاصل از انتخاب الگوریتم K-medoids را بجای الگوریتم K-means نشان می‌دهد.

جدول ۱- نتایج مقایسه عملکرد الگوریتم K-medoids و K-means از لحاظ تعیین تعداد بهترین خوشه

K-medoids	
# Generated by Log[com.rapidminer.datatable.SimpleDataTable]	
# kdb	
2	2.038455301456895
3	2.0755253687899238
4	1.9144175531970258
5	1.9424645897881654
6	1.9514199199455575

K-means	
# Generated by Log[com.rapidminer.datatable.SimpleDataTable]	
# k	**db**
2	2.533608807984115
3	3.068261052728064
4	3.140990808692302
5	3.189741241276476
6	3.021981090479843



شکل ۲- نمودار شاخص db در الگوریتم K-medoids

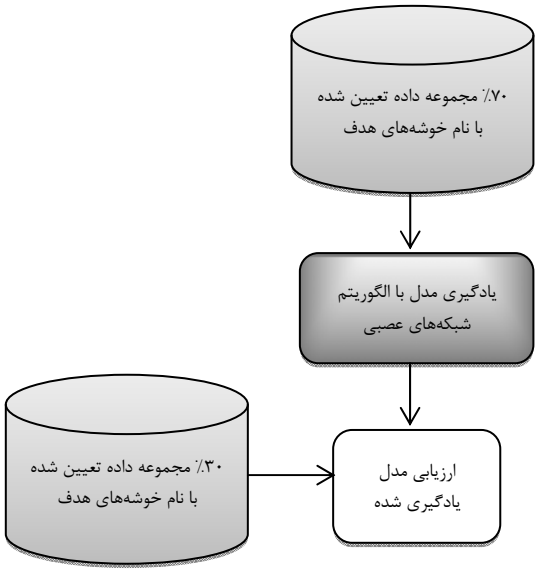
شکل ۲ نمودار تعیین، بهترین تعداد خوشه را با توجه به شاخص دیویس بولدین در الگوریتم K-medoids نشان می‌دهد.

با مشخص شدن روش خوشه‌بندی K-medoids به عنوان مناسب‌ترین روش خوشه‌بندی برای مجموعه داده موردنظر، نتایج خوشه‌بندی به این صورت حاصل شده:

Cluster Model
Cluster 0: 147 items
Cluster 1: 130 items
Cluster 2: 64 items
Cluster 3: 73 items
Total number of items: 414

با توجه به بهترین حالت خوشه‌بندی مطابق با شاخص دیویس بولدین، داده‌ها در بهترین وضعیت در ۴ خوشه، خوشه‌بندی می‌شوند. سپس بعد از بررسی‌های به عمل آمده توسط کارشناسان شرکت بیمه بر روی محتویات هر خوشه، برای هر خوشه نامی تعیین می‌شود تا معرف آن خوشه باشد. به عنوان مثال: مشتریان خوشه شماره ۳ اکثراً سنی بین ۵۰ تا ۶۰ سال دارند، شغلشان مدیر است، درآمد آن‌ها بیش از ۳ میلیون تومان در ماه است، مدرک تحصیلی آن‌ها فوق‌لیسانس است، خانه خود را علاوه بر پرتکرارترین نوع بیمه‌نامه خریداری شده، بر تمامی بیمه‌نامه‌ها نیز بیمه کرده‌اند و معیارهای دیگری که با توجه به اطلاعات مربوط به هر خوشه، کارشناسان این خوشه را بنام خوشه مشتریان طلایی نام‌گذاری کرده‌اند.

همچنین خوشه‌های دیگر نیز با توجه به تجزیه و تحلیل هر خوشه توسط کارشناسان شرکت بیمه بنام مشتریان وفادار، مشتریان حساس و مشتریان ضعیف نام‌گذاری شده‌اند. پس از نام‌گذاری هر خوشه و مشخص شدن محل قرار گرفتن هر رکورد در خوشه‌بندی انجام شده، این بار همین مجموعه داده با اضافه شدن ستون نام خوشه به عنوان مشخصه یا متغیر هدف، برچسب‌گذاری شده و برای پیش‌بینی خوشه برای مشتریان جدید وارد الگوریتم شبکه‌های عصبی می‌شوند. نهایتاً با استفاده از ۷۰ درصد مجموعه داده به عنوان مجموعه داده آموزش جهت یادگیری مدل و انتخاب ۳۰ درصد از مجموعه داده به عنوان مجموعه داده آزمایش، الگوریتم شبکه‌های عصبی بر مجموعه داده حاصل شده اجرا می‌شود. شبکه‌های عصبی تا حد زیادی به عنوان جعبه سیاهی دیده شده‌اند که الگوی پیچیده در داده‌ها را مشخص می‌کنند و یادگیری از طریق آموزش از ویژگی‌های اساسی آن‌ها است [۱۶]. پس از اجرای الگوریتم بر روی مجموعه داده، خوشه‌ها برای مشتریان جدید با دقت ۸۷/۱۰ درصد پیش‌بینی می‌شوند. هر چند شبکه‌های عصبی مصنوعی محدودیت‌های خاص خود را دارند، اما آن‌ها دارای محاسن ویژه‌ای، همچون قدرت یادگیری، انعطاف‌پذیری، انطباق و کشف دانش هستند [۱۷]. نمودار کلی نحوه اعمال مدل پیشنهادی بر روی مجموعه داده در شکل ۳ ارائه شده است.



شکل ۳- نحوه اعمال مدل بر روی مجموعه داده (مرحله یادگیری)

۴- یافته‌ها و ارزیابی پژوهش

در راه کار پیشنهادی، برای بررسی شاخص دقت در نتیجه مدل‌سازی شبکه‌های عصبی از معیار دقت یا نرخ دسته‌بندی استفاده می‌شود:

$$\text{precision} = \frac{\text{TP}}{\text{TP+FP}} \quad (۱)$$

$$\text{recall} = \frac{\text{TP}}{\text{FN+TP}} \quad (۲)$$

TN: این مقدار بیانگر تعداد رکوردهایی است که دسته واقعی آن‌ها منفی بوده و الگوریتم دسته‌بندی نیز دسته آن‌ها را به درستی منفی تشخیص داده است.
FP: این مقدار بیانگر تعداد رکوردهایی است که دسته واقعی آن‌ها منفی بوده و الگوریتم دسته‌بندی، دسته آن‌ها را به اشتباه مثبت تشخیص داده است.
FN: این مقدار بیانگر تعداد رکوردهایی است که دسته واقعی آن‌ها مثبت بوده و الگوریتم دسته‌بندی، دسته آن‌ها را به اشتباه منفی تشخیص داده است.
TP: این مقدار بیانگر تعداد رکوردهایی است که دسته واقعی آن‌ها مثبت بوده و الگوریتم دسته‌بندی نیز دسته آن‌ها را به درستی مثبت تشخیص داده است.

معیار Recall دقت دسته‌بندی دسته x را با توجه به کل رکوردهای با برچسب x نشان می‌دهد. معیار Precision دقت دسته‌بندی دسته x را با توجه به مواردی نشان می‌دهد که برچسب x برای رکورد مورد بررسی توسط دسته‌بند پیشنهاد شده است [۱۸]. جدول ۲ میزان دقت در مدل شبکه‌های عصبی را نشان می‌دهد که برابر است با:

جدول ۲- میزان دقت در مدل شبکه‌های عصبی

	Weighted mean precision	Weighted mean recall	Accuracy
مدل شبکه عصبی	**۹۵.۷۳** درصد	**۹۳.۵۲** درصد	**۹۷.۱۰** درصد

همچنین جدول ۳ نتایج حاصل از میزان دقت در پیش‌بینی خوشه هر رکورد از مجموعه داده آزمایش را نشان می‌دهد.

جدول۳- میزان دقت در پیش‌بینی هر خوشه

	True cluster 1 (مشتریان وفادار)	True cluster 0 (مشتریان ضعیف)	True cluster 2 (مشتریان حساس)	True cluster 3 (مشتریان طلایی)	Class precision
پیش‌بینی خوشه ۱	۳۳	۱	۴	۳	۹۰.۴۹ درصد
پیش‌بینی خوشه ۲	۱	۴۵	۱	۰	۹۷.۷۴ درصد
پیش‌بینی خوشه ۳	۱	۱	۱۴	۱	۹۲.۳۵ درصد
پیش‌بینی خوشه ۴	۱	۰	۲	۱۶	۹۴.۲۱ درصد
Class recall	۹۸.۶۷ درصد	۹۹.۷۴ درصد	۷۶.۶۷ درصد	۹۰.۰۰ درصد	

ارزیابی خوشه‌ای اندازه‌گیری، میزان برتری یک خوشه‌بندی نسبت به خوشه بندی‌های دیگر به وسیله الگوریتم‌های متفاوت خوشه‌بندی یا الگوریتم‌های مشابه ولی با مقدار پارامترهای متفاوت است [۱۹]. با توجه به نتایج و معیارهای حاصل شده، تنها با داشتن برخی اطلاعات جمعیت شناختی از قبیل سن، میزان درآمد و میزان تحصیلات یا میزان استفاده از اینترنت در روز با استفاده از مدل‌سازی شبکه‌های عصبی، خوشه موردنظر مطابق با محتوای اطلاعاتی درون خوشه برای مشتریان جدید پیش‌بینی و براساس سیاست‌ها و استراتژی‌های بازاریابی صورت می‌گیرد.

همچنین تبلیغات موردنظر برای ارائه محصولات و بیمه‌نامه‌های جدید متناسب با وضعیت مشتریان هر خوشه به آنها ارائه می‌شود. همچنین با بررسی پرتکرارترین نوع بیمه‌نامه‌های خریداری شده در هر خوشه، همان بیمه‌نامه‌ها و بیمه‌نامه‌های جدید بدون آگاهی از نیاز و انتخاب مشتری، متناسب با وضعیت خوشه‌های پیش‌بینی شده برای مشتریان ارائه می‌شود.

این کار فقط با داشتن اطلاعات جمعیت شناختی مشتریان انجام می‌شود و بهترین راه تبلیغ محصولات و نوع بیمه‌نامه‌های جدید برای مشتریان و مشتریان جدید است. برای پیاده‌سازی مدل پیشنهاد شده از نرم‌افزار RapidMiner استفاده کرده‌ایم که یک نرم‌افزار مجتمع و مطلوب برای انجام تکنیک‌های خوشه‌بندی و دسته‌بندی است.

۴-۱- تحلیل خوشه‌ها بر روی مجموعه داده‌های هدف

واژه تحلیل خوشه‌ای برای اولین بار توسط ترایون استفاده شد. تحلیل خوشه‌ای، شامل مجموعه‌ای از الگوریتم‌ها و روش‌ها است که جهت گروه‌بندی موضوعات یا اشیای مشابه در طبقه‌های مرتبط استفاده می‌شود [۲۰].

براساس اهداف و معیارهای مختلف، تحلیل‌های خاصی بر روی خوشه‌ها انجام می‌گیرد که در نهایت با توجه به نتایج و خروجی خوشه‌ها، کارشناسان و متخصصان دست به اتخاذ تصمیم می‌زنند و برنامه عملیاتی خود را انجام می‌دهند. با توجه به خوشه‌بندی‌های انجام شده بر روی مجموعه داده‌های این پژوهش می‌توان براساس نیاز، تصمیمات مهمی را در امر بازاریابی و تبلیغات اتخاذ کرد:

- طبق خوشه‌بندی‌هایی که بر پایه اطلاعات جمعیت شناختی حاصل شده بهترین بستر تبلیغات بیمه را تشخیص و آن راه را برای انجام تبلیغ انجام داد و یا از طریق همین خوشه‌بندی از میزان سطح سواد و درآمد افراد آگاهی پیدا کرد و نوع بیمه و شرایط آن را جهت گرفتن بیمه طی تبلیغات هدفمندی به افراد پیشنهاد کرد و موارد گوناگونی که به هدف شرکت و یا سازمان مرتبط است.
- با استفاده از خوشه‌بندی‌هایی که بر پایه اطلاعات نوع و انتخاب بیمه توسط افراد مختلف انجام شده به کمترین و بیشترین انتخاب نوع بیمه دست پیدا کرد و بسته به شرایط افراد و موقعیت‌شان نوع بیمه را به مشتریان قدیمی و یا مشتریان جدید که با محصولات بیمه آشنایی ندارند از طریق تبلیغات طبقه‌بندی شده‌ای پیشنهاد کرد.
- مطابق با خوشه‌بندی‌هایی که بر پایه اطلاعات شعبات مختلف بیمه که در سطح شهر پراکنده شده‌اند به میزان دسترسی و مراجعه افراد به شعبات مختلف می‌توان پی برد و تبلیغات مختص با نقاط پر رفت و آمد و نقاط خلوت انجام داد.

این‌ها تنها مواردی از ارائه تبلیغ بر پایه خوشه‌بندی بود، براساس نیازهای مختلف و تحلیل‌های خاصی که بر روی خوشه‌ها انجام می‌شود هدف‌های راهبردی خاصی انجام می‌شود. همچنین از تکنیک‌های مختلف داده کاوی می‌توان در قسمت دسته‌بندی برای انجام پیش‌بینی‌ها گوناگون استفاده کرد که منجر به نتایج جالب توجه و مفیدی در بحث بازاریابی و ارائه تبلیغات می‌شود.

۵- نتیجه‌گیری و پیشنهادها

به دلیل افزایش رقابت بین سازمان‌ها و فعالیت‌های تبلیغاتی، بخش‌بندی بازار اهمیت ویژه‌ای برای تمرکز روی گروه خاصی از مشتریان و لذا کاهش هزینه‌ها پیدا کرده است. روش پیشنهادی ترکیب روش K-medoids با روش شبکه‌های عصبی در بازاریابی شرکت بیمه برای پیش‌بینی خوشه مشتریان استفاده می‌شود. نتایج آزمایش انجام شده بر رویداده‌های یک شرکت بیمه نشان می‌دهد که تکنیک

پیشنهادی راهکاری جدید با دقت نسبتاً خوب برای پیش‌بینی عملکرد رفتاری مشتریان برای ارائه تصمیمات مختلف است.

همچنین با این تکنیک دقت در پیش‌بینی به مراتب بالاتر از مواقعی است که متغیر هدفی در مجموعه داده نیست. لذا، استفاده از روش K-medoids تنها برای تشریح برخی مسائل یا شبکه‌های عصبی برای پیش‌بینی یک عمل به تنهایی دقت بالایی ندارد و تکنیک‌های اختصاصی‌تری نظیر ترکیب K-medoids با الگوریتم درخت تصمیم و یا تکنیک‌های قواعد وابستگی بسیار مناسب‌تر از الگوریتم‌های عام منظوره هستند. در آینده می‌توان کارایی این روش پیشنهادی را در کاربردهای دیگر بازاریابی مورد ازریابی قرار داد. به علاوه، ترکیب روش پیشنهادی با سایر روش‌های موجود، با هدف افزایش کارایی و یا دقت آن نیز می‌تواند موضوع پژوهشی مناسبی برای کار در آینده باشد. می‌توان شکل‌های مختلف ترکیب تکنیک را نیز به مجموعه قواعد مورد استفاده اضافه کرد و شکل پیچیده‌تری را نیز برای فرضیه‌ها تعریف کرد. بدین صورت فضای فرضیات بزرگتری را مورد جستجو قرار داده و لذا احتمال یافتن فرضیه‌های بهتر با دقت پیش‌بینی بالاتر افزایش می‌یابد.

مراجع

[1] Y. Wind, "Issues and advances in segmentation research," Journal of marketing research, vol. 15, no. 3, pp. 317-337, 1978.

[2] V. Mahajan, and A. Kjain, "An approach to normative segmentation," Journal of Marketing Research, vol. 15, no. 3, pp. 338-345, 1978.

[3] D. Chaffey, and P. R. Smith, "eMarketingeXcellence: Planning and optimizing your digital marketing," Routledge, 2013.

[4] D. Chaffey, and et. al.,"Internet marketing: strategy, implementation and practice," Pearson Education, 2009.

[5] S.Goyat, "The basis of market segmentation: a critical review of literature," 2011.

[6] T. Hong, andE. Kim, "Segmenting customers in online stores based on factors that affect the customer's intention to purchase," Expert Systems with Applications, vol. 39, no. 2, pp. 2127-2131, 2012.

[7] Y. Chen, CH. Yi Wang, and Y. Feng, "Application of a 3NN+ 1 based CBR system to segmentation of the notebook computers market," Expert Systems with Applications, vol.37, no.1, pp. 276-281, 2010.

[8] C-Y. Tsai, and C-C. Chiu, "A purchase-based market segmentation methodology," Expert Systems with Applications, vol. 27, no. 2, pp. 265-276, 2004.

[9] M. Sergio, R. Laureano, and P. Cortez, "Using data mining for bank direct marketing: An application of the crisp-dm methodology," Proceedings of European Simulation and Modelling Conference-ESM, Eurosis, 2011.

[10] W. Kuang-Wei, and K. Peng. "Market segmentation via structured click stream analysis," Industrial Management & Data Systems, vol. 102, no. 9, pp. 493-502, 2002.

[11] J. A. Mazanec, "Classifying tourists into market segments: A neural network approach," Journal of Travel & Tourism Marketing, vol. 1, no.1, pp. 39-60, 1992.

[12] J. Kim, SH. Wei, and H. Ruys, "Segmenting the market of West Australian senior tourists using an artificial neural network," Tourism Management, vol. 24, no. 1, pp. 25-34, 2003.

[13] R. Law, and N. Au, "A neural network model to forecast Japanese demand for travel to Hong Kong," Tourism Management, vol. 20, no. 1, pp. 89-97, 1997.

[14] H. Lijun, and et. al., "Comparison and Analysis of algorithms for association rules," Database Technology and Applications, First International Workshop on, IEEE, 2009.

[15] D. L. Davies, and D. W. Bouldin, "A cluster separation measure," IEEE transactions on pattern analysis and machine intelligence, vol. 1, pp. 224-227, 1979.

[16] S. Walczak, and N. Cerpa, "Heuristic principles for the design of artificial neural networks," Information and software technology, vol. 41, no. 2, pp. 107-117, 1999.

[17] S. Goonatilake, and C. P. Treleaven, Intelligent systems for finance and business, New York, USA: John Wiley & Sons, Inc, 1995.

[18] D. Powers, andD. Martin, "Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation," 2011.

[19] P. K. Roy, andet. al., "Automated Segmentation of White Matter Lesions Using Global Neighbourhood Given Contrast Feature-Based Random Forest and Markov Random Field," Healthcare Informatics (ICHI), IEEE International Conference on, IEEE, 2014.

[20] K. Hornik, M. Stinchcombe, andH. White, "Multilayer feedforward networks are universal approximators," Neural networks, vol. 2, no. 5, pp. 359-366, 1989.

احسان مختاری فارغ‌التحصیل رشته فناوری اطلاعات گرایش تجارت الکترونیکی از دانشگاه گیلان هستند و زمینه‌های مورد علاقه ایشان داده‌کاوی، متن‌کاوی و هوش تجاری هستند.

آدرس پست‌الکترونیکی ایشان عبارت است از:
ehsan_m63000@yahoo.com

سید ابوالقاسم میرروشندل فارغ‌التحصیل از دانشکده فنی دانشگاه تهران در مقطع کارشناسی در رشته مهندسی کامپیوتر با گرایش نرم‌افزار و دانشگاه صنعتی شریف در مقاطع کارشناسی‌ارشد و دکتری در رشته مهندسی کامپیوتر گرایش هوش مصنوعی. از سال ۱۳۹۱ عضو هیات

علمی گروه مهندسی کامپیوتر دانشگاه گیلان بوده و زمینه‌های مورد علاقه ایشـان پردازش زبان‌های طبیعی، داده‌کاوی، یادگیری ماشینی و پردازش تصویر هستند.
آدرس پست‌الکترونیکی ایشان عبارت است از:
mirroshandel@guilan.ac.ir

اطلاعات بررسی مقاله:

تاریخ ارسال: ۱۳۹۵/۱۰/۰۱
تاریخ اصلاح: ۱۳۹۵/۱۱/۰۸
تاریخ قبول شدن: ۱۳۹۵/۱۱/۳۰
نویسنده مرتبط: دکتر سید ابوالقاسم میرروشندل، دانشـکده فنـی، دانشـگاه گیلان، رشت، ایران.

---

[1] Target Marketing
[2] Mas Marketing
[3] Market Segmentation
[4] Clustering
[5] Behavioral
[6] Demographic
[7] Geographic
[8] Psychographic
[9] Self-Organizing Map
[10] K-Nearest Neighbors
[11] Accuracy
[12] Case-Based Reasoning
[13] Recency, Frequency, Monetary
[14] Artificial Neural Networks
[15] Association Rule

علوم رایانش و فناوری اطلاعات
نشریه علمی انجمن کامپیوتر ایران
مجلد ۱۴، شماره ۲، ۱۳۹۵
صفحات ۳۰- ۳۸
مقاله عادی

مدیریت سیستمی دمای پردازنده‌های چندهسته‌ای برای زبان‌های موازی مبتنی بر زمانبند ربایش کار

حمید گوهرجو[۱] مرتضی مرادی[۱و۲] حمید نوری[۱و۲]

[۱]دانشکده مهندسی، دانشگاه فردوسی مشهد، مشهد، ایران
[۲]پژوهشکده علوم کامپیوتر، پژوهشگاه دانش‌های بنیادی (IPM)، تهران، ایران

چکیده

در سال‌های اخیر، دمای بالا و توان مصرفی زیاد در پردازنده‌های چندهسته‌ای به یک چالش اساسی برای سازندگان و کاربران این پردازنده‌ها تبدیل شده است. با رشد دمای پردازنده، هزینه‌های خنک‌سازی و مصرف توان افزایش یافته و طول عمر پردازنده کاهش می‌یابد. در این تحقیق، یک الگوریتم مدیریت دمای پویا در سطح سیستم عاملپیشنهاد شده است که در اجرای برنامه‌های موازی مبتنی بر زمانبند ربایش کار[۱]، دمای پردازنده را در محدودیت درخواستی کاربر مدیریت می‌کند. از این رو، ما دو مدل دمایی و کارایی را جهت پیش‌بینی دمای آینده و تخمین میزان تغییرات کارایی برنامه پیشنهاد دادیم. با استفاده از مدل‌های پیشنهادی، الگوریتم پیشنهادی تعداد هسته‌های فعال و فرکانس پردازنده را به نحوی تعیین می‌کند که دما از محدودیت تعیین شده پایین‌تر نگه داشته شده و کمترین آسیب ممکن به کارایی برنامه وارد گردد. آزمایشات بر روی سیستم واقعی نشان داد که الگوریتم پیشنهادی به طور میانگین ۲۸ درصد کارایی بالاتری از الگوریتم آگاه از همسایگی داشته و برخلاف این الگوریتم، هرگز از محدودیت دمایی تعیین شده تخطی نمی‌کند.

کلمات کلیدی: اجرای موازی، پردازنده چندهسته‌ای، ربایش کار، زمانبند، سیستم عامل، مدیریت دما.

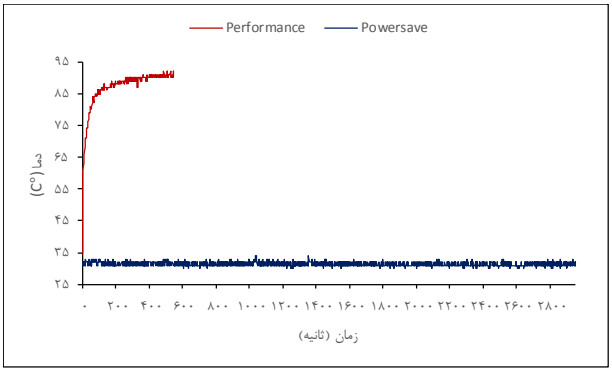
۱- مقدمه

در سال‌های اخیر با افزایش تعداد و چگالی ترانزیستورها بر روی تراشه، توان مصرفی و دمای پردازنده‌های چندهسته‌ای افزایش یافته است. با رشد دمای پردازنده، هزینه‌های خنک‌سازی و مصرف توان افزایش یافته و طول عمر پردازنده کاهش می‌یابد [۱]. در مواجهه با این مشکل، راهکارهای مختلفی با محوریت مدیریت دمای پردازنده مطرح شده‌اند که از این میان، دسته‌ای با عنوان مدیریت پویای پیشگیرانه‌ی دما[۲]، رسیدن پردازنده به آستانه‌ی دمایی نامطلوب را با استفاده از مدل‌هایی پیش‌بینی کرده و با انجام اقدامات پیشگیرانه‌ای مانند کاهش فرکانس ساعت پردازنده[۳]، مهاجرت پردازه‌ها[۴] و کاستن از هسته‌های فعال پردازنده ضمن ممانعت از کاهش شدید کارایی از بروز دمای نامطلوب اجتناب می‌کنند [۳].

با ظهور پردازنده‌های چندهسته‌ای، برنامه‌ها ناگزیرند تا با استفاده از زبان‌های برنامه‌نویسی موازی پیاده‌سازی شوند تا بتوانند کارایی بالایی را در اجرا بر روی پردازنده‌های چندهسته‌ای بدست آورند. تعداد زیادی از زبان‌های موازی مانند OpenMP، Intel Cilk Plus و Intel TBB، از زمانبند ربایش‌کار برای ایجاد موازی‌سازی در سطح برنامه بهره می‌برند [۲]. این در حالی است که سیستم عامل‌ها از وجود چنین زمانبندی درون برنامه اطلاعی ندارند. در اجرا به وسیله زمانبند ربایش کار، چندین نخ به‌عنوان کارگر[۵] در سطح سیستم عامل ایجاد می‌شود که تمامی آن‌ها به صورت همزمان به اجرای وظایف تعریف شده در برنامه می‌پردازند. از آنجایی که تخصیص وظایف به نخ‌های کارگر در زمان اجرا به صورت پویا و بدون پیش‌فرضی در برنامه صورت می‌گیرد، در صورت توقف یکی از نخ‌های کارگر، دیگر نخ‌های فعال در برنامه به اجرای وظایف باقیمانده خواهند پرداخت. این در حالی است که در برخی مدل‌های برنامه‌نویسی دیگر در صورت توقف یک نخ ممکن است ادامه‌ی اجرای کل برنامه به چالش کشیده شود. در [۴] و [۵] به بهبود صف نگهداری وظایف در ربایش کار پرداخته شده است. و [۶]، [۷] و [۸]

مسئله‌ی نحوه‌ی انتخاب صف قربانی را جهت رعایت وظیفه بهبود داده‌اند. در [۹] در هنگام رعایت وظیفه به جای یک وظیفه، نصف وظایف موجود در صف قربانی رعایت می‌شود. در [۱۰] رعایت‌کننده خودش حق برداشتن وظیفه از صف قربانی را ندارد بلکه درخواست خود را به قربانی ارسال می‌کند و قربانی در صورت داشتن وظیفه آن را برای رعایت‌کننده ارسال می‌کند. در [۱۱] با استفاده از ساختمان داده لیست پیوندی، یک صف با اندازه پویا ارائه شده است. در [۱۲] به زمان‌بندی وظایف با درنظر گرفتن اولویت در رعایت کار پرداخته شده است. در هیچ کدام از این کارها مساله دما مورد توجه قرار نگرفته است.

در این تحقیق، ما اجرای موازی برنامه‌ی ضرب استاندارد ماتریس را بر روی پردازنده Intel Core i7-4790K توسط دو مدیریت فرکانس powersave و performance موجود در سیستم عامل لینوکس آزمایش کردیم. نمودار ارائه شده در شکل ۱، دمای پردازنده را در این آزمایش نشان می‌دهد که در آن محور افقی زمان اجرا و محور عمودی دمای پردازنده است. همان‌طور که در شکل دیده می‌شود، دمای پردازنده در مدیریت با powersave به طور میانگین حدود ۵۵ درجه کمتر از میانگین دمای performance است. در مقابل زمان اجرای performance تقریباً یک‌چهارم زمان اجرای powersave است. بر این اساس، دریافتیم که با امکانات موجود در سیستم عامل لینوکس نمی‌توان با صرف هزینه کارایی پایینی، دمای پردازنده را در محدوده‌ی دلخواه مدیریت کرد.



شکل ۱- دمای پردازنده تحت مدیریت فرکانس در سیستم عامل لینوکس

تا کنون تحقیقات زیادی برای مسئله‌ی مدیریت پویای دمای پردازنده‌ی چندهسته‌ای صورت گرفته است. در این کارها تمرکز روی معیارهایی از جمله کاهش میانگین دمای پردازنده، کاهش رخداد نقاط داغ[۶] بر روی پردازنده و نیز کنترل دما زیر محدودیت درخواستی کاربر بوده است [۱۳-۱۸]. از میان کارهای صورت گرفته، [۱۳] و [۱۴] عملکرد بهتری از دیگران داشته‌اند. در این کارها، از مدل‌های پیش‌بینی دمای پردازنده بهره‌برده شده است تا دمای آینده‌ی پردازنده پیش‌بینی شود و در صورت لزوم برای اجتناب از گذر دما از محدودیت درخواستی کاربر، عمل پیش‌دستانه‌ای صورت می‌گیرد. تحقیقات دیگری مشابه این کارها صورت گرفته است اما در اکثر این کارها ویژگی‌هایی از جمله موازی‌سازی در برنامه‌های کاربردی مورد توجه قرار گرفته نشده است و لذا کارایی آن‌ها برای چنین کاربردهایی مورد تردید است.

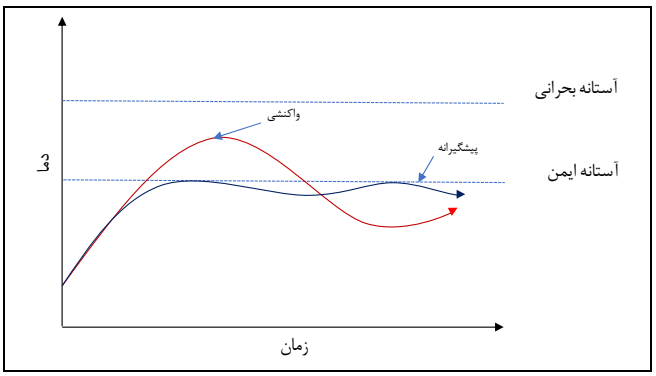
اطلاعات ما نشان می‌دهد که تاکنون هیچ یک از راهکارهای مدیریت دمای مطرح شده در سطح سیستم عامل، آگاهی از وجود زمانبند رعایت کار را در برنامه‌های موازی در نظر نگرفته‌اند. در حالی که با وجود این زمانبند، امکان تغییر تعداد نخ‌های فعال در زمان اجرا به صورت پویا و بدون نیاز به تغییر ساختار برنامه وجود دارد. از این‌رو، در این تحقیق یک الگوریتم مدیریت دمای پویا در سطح سیستم عامل را پیشنهاد دادیم که در اجرای برنامه‌های موازی، دمای پردازنده را کمتر از آستانه‌ی درخواستی کاربر مدیریت می‌کند. در الگوریتم ما، فرض بر این است که برنامه‌ی موازی در حال اجرا با استفاده از زمانبند رعایت کار پیاده‌سازی شده است و در اثر کاهش و یا افزایش تعداد نخ‌ها، روند اجرای برنامه مختل نخواهد شد. در ادامه، نوآوری‌های این پژوهش به صورت خلاصه بیان شده است:

- ارائه الگوریتم مدیریت دما در سطح سیستم عامل برای برنامه‌های موازی مبتنی بر زمانبند رعایت کار برای اولین بار به نحوی که تضمین‌کننده‌ی عدم‌تجاوز دما از سطح درخواستی کاربر باشد.
- پیشنهاد دو مدل دمایی و کارایی با دقت مناسب جهت استفاده در الگوریتم‌های مدیریت دما.
- بررسی دقت مدل دمایی و تاثیر افزایش دقت مدل‌ها در الگوریتم مدیریت دما، بر کارایی برنامه و دمای پردازنده.
- عدم‌تجاوز از محدودیت دمایی کاربر با صرف هزینه‌ی کارایی پایین‌تر از دیگران، چنانچه در کارهای دیگران با وجود تخطی از محدودیت دمایی تعیین شده، کارایی پایین‌تری بدست آمده است.

در ادامه، بخش ۲ پیشینه‌ی تحقیق و کارهای انجام شده در زمینه‌ی مدیریت دما را بررسی کرده است. سپس، مدل‌های دمایی، مدل کارایی و الگوریتم مدیریت دمای پیشنهادی در بخش ۳ ارائه شده‌اند. همچنین، بخش ۴ به ارزیابی الگوریتم مدیریت دمای پیشنهادی و بررسی دقت مدل‌های دمایی پرداخته است. در نهایت، نتیجه‌گیری و پیشنهاداتی برای ادامه‌ی تحقیق در بخش ۵ بیان شده است.

۲- پیشینه

تا کنون در تحقیقات صورت گرفته در مدیریت پویای دمای پردازنده، از ابزار کنترلی دما مانند تنظیم فرکانس پردازنده، مهاجرت پردازه‌ها و تغییر درجه همروندی اجرای پردازه‌ها در یک زمانبند زمان اجرا برای دستیابی به اهداف استفاده شده است [۱]. به‌طور کلی، رویکرد حل مسئله در این تحقیقات را می‌توان در دو دسته‌ی واکنشی[۷] و پیشگیرانه[۸] تقسیم کرد (شکل ۲). در رویکرد واکنشی، پس از رسیدن دمای پردازنده به آستانه‌ی نامطلوب از ابزار کنترلی برای کاهش دما استفاده می‌شود. مشکل اساسی این روش این است که عدم‌تخطی از آستانه را تضمین نمی‌کند. علاوه بر این، نوسانات زیاد دمایی ایجاد شده در این روش، موجب کاهش عمرپردازنده می‌شود [۱۵]. در مقابل، رویکرد پیشگیرانه با بهره‌گیری از مدل‌های پیش‌بینی دما و انجام اقدامات کنترلی پیشگیرانه از رسیدن دما به محدوده‌ی نامطلوب جلوگیری می‌کند. مطالعات اخیر نشان داده است که رویکرد پیشگیرانه کارآمدی بیشتری در دستیابی به اهداف مدیریت دما از جمله عدم‌تخطی از آستانه‌ی دمای نامطلوب و اجتناب از کاهش کارایی برنامه دارد [۳].



شکل ۲- رویکردهای مدیریت دما

به طور کلی کارهای انجام شده در رویکرد پیشگیرانه‌ی دما را می‌توان در دو

دسته قرار داد. دسته‌ی اول، کارهایی هستند که در آن‌ها آگاهی از وجود برنامه‌های موازی مورد توجه قرار گرفته نشده است و مدیریت دما، هر یک از نخ‌های تولید شده توسط یک برنامه موازی را به عنوان برنامه‌ی مستقلی تلقی می‌کند [۱۳-۱۷]. براساس آزمایشات ما، عدم‌آگاهی از موازی‌سازی در برنامه‌ها می‌تواند هزینه‌های کارایی را در مدیریت دما افزایش دهد. دسته‌ی دوم از این کارها، آگاهی از وجود موازی‌سازی در برنامه‌ها را در نظر گرفته‌اند [۱۶-۲۵]. در این دسته، برخی از کارها به مسئله‌ی کنترل دما پرداخته [۱۶،۱۷، ۱۹] و برخی دیگر نیز در جهت بهبود توان و انرژی مصرفی پردازنده تلاش کرده‌اند [۲۰-۲۵]. گرچه بهبود توان مصرفی و یا انرژی ممکن است در راستای اهداف این پژوهش یعنی عدم‌تخطی از محدودیت دمایی نباشد، اما ایده‌های به کار رفته در این کارها ممکن است برای کنترل دمای پردازنده هم کارایی داشته باشند.

از میان تحقیقات اخیر، الگوریتم آگاه از همسایگی به‌عنوان یک رویکرد پیشگیرانه مورد توجه زیادی بوده است [۱۳]. در این الگوریتم، دمای هر کدام از هسته‌های پردازنده با در نظر گرفتن تاثیر هسته‌های همسایه به‌صورت جداگانه پیش‌بینی می‌شود. در این کار، چنانچه دمای پیش‌بینی شده برای یک هسته از محدودیت دمایی بیشتر باشد، هسته‌ای به عنوان مقصد جهت مهاجرت پردازهای هسته‌ی جاری انتخاب می‌شود. برای انتخاب هسته‌ی مقصد از پایین‌ترین دمای پیش‌بینی شده‌ی هسته‌ها استفاده می‌شود که از محدودیت مشخص شده پایین‌تر است. شکل ۳، نمودار گردشی الگوریتم آگاه از همسایگی را نمایش می‌دهد. در این الگوریتم، T_i^{prev}، T_i^{cur} و T_i^{pred}، به ترتیب دمای گذشته، دمای حال و دمای پیش‌بینی شده برای هسته‌ی i هستند و T^{thr}، $\Delta(t)$ و S_i^{safe}، به‌ترتیب، آستانه محدودیت دمایی، فاصله‌ی زمانی بین دو اجرای متوالی الگوریتم و فرکانس مطمئن برای هسته‌ی i، می‌باشند. همچنین، $H(P_k)$، فاکتور گرمایی هسته‌ی مقصد می‌باشد که از فرمول زیر بدست می‌آید:

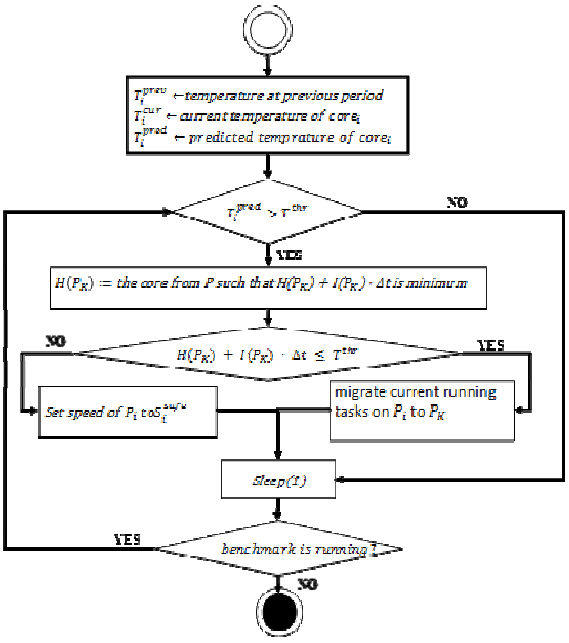
$$H(P_k) = \frac{\sum P_j \in P_k^{NB} \cup \{P_k\}^{T_j}}{|P_k^{NB} \cup \{P_k\}|} \quad (۱)$$

در واقع فرمول (۱)، میانگین دمای هسته‌ی k و هسته‌های مجاور آن را محاسبه می‌کند. از دید الگوریتم آگاه از همسایگی، هسته‌ای که کمترین مقدار $H(P_k)$ را دارد برای انتخاب به عنوان مقصد مناسب است. افزون بر این، فاکتور $I(P_k)$ که بیانگر میزان شیب افزایش دمای هسته‌ی k با درنظر گرفتن هسته‌های همسایه است، در الگوریتم آگاه از همسایگی از معیارهای انتخاب مقصد محسوب می‌شودکه مقادیر کمتر برای آن مطلوب‌تر هستند. $I(P_k)$ را می‌توان از فرمول زیر محاسبه کرد:

$$I(P_k) = \frac{\sum P_j \in P_k^{NB} \cup \{P_k\} \frac{T_j^{cur} - T_j^{prev}}{t_j^{cur} - t_j^{prev}}}{|P_k^{NB} \cup \{P_k\}|} \quad (۲)$$

در الگوریتم آگاه از همسایگی، چنانچه $H(P_k) + I(P_k).\Delta(t)$ برای هیچ یک از هسته‌ها پایین‌تر از محدودیت دمایی نباشد، فرکانس هسته‌ی جاری به سطح ایمن (S_i^{safe}) کاهش خواهد یافت تا دمای هسته‌ی جاری در آینده از محدودیت تعیین شده تجاوز نکند.

مزیت الگوریتم آگاه از همسایگی نسبت به دیگران، درنظرگرفتن تاثیر هسته‌های مجاور در مدل دمایی جهت پیش‌بینی دمای آینده برای یک هسته است تا تمایز میان دمای هسته‌ها مبنای مهاجرت پردازه‌ها قرار گیرد. به نظر می‌رسد، این روش در اجرای همزمان چندین برنامه‌ی غیرموازی کارایی مناسبی داشته باشد. در مقابل، در اجرای برنامه موازی ممکن است تفاوت چندانی میان فعالیت هسته‌های مختلف نباشد و مهاجرت نخ‌ها موجب کاهش دما نشود. در چنین حالتی کاهش شدید فرکانس در دستور کار قرار خواهد گرفت که بدون شک هزینه کارایی سنگینی را به مدیریت دما تحمیل خواهد کرد.



شکل ۳- نمودار گردشی الگوریتم آگاه از همسایگی [۱۳]

در [۱۴] برای پیش‌بینی دما علاوه بر دما و ویژگی فیزیکی هسته‌ها، ویژگی‌های بار کاری[۹] نیز درنظر گرفته شده است. در این روش یک مدل دمایی برای پیش‌بینی دما براساس بار کاری و یک مدل دمایی دیگر برای پیش‌بینی دما براساس ساختار پردازنده پیشنهاد شده است. در نهایت پیش‌بینی نهایی براساس مجموع ضرایبی از این دو، شکل می‌گیرد. این ضرایب براساس مشاهدات پیشین و به صورت تجربی بدست می‌آیند. زمانی که دمای یک هسته از یک مقدار مشخص بیشتر شود، هسته‌ای که کمترین پیش‌بینی دما را در آینده خواهد داشت، به عنوان مقصد مهاجرت انتخاب می‌شود.

در برخی از کارها از شمارنده‌های کارایی[۱۰] جهت پیش‌بینی دمای آینده استفاده شده است. در [۱۶]، با استفاده از یک آزمایش تجربی مشاهده شده که ترتیب مختلف اجرای وظایف در دما موثر است. به‌عنوان مثال اگر ابتدا یک وظیفه گرم و سپس یک وظیفه سرد اجرا شود، دمای هسته نسبت به حالت عکس کمتر خواهد بود. این مقاله به کمک شمارنده‌های کارایی یک تخمین زننده‌ی دما ارائه داده است که از آن در سیاستی استفاده شده که وظایف را به ترتیب دما مرتب کرده و داغ‌ترین وظیفه‌ای که دمای پردازنده را از حد آستانه بالاتر نمی‌برد را به عنوان وظیفه بعدی جهت اجرا انتخاب می‌کند.

یک پردازنده از واحدهای محاسباتی متفاوتی تشکیل شده است و یک برنامه و یا وظیفه ممکن است از تمامی این واحدهای محاسباتی استفاده نکند. و هر یک از وظایف ممکن است از واحدهای محاسباتی متفاوتی در پردازنده استفاده کنند. به عنوان مثال، یک وظیفه که عمده‌ی دستوراتش، اعمال ریاضی صحیح هستند، ALU را تحت تاثیر قرار می‌دهند در حالی که وظیفه‌ای که تمامی دستوراتش ممیز شناور است، واحد ممیز شناور را تحت تاثیر قرار داده و موجب افزایش دما در آن واحد می‌شود. در [۱۷]، سعی شده است تا در یک محیط چند نخی همزمان[۱۱]، وظایفی که از واحدهای محاسباتی متفاوت استفاده می‌کنند پشت سر هم اجرا شوند و یا وظایفی که از این نظر متمایز هستند همزمان اجرا شوند.

در [۱۸، ۱۹] یک الگوریتم توازن بار[۱۲] آگاه از دما ارائه شده است. که با استفاده از تغییر فرکانس و مهاجرت سعی می‌کند دما را زیر حد آستانه حفظ کند.

این روش در محیط برنامه‌نویسی charm++ پیاده‌سازی شده است. یکی از معایب مهم این کار، نیاز به آگاهی از زمان اجرای وظایف در الگوریتم زمانبندی است. باید توجه شود که در برنامه‌ها چنین اطلاعاتی به راحتی در دسترس نیست و به دست آوردن آن‌ها نیازمند پایش‌های آماری از زمان اجرا و استفاده از مدل‌های تخمینی است. در مقابل، از آنجایی که سیستم عامل از وجود وظایف موجود در برنامه مطلع نیست و در مدیریت اجرای آن‌ها نقشی ندارد، استفاده از این ایده برای مدیریت دما در سطح سیستم عامل، در عمل ممکن نیست.

یک برنامه‌ی موازی معمولاً علاوه بر قسمت‌هایی که موازی اجرا می‌شوند دارای قسمت‌هایی است که در آن موازی‌سازی صورت نگرفته و اجرا به صورت سریال است. بر این اساس، در برخی از کارها سعی شده است تا قسمت‌های سریال برنامه روی یک هسته‌ی قوی از پردازنده و قسمت‌های موازی برنامه روی دیگر هسته‌های ضعیف‌تر اجرا شوند تا ضمن افزایش کارایی برنامه، در مصرف انرژی صرفه‌جویی صورت گیرد. [۲۰، ۲۱]، مصرف انرژی در قسمت‌های سریال و موازی برنامه را با رویکردی در سطح کامپایلر کاهش داده‌اند. در اینکارها، در زمان کامپایل، قطعه کدی به قسمت‌های سریال برنامه افزوده می‌شود که فرکانس ساعت پردازنده را در حداکثر قرار می‌دهد و در قسمت‌های موازی برنامه قطعه کدی قرار داده می‌شود که سطح پایینی از فرکانس پردازنده را درخواست می‌کند. بنابراین، همواره قسمت‌های سریال برنامه با بیشترین فرکانس و قسمت‌های موازی با فرکانس پایین اجرا شده و در توان مصرفی صرفه‌جویی خواهد شد.

در تعداد زیادی از کارهای انجام شده در زمینه کاهش مصرف انرژی، تلاش شده است تا بهترین پیکربندی پردازنده برای اجرای برنامه پیش‌بینی شود تا ضمن کمترین تخریب کارایی، بیشترین صرفه‌جویی در مصرف انرژی صورت گیرد. به عبارت دیگر، در زمان اجرا و یا قبل از آن، مصرف انرژی و میزان کارایی برنامه در ترکیب‌های مختلفی از تعداد هسته‌ها و فرکانس ساعت پردازنده پیش‌بینی می‌شود و هر پیش‌بینی به عنوان یک نقطه‌ی عملیاتی[۱۳] سیستم تلقی می‌گردد. سپس، یکی از حالات که به اهداف موردنظر کاربر نزدیک‌تر است از میان نقاط عملیاتی انتخاب می‌شود.

در این راستا، [۲۲، ۲۳] از یک روش رگرسیون خطی[۱۴] برای دسته‌بندی و پیدا کردن نقاط عملیاتی سیستم استفاده کرده‌اند. در این کارها، داده‌های ورودی رگرسیون می‌تواند دما، شمارنده‌های کارایی، زمان اجرای برنامه، تعداد هسته‌ها و فرکانس پردازنده باشد. همچنین، خروجی زمان اجرا و انرژی مصرفی است. در این کارها، با افزایش تعداد انتخاب‌ها از میان نقاط عملیاتی ممکن، فضای جستجو بزرگ می‌شود و بررسی تمام این نقاط سربار زمان اجرا زیادی خواهد داشت. در [۲۴]، با استفاده از الگوریتم تپه‌نوردی[۱۵] و جستجوی دودویی فضای جستجوی نقاط عملیاتی هرس شده و بنابراین بار محاسباتی انتخاب نقطه‌ی عملیاتی بهینه کاهش می‌یابد گرچه ممکن است لزوماً بهترین حالت انتخاب نشود.

در اکثر کارهای انجام شده، زمانی که اهداف مختلفی از جمله کاهش دما، مصرف انرژی و زمان اجرای برنامه به صورت توامان موردنظر است، از روش بهینه‌سازی مبتنی بر محدودیت برای انتخاب تصمیم زمانبندی بهینه استفاده می‌شود. در این روش، یک هدف بیشینه یا کمینه می‌شود و دیگر اهداف به صورت محدودیت‌هایی در نظر گرفته می‌شوند.

به عنوان مثال، برای کمینه کردن زمان اجرا تلاش می‌شود که دما از محدودیت مشخصی فراتر نرود. با این حال، در معدودی از کارها سعی شده است تا تمامی اهداف به صورت همزمان بیشینه یا کمینه شوند. از این دست، در [۲۵] با استفاده از الگوریتم ژنتیک بهینه‌سازی توام کارایی، دما و توان مصرفی صورت گرفته است. با این وجود، سربار بالای زمان اجرای محاسبات تکاملی مانع کارآمدی راه حل پیشنهادی در این مقاله شده است.

۳- سیاست پیشنهادی مدیریت دما

سیاست پیشنهادی مدیریت پویای دما در این تحقیق بر مبنای یک مدل دمایی برای پیش‌بینی دمای پردازنده، یک مدل کارایی جهت پیش‌بینی میزان افزایش سرعت و یک الگوریتم انتخاب پیکربندی است که به کمک این دو مدل ضمن تضمین رعایت محدودیت دما، سعی در کاهش تخریب کارایی دارد. پس از انجام آزمایشات مختلف بر روی برنامه‌های محک موازی، پی‌بردیم که با تغییر تعداد نخ‌های فعال برنامه و فرکانس پردازنده، زمان اجرای برنامه تقریباً به صورت خطی و دمای پردازنده به صورت غیرخطی تغییر می‌کنند. بنابراین، ما با پیشنهاد دو مدل کارایی و دمایی، نقش عواملی همچون فرکانس ساعت پردازنده و تعداد هسته‌های فعال را در کارایی برنامه و دمای پردازنده تبیین کرده و الگوریتم مدیریت دمایی خود را بر پایه این مدل‌ها پیشنهاد می‌دهیم.

۳-۱- مدل پیش‌بینی دما

یک هدف اساسی در مدل پیش‌بینی دما، افزایش دقت در مقادیر پیش‌بینی شده است. از این رو، برای افزایش دقت پیش‌بینی مدل دما بایستی ویژگی‌های موثر در دمای پردازنده شناسایی و در مدل گنجانده شوند. با این وجود، شناخت این ویژگی‌ها نیازمند صرف آزمایشات متعدد و آزمودن ترکیبات مختلفی از ویژگی‌های آماری دمای پردازنده است. گرچه روش‌های متعددی برای ایجاد مدل‌های دمایی با استفاده از ویژگی‌های آماری موجود است، در این تحقیق ما از رگرسیون خطی برای تخمین ضرایب ویژگی‌های موثر مدل دمایی خود بهره بردیم که در تحقیقات دیگر مرسوم بوده است [۱۴]. فرمول (۳) یک معادله درجه اول از مدل دمایی پیشنهادی را نشان می‌دهد:

$$T_{new} = \alpha_1 T_{cur} + \alpha_2 f_{new} + \alpha_3 c_{new} + \alpha_4 \qquad (۳)$$

که در آن T_{new} پیش‌بینی دمای جدید، T_{cur} دمای حال، f_{new} فرکانس آینده پردازنده، c_{new} تعداد نخ‌های کارگر آینده و مقادیر α_i برای $i = \{1,2,\dots,4\}$ ضرایب تعیین شده توسط رگرسیون خطی هستند. پس از انجام آزمایشات متعدد روی برنامه‌های محک مختلف، به این نتیجه رسیدیم که علاوه بر ویژگی‌های استفاده شده در فرمول (۳) عوامل دیگری چون شیب تغییرات ویژگی‌ها نیز بر عملکرد الگوریتم مدیریت دمایی موثر است. فرمول (۴)، با دخیل کردن پارامترهای بیشتر را در یک معادله‌ی درجه دوم برای مدل دمایی پیشنهاد می‌دهد که در آن علاوه بر ویژگی‌های فرمول (۴)، میزان تغییرات ویژگی‌ها نیز جهت افزایش دقت مدل، گنجانده شده است.

$$\begin{aligned} T_{new} = {} & \alpha_1 T_{cur} + \alpha_2 f_{new} + \alpha_3 c_{new} + \alpha_4 \nabla T_{cur} + \alpha_5 \nabla f_{new} + \\ & \alpha_6 \nabla c_{new} + \alpha_7 T_{cur}^2 + \alpha_8 f_{new}^2 + \alpha_9 c_{new}^2 + \alpha_{10} \nabla T_{cur}^2 + \\ & \alpha_{11} \nabla f_{new}^2 + \alpha_{12} \nabla c_{new}^2 + \alpha_{13} \end{aligned} \qquad (۴)$$

عملگر گرادیان (∇) برای یک ویژگی به معنی تفاضل مقدار جدید آن ویژگی با مقدار قبلی آن در دو نمونه‌ی متوالی از زمان اجرا است. به عبارت دیگر برای ویژگی فرضی χ خواهیم داشت:

$$\nabla\chi_{cur} = \chi_{cur} - \chi_{past} \quad , \quad \nabla\chi_{new} = \chi_{new} - \chi_{cur}$$

که در آن، χ_{past}، χ_{cur} و $\nabla\chi_{new}$ به ترتیب مقدار پیشین، حال و آینده‌ی ویژگی فرضی χ است.

۳-۲- مدل پیش‌بینی کارایی

در مدل پیش‌بینی کارایی ما به دنبال تخمین تغییر در زمان اجرای برنامه در اثر تغییر پیکربندی سیستم هستیم. از این رو، تخمین توان پردازشی سیستم را به صورت زیر تبیین می‌کنیم:

$$\text{ProcessingPower} = \text{f} \times \text{c}$$

که در آن f فرکانس ساعت و c تعداد هسته‌های فعال پردازنده است.

در بسیاری از الگوریتم‌های مدیریت دمای پردازنده، از تنظیم در فرکانس ساعت و تغییر در تعداد هسته‌های فعال در پردازنده استفاده می‌شود. بنابراین، میزان تغییر در کارایی برنامه را پس از اعمال تغییرات نسبت به حالت کنونی را می‌توان از تقسیم تخمین توان پردازشی کنونی بر تخمین توان پردازشی جدید محاسبه کرد که خواهیم داشت:

$$\text{Speedup} = \frac{\text{f}_{\text{new}} \times \text{c}_{\text{new}}}{\text{f}_{\text{cur}} \times \text{c}_{\text{cur}}} \quad (۵)$$

که در آن f_{cur} و c_{cur} به ترتیب فرکانس فعلی و تعداد هسته‌های فعال فعلی پردازنده است. همچنین، f_{new} و c_{new} به‌ترتیب فرکانس آینده و تعداد هسته‌های فعال آینده پردازنده است.

بدون شک، هر چه مقدار Speedup در فرمول (۵) بزرگتر باشد، افزایش سرعت بیشتری در اثر تغییر در فرکانس ساعت و تعداد هسته‌های فعال پردازنده حاصل خواهد شد. بنابراین، افزایش مقدار Speedup می‌تواند از اهداف الگوریتم مدیریت دما باشد.

۳-۳- الگوریتم مدیریت دما

ما در الگوریتم پویای دمای پیشنهادی، به نحوی تعداد هسته‌های فعال و فرکانس ساعت پردازنده‌ی چندهسته‌ای را تعیین می‌کنیم که ضمن تضمین عدم‌تخطی دما از میزان آستانه، حداقل آسیب کارایی به برنامه وارد شود. در شکل ۴، الگوریتم مدیریت دمای پیشنهادی نمایش داده شده است. الگوریتم پیشنهادی ما شامل دو بخش مقداردهی اولیه و اجرای متوالی است. در بخش مقداردهی اولیه که در خطوط یک تا پنج در شکل ۴ مشاهده می‌شود، مجموعه‌ی فرکانس‌های ساعت و تعداد هسته‌های پردازنده در متغیرهایی ذخیره شده و نیز تعداد هسته‌های فعال برابر با تمام هسته‌ها و فرکانس ساعت برابر با حداکثر فرکانس ممکن مقداردهی می‌شوند. همچنین، محدودیت دمایی کاربر به آستانه‌ی دمایی ایمن تبدیل می‌شود. آستانه دمایی ایمن، عدم‌تخطی از محدودیت دمایی کاربر را تضمین می‌کند و همواره مقداری کمتر یا مساوی آن است. برای بدست آوردن مقدار آستانه دمایی ایمن، می‌توان ابتدا در الگوریتم مدیریت دمای پیشنهادی مقدار آستانه را برابر با محدودیت دمایی کاربر فرض کرد. سپس، با اجرای برنامه‌های محک مختلف مقدار حداکثر دمای اندازه‌گیری شده برای هر آستانه را می‌توان به عنوان محدودیت دمایی قابل تضمین با آن آستانه محسوب کرد. به عنوان مثال، ممکن است نیاز باشد برای تضمین عدم‌تخطی از محدودیت دمایی ۶۰ درجه، میزان آستانه دمایی ایمن را برابر با ۴۸ درجه در نظر گرفت.

در بخش دوم که خطوط شش تا ۱۴ الگوریتم را شامل می‌شود، فرآیندی توصیف می‌شود که در طی مدت اجرای برنامه‌ی موازی در سیستم ادامه می‌یابد. در این فرآیند، ابتدا (خط شش) تنظیمات جدید فرکانس و تعداد هسته‌های فعال پردازنده با محاسباتی تعیین می‌شود که در آن بیشینه‌ی کارایی با شرط عدم‌تجاوز از آستانه ایمن دما حاصل می‌آید. برای انجام چنین محاسباتی از بین تمامی ترکیبات تعداد هسته‌ها و فرکانس پردازنده که در آن‌ها دمای پردازنده با استفاده از فرمول (۳) یا (۴) کمتر از مقدار آستانه تخمین زده می‌شود، گزینه‌ای انتخاب شود که براساس فرمول (۵) بیشترین کارایی را دارد. در گام بعدی، اگر تعداد هسته‌های فعال جدید کمتر از هسته‌های فعال کنونی باشند، به تعداد تفاضل این دو، از نخ‌های کارگر فعال موجود معلق خواهند شد تا در عمل از هسته‌های فعال پردازنده کاسته شود. همچنین، اگر تعداد هسته‌های فعال جدید بیشتر از هسته‌های فعال کنونی باشد، به تعداد تفاضل این دو از نخ‌های کارگری که قبلا تعلیق شده‌اند برای ادامه‌ی اجرا بیدار خواهند شد. در نهایت، فرکانس پردازنده با فرکانس جدید تنظیم خواهد شد. لازم به ذکر است که در عموم پیاده‌سازی‌های رابایش کار، تعداد نخ‌های کارگر در ابتدای اجرا برابر با تعداد هسته‌های پردازنده است.

```
1: Φ ← set of processor frequencies
2: P ← set of processor cores
3: f_cur ← maximum processor frequency
4: c_cur ← the number of processor cores
5: T_safe ← getSafeThreshold(T_constraint )
6: while(a parallel program is running) do
7:   f_new , c_new ← argmax_{f_new ∈Φ, c_new ∈P} {Speedup | T_new ≤ T_safe }
8:   if c_new ≤ c_cur then
9:     Suspend (c_new − c_cur ) worker threads
10:  else
11:    Awake (c_cur − c_new ) worker threads
12:  end if
13:  f_cur ← f_new
14:  c_cur ← c_new
15:  sleep(1)
16: end while
```

شکل ۴- الگوریتم مدیریت دمای پیشنهادی

۴- ارزیابی

ما الگوریتم مدیریت دمای پیشنهادی خود را بر روی یک سیستم واقعی دارای پردازنده‌ی Intel Core i7-4790K که دارای چهار هسته‌ی فیزیکی، هشت هسته منطقی و ۱۶ پله‌ی فرکانسی مختلف است آزمایش کردیم. این سیستم دارای ۸ گیگابایت حافظه اصلی بود و از سیستم عامل Ubuntu 14.04 به عنوان میزبان استفاده شد. برای پیاده‌سازی الگوریتم پیشنهادی از زبان C++ و کامپایلر GCC استفاده شد. همچنین، برنامه‌های محک مورد آزمایش شامل مرتب‌سازی ادغامی، ضرب استاندارد ماتریس، تبدیل فوریه سریع، ترانهاده ماتریس، مرتب‌سازی سریع و مرتب‌سازی رقمی با استفاده از ابزار Intel Cilk Plus [۲۸] پیاده‌سازی شدند. افزون بر این، ما برای خواندن دمای هسته‌های پردازنده از ابزار lm-sensors [۲۶] و برای تنظیم فرکانس پردازنده از ابزار CPUFreq [۲۷] استفاده کردیم. در طول زمان آزمایشات تلاش شد تا دمای اتاق ثابت و برای تمامی آزمایشات یکسان نگاه داشته شود. همچنین، ما با استفاده از ابزار lm-sensors سرعت خنک‌کننده‌ی پردازنده را در تمامی آزمایشات ثابت نگاه داشتیم تا تغییر رفتار خنک‌کننده در نتایج اثری نداشته باشد.

۴-۱- ارزیابی دقت مدل دمایی

ما هر یک از مدل‌های دمایی ارئه شده در بخش ۱-۳ را در الگوریتم مدیریت دمایی پیشنهادی قرار داده تا اثر افزایش دقت مدل‌ها بر عملکرد سیاست پیشنهادی مدیریت دما را ارزیابی کنیم. در هر اجرا، دمای پردازنده به صورت لحظه‌ای اندازه‌گیری شد و با دمای پیش‌بینی شده برای آن لحظه مقایسه گردید. شکل ۵، نتایج حاصل از این آزمایش را برای برنامه‌های محک مختلف نشان

می‌دهد. در نمودارهای ارائه شده در شکل ۵، محور افقی محدودیت‌های دمایی مختلف و محور عمودی میزان خطای مطلق پیش‌بینی دمای پردازنده برحسب درجه سانتیگراد است. همان‌طور که در این شکل مشاهده می‌شود، برای بیشتر برنامه‌های محک، مدل دمایی فرمول (۳) از نظر میزان قدر مطلق خطای پیش‌بینی، عملکرد بهتری نسبت به مدل دمایی فرمول (۲) دارد. به طور میانگین، مدل دمایی فرمول (۳) از نظر زمان اجرا به میزان ۴۴ درصد و از نظر میانگین خطای پیش‌بینی ۱۶ درصد عملکرد بهتری نسبت به مدل دمایی فرمول (۲) دارد. این نتایج نشان می‌دهد که افزایش دقت مدل به میزان قابل توجهی در کارایی برنامه و عملکرد الگوریتم مدیریت دمایی موثر است. بنابراین، ما مدل پیش‌بینی دمای فرمول (۳) را برای استفاده در سیاست پیشنهادی مدیریت دمای خود انتخاب کردیم.

۴-۲- آستانه دمایی ایمن

همان‌طور که در بخش ۳-۳ بیان شد، در ابتدای اجرای الگوریتم پیشنهادی مدیریت دما نیاز است تا محدودیت دمایی درخواستی کاربر به آستانه دمایی ایمن تبدیل شود. برای این منظور، ما آستانه‌های دمایی ایمن را با اجرای برنامه‌های محک بر روی بستر آزمایش و براساس توضیحات بخش ۳-۳ محاسبه کردیم. جدول ۱، مقادیر آستانه محاسبه شده را برای تعدادی از محدودیت‌های دمایی کاربر نمایش می‌دهد. همان‌طور که در این جدول مشاهده می‌شود، همواره آستانه دمایی ایمن برای هر محدودیت، مقداری کمتر از آن است.

۴-۳- ارزیابی روش پیشنهادی

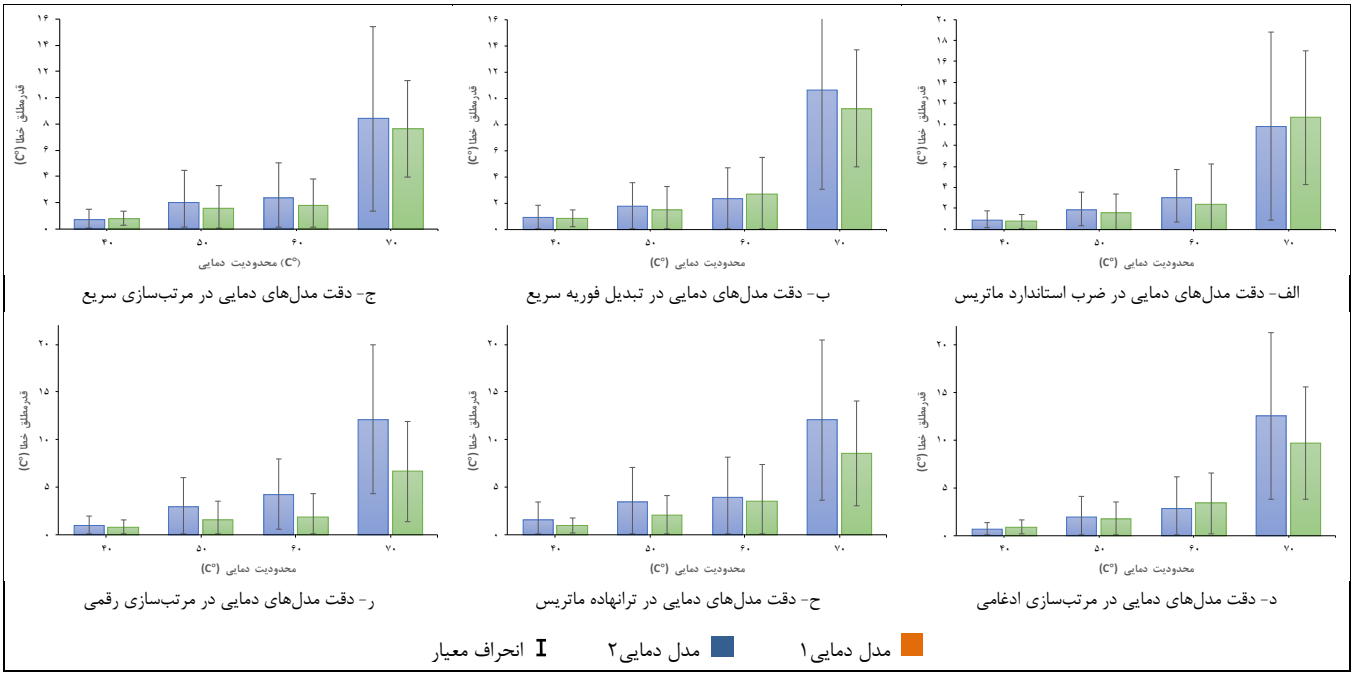
برای ارزیابی الگوریتم پیشنهادی، ما الگوریتم مدیریت پویای دمای آگاه از همسایگی [۴] را پیاده‌سازی کردیم. سپس، در اجرای آزمایشات دمای پردازنده را به صورت لحظه‌ای نمونه‌گیری کرده و مدت زمان اجرای برنامه‌های محک ثبت نمودیم. محدودیت‌های دمایی در نظر گرفته شده در آزمایشات شامل ۴۰، ۵۰، ۶۰ و ۷۰ درجه سانتیگراد بودند.

جدول ۱- آستانه دمایی ایمن برای محدودیت کاربر

محدودیت دمایی کاربر	آستانه دمایی ایمن
۴۰	۳۵
۴۵	۳۸
۵۰	۴۰
۵۵	۴۱
۶۰	۴۷
۶۵	۵۰
۷۰	۵۵

نمودارهای ارائه شده در شکل ۶، نتایج حاصل از این آزمایشات را به‌صورت خلاصه بیان کرده است. در هر یک از این نمودارها، محور افقی زمان اجرای برنامه محک بر حسب ثانیه و محور عمودی دمای پردازنده بر حسب درجه سانتیگراد است. براساس این نتایج، در تمامی موارد زمان اجرای برنامه محک تحت مدیریت الگوریتم پیشنهادی بسیار کمتر از زمان اجرا تحت مدیریت الگوریتم آگاه از همسایگی است. همچنین، الگوریتم پیشنهادی تمامی محدودیت‌های دمایی را رعایت کرده و هرگز اجازه نمی‌دهد که دمای پردازنده از محدودیت تعیین شده فراتر رود. این در حالی است که الگوریتم آگاه از همسایگی به‌طور میانگین در ۲۹ درصد از زمان اجرا از محدودیت دمایی تعیین شده تخطی کرده و نوسانات شدیدی در دمای پردازنده ایجاد می‌کند.

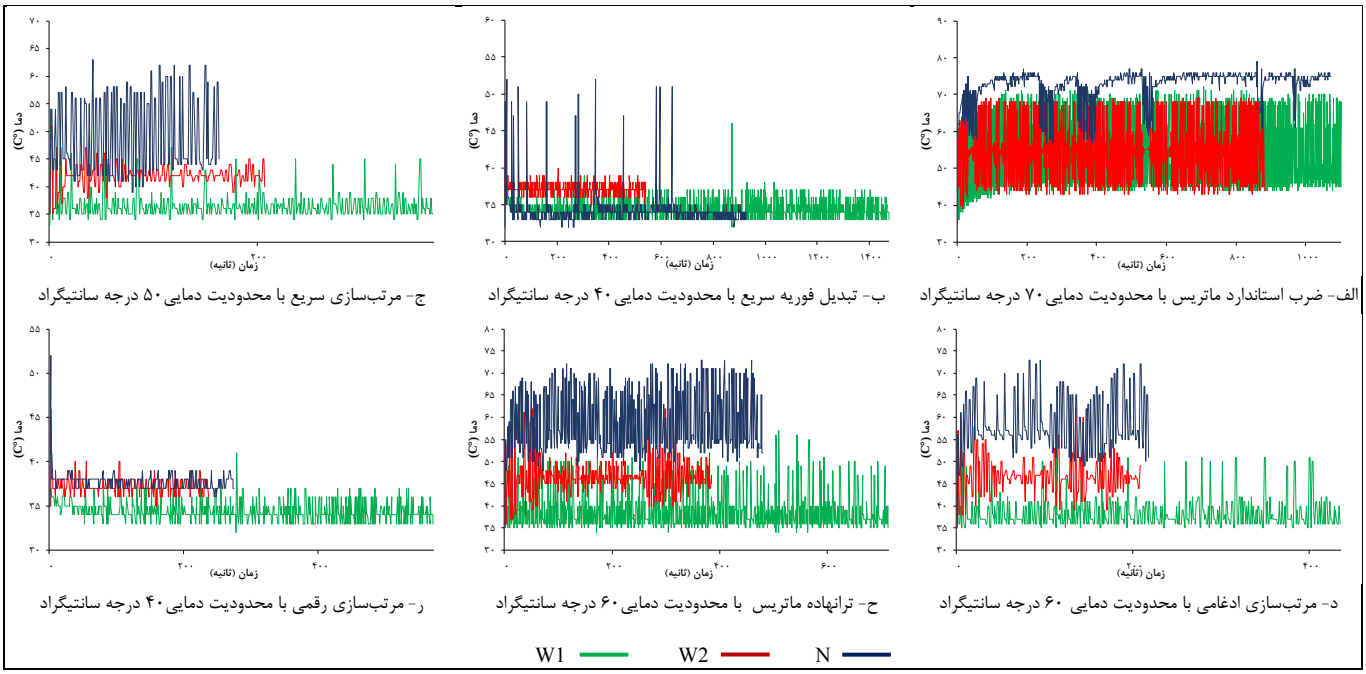
شکل ۷، میانگین زمان اجرای برنامه‌های محک مختلف در محدودیت‌های دمایی ۴۰، ۵۰، ۶۰ و ۷۰ درجه را نمایش می‌دهد. در این شکل، ستون مربوط به میانگین زمان اجرا در الگوریتم پیشنهادی با W1 و W2، و برای الگوریتم آگاه از همسایگی با حرف N نمایش داده شده است. همچنین، از حرف P برای نمایش زمان اجرا توسط زمانبند performance استفاده شده است. در هر ستون، ارتفاع آبی رنگ مدت زمانی است که دما در محدوده‌ای کمتر از محدودیت تعیین قرار داشته است. همچنین، میانگین زمان تخطی از محدودیت دمایی با رنگ نارنجی نمایش داده شده است.



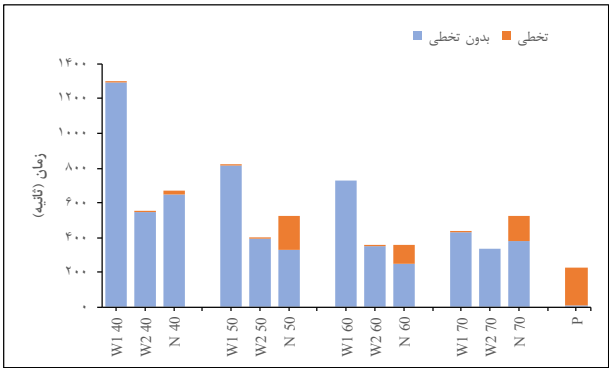
شکل۵- دقت مدل‌های دمایی در برنامه‌های محک مختلف. محور افقی آستانه‌های ایمن دمایی مختلف و محور عمودی قدرمطلق خطای پیش‌بینی دما

بر این اساس، مشاهده می‌گردد که توسط الگوریتم آگاه از همسایگی در محدودیت‌های دمایی مختلف به طور میانگین برای ۱۵ درصد از زمان اجرا، دمای پردازنده از محدودیت تعیین شده تخطی کرده است. این در حالی است که الگوریتم پیشنهادی تقریباً در تمامی موارد توانسته است دمای پردازنده را کاملا زیر محدودیت تعیین شده مدیریت کند و به‌طور میانگین ۲۸ درصد نسبت به روش آگاه از همسایگی و ۴۶ درصد نسبت به روش پایه با مدل دمایی یک، کارایی بهتری داشته باشد.

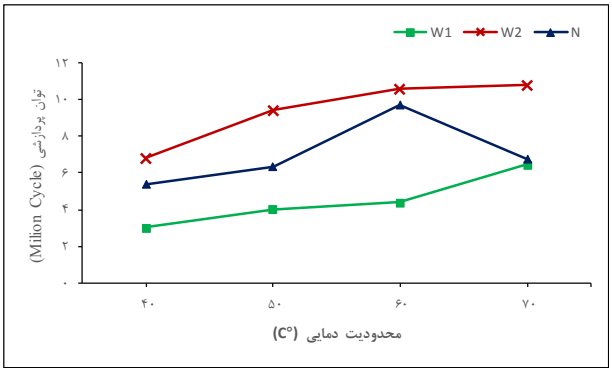
همچنین، ما مشاهده کردیم که همواره میانگین توان پردازشی در الگوریتم پیشنهادی نسبت به الگوریتم آگاه از همسایگی بیشتر بوده ولی این توان بیشتر با ترکیب تعداد هسته‌های فعال بیشتر و فرکانس پایین‌تر حاصل آمده است. شکل ۸، میانگین توان پردازشی پردازنده را برای الگوریتم پیشنهادی و الگوریتم آگاه از همسایگی در اجرای برنامه‌های محک نمایش داده است. با توجه به این شکل، هر چه سطح محدودیت دما بیشتر باشد، الگوریتم‌های مدیریت دما ناگزیر به اعمال سطح پایین‌تری از توان پردازشی هستند که منجر به افزایش زمان اجرای برنامه‌های موازی خواهد شد.



شکل ۶- مقایسه عملکرد روش‌ها در برنامه‌های محک مختلف. محور افقی زمان برحسب ثانیه و محور عمودی دمای لحظه‌ای برحسب درجه سانتیگراد



شکل۷- میانگین زمان اجرای برنامه‌های محک تحت محدودیت‌های دمایی ۴۰، ۵۰، ۶۰ و ۷۰ درجه سانتیگراد



شکل۸- میزان توان پردازشی الگوریتم پیشنهادی در مقایسه با الگوریتم آگاه از همسایگی

۵- نتیجه‌گیری

در این پژوهش، ما یک الگوریتم مدیریت پویای دما را در سطح سیستم عامل پیشنهاد دادیم که دمای پردازنده‌ی چندهسته‌ای را در اجرای برنامه‌های موازی پیاده‌سازی شده توسط زمانبند رایش کار مدیریت می‌کند. الگوریتم ما مبتنی بر دو مدل دمایی و کارایی پیشنهادی است که دمای پردازنده و تغییرات زمان اجرای را پیش‌بینی می‌کنند. همچنین تاثیر افزایش دقت مدل دمایی را نیز مورد بررسی قرار دادیم و مشاهده کردیم که افزایش دقت مدل نه‌تنها زمان اجرای برنامه را کاهش می‌دهد بلکه برعملکرد الگوریتم مدیریت دما نیز موثر بوده و خطای پیش‌بینی دما را نیز کاهش می‌دهد. آزمایشات بر روی سیستم واقعی نشان داد که الگوریتم پیشنهادی ما عملکرد بسیار بهتری از الگوریتم آگاه از همسایگی دارد. بنابراین، ما الگوریتم پیشنهادی خود را راه حلی مناسب برای مدیریت دمای پردازنده‌ی چندهسته‌ای در اجرای برنامه‌های موازی رایش کار می‌دانیم.

در آینده تصمیم داریم دقت مدل‌های دمایی و کارایی را بهبود بخشیم و اثر این بهبود را در عملکرد الگوریتم مدیریت دمای پیشنهادی ارزیابی کنیم. همچنین، قصد داریم تا ارزیابی‌های خود را بر روی پردازنده‌ها و برنامه‌های محک متنوع‌تری گسترش دهیم.

مراجع

[1] J. Kong, S. W. Chung, and K. Skadron, "Recent thermal management techniques for microprocessors,"In Computing Surveys (CSUR), vol. 44, no. 3, pp.1-42, 2012.

[2] J. Diaz, C. Munoz-Caro, and A. Nino, "A survey of parallel programming models and tools in the multi and many-core era," IEEE Transactions on parallel and distributed systems, vol. 23, no. 8, pp. 1369-1386, 2012.

[3] S. Zhuravlev, J. C.Saez, Blagodurov, S. Fedorova, and M. Prieto, "Survey of energy-cognizant scheduling techniques," IEEE Transactions on Parallel and Distributed Systems, vol. 24, no. 7, pp. 1447-1464, 2013.

[4] T. van Dijk, and C. Jaco, "Lace: non-blocking split deque for work-stealing," In European Conference on Parallel Processing, pp. 206-217, 2014.

[5] A. Morrison, and Y. Afek, "Fence-free work stealing on bounded TSO processors," ACM SIGPLAN, vol. 49, no. 4, pp. 413-426, 2014.

[6] J. Nakashim, S.Nakatani, and K. Taura, "Design and implementation of a customizable work stealing scheduler," In Proceedings of the 3rd International Workshop on Runtime and Operating Systems for Supercomputers, pp. 1-9, 2013.

[7] M. Wimme, D. Cederma, J. L. Träff, and P. Tsigas, "Work-stealing with configurable scheduling strategies," In ACM SIGPLAN, vol. 48, no. 8, pp. 315-316, 2013.

[8] A. Bhattacharjee, and M. Martonosi, "Thread criticality predictors for dynamic performance, power, and resource management in chip multiprocessors," InACM SIGARCH Computer Architecture, vol. 37, no. 3, pp. 290-301, 2009.

[9] D. Hendler, and N. Shavi, "Non-blocking steal-half work queues," In Proceedings of the twenty-first annual symposium on Principles of distributed computing, pp. 280-289, 2002.

[10] U. Acar, A. Charguéraud, and M. Rainey, "Scheduling parallel programs by work stealing with private deques," In ACM SIGPLAN, vol. 48, no. 8, pp. 219-228, 2013.

[11] D. Hendler, Y. Lev, M. Moir, and N. Shavit, "Adynamic-sized nonblocking work stealing deque," Technical report, Sun Microsystems, 2005.

[12] S. Imam, and V. Sarkar, "Load balancing prioritized tasks via work-stealing," In Euro-Par'15, pp. 222-234, 2015.

[13] G. Liu, M. Fan, G. Quan, and M. Qiu, "On-Line predictive thermal management under peak temperature constraints for practical multi-core platforms,"In Journal of Low Power Electronics, vol. 8, no. 5, pp. 565-578, 2012.

[14] I. Yeo, C. C. Liu, and E. J. Kim, "Predictive dynamic thermal management for multicore systems," In Proceedings of the 45th annual Design Automation Conference, pp. 734-739, 2008.

[15] R. Cochran, and R. Sherie, "Thermal prediction and adaptive control through workload phase detection," ACM Transactions on Design Automation of Electronic Systems, vol. 18, no. 7, 2013.

[16] J. Yang, X. Zhou, M. Chrobak, Y. Zhang, and L. Jin, "Dynamic thermal management through task scheduling," IEEE International Symposium on Performance Analysis of Systems and software (ISPASS'08), pp. 191-201, 2008.

[17] A. Merkel, and F. Bellosa, "Task activity vectors: a new metric for temperature-aware scheduling," In SIGOPS Operating Systems Review, vol. 42, no. 4, pp.1-12, 2008.

[18] O. Sarood, P. Miller, E. Totoni, and L.V. Kale, "Load Balancing for High Performance Computing Data Centers," IEEE Transactions on Computers, vol. 61, no. 12, pp. 1752-1764, 2012.

[19] O. Sarood, "Optimizing performance under thermal and power constraints for HPC data centers,"Doctoral dissertation, University of Illinois at Urbana-Champaign, 2014.

[20] T. Lu, P. P. Pande, and B. Shirazi,"A Dynamic, Compiler Guided DVFS Mechanism to Achieve Energy-Efficiency in Multi-coreProcessors," Sustainable Computing: Informatics and Systems, vol. 12, pp. 1-9, 2016.

[21] R. A. Shafik, A. Das, S. Yang, G. Merrett, and B. M. Al-Hashimi, "Adaptive energy minimization of OpenMP parallel applications on many-core systems," In Proceedings of the 6th Workshop on Parallel Programming and Run-Time Management Techniques for Many-core Architectures, pp. 19-24, 2015.

[22] R. Cochran, C. Hankendi, and A. Coskun, "Identifying the optimal energy-efficient operating points of parallel workloads," In IEEE/ACM International Conference on Computer-Aided Design, pp. 608-615, 2011.

[23] D. D. Sensi, "Predicting Performance and Power Consumption of Parallel Applications," International Conference on Parallel, Distributed, and Network-Based Processing, pp. 200-207, 2016.

[24] J. Li, and J. F. Martinez, "Dynamic power-performance adaptation of parallel computation on chip multiprocessors," In The Twelfth International Symposium on High-Performance Computer Architecture, pp. 77-87, 2006.

[25] H. F. Sheikh, I. Ahmad, and D. Fan, "An Evolutionary Technique for Performance-Energy-Temperature Optimized

Scheduling of Parallel Tasks on Multi-Core Processors," IEEE Transactions on Parallel and Distributed Systems, vol. 27, no. 3, pp. 668-681, 2016.

[26] lm-sensors Linux hardware monitoring [Online]. Available: http://www.lm-sensors.org, Jan 2017.

[27] Linux cpufreq governors, Linux Kernel [Online]. Available:https://www.kernel.org-/doc/Documentation/cpu-freq/governors.txt. Jan 2017.

[28] Intel Cilk Plus [Online]. Available: http://www.cilkplus.org, Jan 2017.

حمید گوهرجو دانش آموخته رشته علوم کامپیوتر دانشگاه گلستان- گرگان در مقطع کارشناسی. در حال حاظر دانشجوی کارشناسی‌ارشد رشته مهندسی کامپیوتر - نرم‌افزار دانشگاه فردوسی مشهد و عضو آزمایشگاه معماری کامپیوتر پیشرفته دانشگاه فردوسی مشهد. علاقه‌مندی‌ها: برنامه‌نویسی موازی، زمانبندی، سیستم عامل و مدیریت دمای پردازنده.
آدرس پست‌الکترونیکی ایشان عبارت است از:
ha.goharjoo@mail.um.ac.ir

مرتضی مرادی در سال ۱۳۸۶ از دانشگاه بیرجند در مقطع کارشناسی رشته مهندسی کامپیوتر فارغ‌التحصیل شد. سپس، در سال ۱۳۹۱ موفق به اخذ مدرک کارشناسی‌ارشد از دانشگاه آزاد اسلامی واحد مشهد شد. او در حال حاضر دانشجوی دکتری مهندسی کامپیوتر- نرم‌افزار گرایش سیستم‌های نرم‌افزاری در دانشگاه فردوسی مشهد است. علایق تحقیقاتی وی شامل طراحی الگوریتم‌های موازی، مدیریت حافظه‌های اشتراکی و توزیع شده، کنترل دما و توان مصرفی پردازنده و پرداختن به مسئله زمانبندی اجرای وظایف است.
آدرس پست‌الکترونیکی ایشان عبارت است از:
morteza.moradi@mail.um.ac.ir

حمید نوری مقطع کارشناسی و کارشناسی‌ارشد خود را به ترتیب در سال ۱۳۷۵ و ۱۳۷۹ در رشته مهندسی کامپیوتر از دانشگاه صنعتی شریف و دانشگاه صنعتی امیرکبیر دریافت کرد. مقطع دکتری خود را در رشته مهندسی کامپیوتر در دانشگاه کیوشو در ژاپن گذرانده است. در سال‌های ۱۳۸۷ تا ۱۳۸۹ استادیار دانشکده مهندسی برق و کامپیوتر دانشگاه تهران بوده است. از سال ۱۳۸۹ تا به اکنون استادیار گروه مهندسی کامپیوتر در دانشگاه فردوسی مشهد و سرپرست آزمایشگاه معماری کامپیوتر پیشرفته می‌باشد. زمینه‌های تحقیقاتی ایشان شامل معماری کامپیوتر، پردازنده‌های چند هسته‌ای، پردازش موازی و سیستم‌های نهفته می‌باشد.
آدرس پست‌الکترونیکی ایشان عبارت است از:
hnoori@um.ac.ir

اطلاعات بررسی مقاله:

تاریخ ارسال: ۱۳۹۵/۱۰/۲۷
تاریخ اصلاح: ۱۳۹۵/۱۱/۲۰
تاریخ قبول شدن: ۱۳۹۵/۱۱/۲۹
نویسنده مرتبط: دکتر حمید نوری، دانشکده مهندسی، دانشگاه فردوسی مشهد، مشهد، ایران.

[1]Work Stealing
[2]Predictive Dynamic Thermal Management (PDTM)
[3]Dynamic Voltage and Frequency Scaling (DVFS)
[4]Task Migration
[5]Worker Thread
[6]Hot Spots
[7]Reactive
[8]Predictive
[9]Workload
[10]Performance Counters
[11]Simultaneous Multi-Threading (SMT)
[12]Load Balancing
[13]Operating Points
[14]Leaner Regression
[15]Hill Climbing

علوم رایانش و فناوری اطلاعات
نشریه علمی انجمن کامپیوتر ایران
مجلد ۱۴، شماره ۲، ۱۳۹۵
صفحات ۳۹- ۴۶
مقاله عادی

یادگیری عمیق در خلاصه‌سازی چندسندی متون فارسی

شیما محرابی حمیدرضا احمدی‌فر سید ابوالقاسم میرروشندل

دانشکده فنی، دانشگاه گیلان، رشت، ایران

چکیده

با پیشرفت علوم و تکنولوژی و در نتیجه افزایش حجم اطلاعات متنی قابل دسترس از طریق اینترنت، وجود سامانه‌های خلاصه‌ساز که چکیده‌ای از اطلاعات موردنظر را در کوتاه‌ترین زمان ممکن در دسترس کاربر قرار دهند، ضروری به‌نظر می‌رسد. خلاصه‌سازی خودکار متون از دیرباز مورد توجه پژوهشگران حوزه‌ی پردازش زبان‌های طبیعی قرار گرفته است. امروزه با بهبود توان پردازشی سیستم‌های موجود و ظهور ابزارهای محاسباتی نوین، تلاش برای افزایش کارایی سیستم‌های خلاصه‌ساز ادامه دارد. در این مقاله به معرفی یک سامانه‌ی خلاصه‌ساز استخراجی چندسندی متون فارسی می‌پردازیم. این سامانه برای امتیازدهی به جملات از نظر میزان اهمیت آنها در سند، از روشی تحت عنوان یادگیری عمیق بهره می‌برد. یادگیری عمیق، روشی برای آموزش ماشین برپایه‌ی شبکه‌های عصبی مصنوعی است. پیش از این، یادگیری عمیق در زمینه‌های پردازش صوت و تصویر و همچنین پردازش زبان‌های طبیعی مورد استفاده قرار گرفته است. نتایج خوب بدست آمده از این روش در مقایسه با دیگر روش‌های مرسوم، انگیزه‌ای در بکارگیری این روش در خلاصه‌سازی خودکار چندسندی متون فارسی گشت. در خلاصه‌ساز پیشنهادی با استفاده از یک شبکه‌ی Autoencoder عمیق، عمل امتیازدهی به جملات انجام می‌گیرد و میزان دقت خلاصه‌ساز در ارزیابی جمله‌ای قابل‌قبول به‌نظر می‌رسد.

کلمات کلیدی: پردازش زبان فارسی، خلاصه‌سازی خودکار چندسندی، یادگیری عمیق، شبکه‌های عصبی مصنوعی.

۱- مقدمه

در دنیای امروزی، با پیشرفت علوم و تکنولوژی، شاهد افزایش روزافزون حجم اطلاعات متنی قابل دسترس هستیم. بنابراین دستیابی به اطلاعات مورد نظر در حجمی کمتر اما کاربردی و دارای جامعیت قابل‌قبول، بسیار مطلوب به‌نظر می‌رسد. خلاصه‌سازی متون[1] توسط انسان باوجود داشتن مزایایی از قبیل صحت و جامعیت مستلزم صرف وقت و هزینه‌ی بالایی است. همچنین خلاصه‌سازی اسناد بزرگ به‌صورت دستی برای انسان کاری دشوار است. این در حالی است که متون موجود برروی اینترنت فراوان است و اینترنت بیشترین اطلاعات مورد نیاز افراد را فراهم می‌کند. لذا وجود یک سیستم خلاصه‌ساز کامپیوتری باعث صرفه‌جویی در زمان و هزینه‌ی مصرفی در تولید متن خلاصه خواهد شد، هرچند که ممکن است از لحاظ صحت و جامعیت با خلاصه‌ی تولید شده توسط انسان برابری نکند.

بطور کلی هدف از خلاصه‌سازی خودکار متن، فشرده‌سازی و کوتاه نمودن متن اصلی با حفظ محتوا و معنای کلی آن است. خلاصه‌سازی متون منجر به استفاده از منابع بیشتر با سرعت بالاتر و در نتیجه دستیابی به اطلاعاتی غنی‌تر می‌شود. یک خلاصه خوب باید موضوعات گوناگون یک یا چند سند را بدون افزونگی دربرداشته باشد.

به شکل‌های مختلفی می‌توان خلاصه‌سازی را طبقه‌بندی کرد. یک روش طبقه‌بندی از نظر شکل و فرم سازماندهی متن خلاصه است. در این حالت متن خلاصه شده به یکی از روش‌های استخراجی[2] و یا چکیده ای[3] حاصل می‌گردد. در روش استخراجی، جملات مهم متن انتخاب شده و به همان شکل اصلی خود در متن خلاصه ظاهر می‌شوند. در خلاصه‌ی چکیده‌ای، خلاصه‌ی مفهومی از متن در خروجی خلاصه‌ساز تولید می‌شود. در واقع ممکن است، فرم جملات تغییر یابد. این روش مشابه روشی است که انسان برای خلاصه‌سازی متون بکار می‌گیرد [۱]. همچنین گونه‌ی دیگری از دسته‌بندی خلاصه، براساس تعداد اسناد شرکت داده شده در تولید خلاصه است. بر این اساس خلاصه‌سازی به دو دسته‌ی تک‌سندی و چندسندی تقسیم‌بندی می‌شود. در خلاصه‌سازی تک‌سندی، از یک سند برای ایجاد متن خلاصه استفاده می‌شود اما در خلاصه‌سازی چندسندی، ورودی سیستم چندین سند با موضوع کلی مشترک است که جنبه‌های مختلفی از آن موضوع را

تحت پوشش قرار می‌دهد

یکی از مهمترین چالش‌ها در امر خلاصه‌سازی متون، انتخاب بهترین جملات متن اصلی است به‌طوری که جنبه‌های مهم و کلیدی متن را شامل شده و در عین حال فاقد افزونگی باشد. در نتیجه لازم است که متن اصلی پیش‌پردازش شود و ویژگی‌هایی که در انتخاب بهترین جملات تاثیرگذارند، استخراج شوند. مرحله‌ی پیش‌پردازش و استخراج ویژگی‌ها نقش بسزایی در حصول نتیجه‌ی مطلوب، ایفا می‌کند. در خلاصه‌سازی استخراجی جملات در قالب بردارهایی قرار می‌گیرند که هر بردار شامل ویژگی‌هایی است که اهمیت یک جمله را از جهات مختلف بیان می‌دارد. یک بردار ویژگی برداری n عضوی است که هر عضو آن یک مقدار عددی دارد. تصمیم‌گیری در مورد میزان اهمیت یک جمله براساس این مقادیر صورت می‌گیرد.

یکی از معروف‌ترین سیستم‌های خلاصه‌سازی چندسندی، تحت عنوان MEAD شناخته می‌شود [۲]. Gistsumm نیز یک خلاصه‌ساز استخراجی است که از ۳ بخش تکه‌سازی متن، امتیازدهی به جملات و ایجاد خلاصه تشکیل شده است [۳]. قدیمی‌ترین کاری که در زبان فارسی برای یک سیستم خلاصه‌ساز صورت گرفته، یک خلاصه‌ساز استخراجی تک‌سندی به‌نام FarsiSum است [۴]. این خلاصه‌ساز یک برنامه‌ی کاربردی تحت وب است و شکل‌گیری آن بر پایه‌ی یک پروژه‌ی خلاصه‌سازی در زبان سوئدی است. در [۵] یک استخراج کننده متن فارسی تک‌سندی ارائه شده است که براساس زنجیره‌ی زبانی و روش‌های مبتنی بر گراف عمل می‌کند. در [۶] یک سیستم خلاصه‌ساز چندسندی- چندزبانی ایجاد و معرفی شده است که براساس SVR[۴] و دسته‌بندی چندسطحی عمل می‌کند. در [۷] یک خلاصه‌ساز براساس منطق فازی معرفی شده است. در این سیستم، ویژگی‌های یک متن نظیر طول جمله، میزان تشابه با عنوان و تشابه با کلمات کلیدی به‌عنوان ورودی سیستم فازی درنظر گرفته شده و عمل خلاصه‌سازی انجام می‌شود.

از سال ۲۰۰۶ حوزه‌ی جدیدی از تحقیقات یادگیری ماشین با نام یادگیری ساختاریافته عمیق و یا بطور مصطلح‌تر یادگیری عمیق[۵] پا به عرصه‌ی وجود نهاد [۸]. در طی چند سال اخیر تکنیک‌های حاصله از تحقیقات یادگیری عمیق، حجم وسیعی از پژوهش‌های مربوط به پردازش سیگنال و اطلاعات را تحت تأثیر خود قرار داده است. یادگیری عمیق عموماً از شبکه‌های عصبی مصنوعی بهره می‌برد. سطوح بالاتر در شبکه به واسطه‌ی اطلاعات حاصله از سطوح پایین‌تر تعریف می‌شوند. یادگیری عمیق را می‌توان تقاطعی میان حوزه‌های تحقیقاتی شامل هوش مصنوعی، شبکه‌های عصبی، تشخیص الگو و پردازش سیگنال دانست. ۳ عامل مهم در محبوبیت امروزی یادگیری عمیق، افزایش روزافزون توان پردازشی تراشه‌های محاسباتی، افزایش قابل‌توجه حجم اطلاعات مورد استفاده برای یادگیری و پیشرفت‌های اخیر در پژوهش‌های مرتبط با پردازش اطلاعات و سیگنال است [۹].

یکی از مقالات مهم در زمینه‌ی یادگیری عمیق در سال ۲۰۰۹ منتشر شد و در آن دلایل اهمیت این روش و مزایای آن برشمرده شد. امکان یادگیری چندسطحی و توانایی استخراج ویژگی‌ها بطور مستقیم از داده‌های ورودی از نقاط قوت این روش محسوب می‌شود [۱۰]. در [۱۱] شبکه‌های عصبی عمیق را در زمینه‌ی پردازش صوت و گفتار مورد آزمایش قرار داده‌اند. ارزیابی‌های صورت گرفته در مقاله‌ی مذکور، نشان داده است که استفاده از شبکه‌های عصبی عمیق در پردازش گفتار نسبت به روش‌های مرسوم نظیر مدل مارکوف پنهان، با افزایش کارایی همراه است.

نتایج بدست آمده توسط یادگیری عمیق در حوزه‌های پردازش صوت و تصویر، پژوهشگران پردازش زبان‌های طبیعی را برآن داشت که از این روش یادگیری در مسائل مربوط به پردازش زبان نیز بهره ببرند. در سال ۲۰۱۱ یک معماری واحد و یکپارچه معرفی شد که قابل‌اعمال بر روی مسائل متنوعی در پردازش زبان است. مسائل قابل حل توسط این معماری شامل تشخیص موجودیت‌های نامدار[۶]، برچسب‌گذاری اجزای کلام[۷]، برچسب‌گذاری نقش معنایی کلمات[۸] و تکه‌بندی گرامری جملات[۹] است [۱۲، ۱۳]. در [۱۴] از یادگیری عمیق در مدل‌سازی زبانی[۱۰] استفاده شده است. در این مقاله مشاهده شده است که میزان نرخ خطای کلمات و همچنین میزان حیرانی در انتخاب کلمه به نسبت روش‌های دیگر بهبود یافته است. در [۱۵] یک چهارچوب خلاصه‌سازی چندسندی با استفاده از یادگیری عمیق ارائه شده است که بردار ویژگی‌ها در آن شامل تعداد تکرار کلمات یک مجموعه لغات از پیش تعیین شده، در اسناد موردنظر برای خلاصه‌سازی است. بنابراین ورودی شبکه، برداری از ویژگی‌ها و خروجی آن مجموعه‌ای از جملات خواهد بود. برای ایجاد خلاصه ساز از یک RBM[۱۱] استفاده می‌شود. ابتدا شبکه سعی در حذف کلمات غیرضروری دارد، سپس کلمات مهم از بین کلمات باقی‌مانده تعیین می‌شوند. جملاتی که حاوی کلمات مهم هستند، استخراج شده و در نهایت با استفاده از برنامه‌نویسی پویا و با در نظر گرفتن محدودیت حجم از پیش تعریف شده متن خلاصه، از بین جملات مهم، جملات مناسب برای شکل‌گیری خلاصه انتخاب می‌گردد.

با توجه به دستاوردهای بدست آمده از یادگیری عمیق در پردازش زبان‌های طبیعی برآن شدیم که از توانایی‌های این روش یادگیری در خلاصه‌سازی خودکار متون فارسی بهره ببریم. در این پژوهش به معرفی یک سیستم خلاصه‌سازی استخراجی چندسندی برای متون فارسی می‌پردازیم که در آن عمل انتخاب جملات مهم براساس امتیازاتی است که به هر جمله توسط شبکه عمیق اختصاص داده شده است. شبکه‌ی عمیق پیشنهادی، یک شبکه‌ی ۹ لایه است که در لایه‌ی ورودی ویژگی‌های هر جمله شامل TF/IDF، میزان تشابه با عنوان سند، مکان رخداد جمله و امتیاز مربوط به برچسب اجزای کلام جمله به شبکه داده می‌شود. در نهایت طی آموزش لایه‌ای شبکه به هر جمله یک امتیاز اختصاص داده می‌شود، سپس جملات براساس امتیازشان مرتب شده و جملات با بالاترین امتیاز برای تشکیل متن خلاصه انتخاب می‌شوند. ساختار این مقاله بدین شرح است: در بخش دوم به معرفی یادگیری عمیق پرداخته می‌شود و در بخش سوم روش پیشنهادی برای خلاصه‌سازی چندسندی متون فارسی ارائه خواهد شد. بخش چهارم ارزیابی روش پیشنهادی را شامل می‌شود و در پایان نیز نتیجه‌گیری ذکر شده است.

۲- معرفی یادگیری عمیق

مکانیزم پردازش اطلاعات توسط انسان نظیر بینایی و شنوایی، به نوعی بازگوکننده‌ی نیاز به معماری عمیق برای استخراج ساختارهای پیچیده‌ی داده‌های ورودی است. به‌عنوان مثال سیستم بینایی انسان ذاتاً از یک ساختار سلسله مراتبی برای درک تصاویر بهره می‌برد و به‌عنوان ورودی رنگ، اندازه، جهت و عواملی از این دست را مدنظر قرار می‌دهد. مفهوم یادگیری عمیق نشأت گرفته از پژوهش‌های حوزه‌ی شبکه‌های عصبی مصنوعی است. شبکه‌های پیش‌خور[۱۲] و یا پرسپترون چند لایه با تعداد لایه‌های مخفی زیاد، نمونه‌های خوبی از مدل‌هایی با معماری عمیق هستند. الگوریتم Back-Propagation که در دهه‌ی ۱۹۸۰ به محبوبیت رسید، یک الگوریتم معروف در یادگیری پارامترهای این شبکه محسوب می‌شود.

آموزش شبکه‌های عمیق دشوار است، روش‌هایی که بطور مؤثر و بهینه بر روی شبکه‌های کم‌عمق اعمال می‌شود، در شبکه‌های عمیق چندان کارآمد نیستند. این مشکل با معرفی روشی تحت عنوان پیش‌آموزش لایه‌ای بدون‌ناظر[۱۳]، در شبکه‌های عمیق مرتفع گشت. بطور دقیق‌تر، در یک ساختار یادگیری عمیق، هر لایه بطور مجزا در نظر گرفته می‌شود. در این روش به محض اینکه لایه‌های قبلی آموزش داده شدند، لایه‌ی بعدی با استفاده از داده‌های حاصل از لایه‌ی قبلی آموزش می‌بینند. سپس برروی کل شبکه یک مرحله تنظیم کلی یا fine-tuning انجام

می‌گیرد [۹].

از مدل‌های مرسوم در یادگیری عمیق می‌توان به ماشین بولتزمن محدود شده و Autoencoderها اشاره کرد. ماشین بولتزمن محدود شده و یا به اختصار RBM مدلی برای نمایش یک توزیع احتمال است. به‌منظور آموزش RBM و با فراهم آوردن مجموعه داده‌های آموزشی، این مدل سعی در تنظیم پارامترهای خود دارد، به قسمی که توزیع احتمال ارائه شده توسط RBM به بهترین شکل ممکن در برگیرنده‌ی داده‌های آموزشی باشد. با معرفی یک معماری چند لایه به‌نام شبکه‌های باور عمیق[۱۴] که متشکل از چندین RBM متصل به هم هستند، RBMها بیش از پیش مورد توجه قرار گرفتند. ایده‌ی این معماری استخراج ویژگی‌های مرتبط با نورون‌های ورودی توسط نورون‌های مخفی است. سپس این ویژگی‌ها به‌عنوان ورودی RBM بعدی عمل می‌نمایند. بدین شکل و با پشته‌سازیRBMها، شبکه قادر به یادگیری ویژگی‌های جدید از ویژگی‌های حاصله‌ی قبلی خواهد بود [۱۶].

نوع خاصی از شبکه‌های عصبی عمیق تحت عنوان Autoencoderها شناخته می‌شود که در آن‌ها بردار خروجی مشابه بردار ورودی است. این نوع شبکه اغلب برای یادگیری ویژگی‌ها با رمزگذاری مؤثر داده‌های ورودی بکار گرفته می‌شود. Autoencoder یک روش استخراج ویژگی به شکل غیرخطی و بدون استفاده از داده‌های برچسب‌دار است. یک Autoencoder دارای یک لایه‌ی ورودی است که نمایش‌دهنده‌ی داده‌های ورودی شبکه است (به‌عنوان مثال، پیکسل‌های یک تصویر). همچنین این مدل شامل یک یا چند لایه‌ی مخفی است که نشان‌دهنده‌ی ویژگی‌های تغییریافته هستند و دارای یک لایه خروجی مطابق لایه‌ی ورودی خود است. تعداد نورون‌های مخفی می‌توانند بیشتر و یا کمتر از تعداد نورون‌های ورودی باشد. یک Autoencoder اغلب توسط یکی از اشکال الگوریتم Back-Propagation آموزش می‌بینند و عموماً روش Stochastic gradient descent را مورد استفاده قرار می‌دهد [۱۷].

۳- روش پیشنهادی خلاصه‌سازی

یکی از مراحل پایه‌ای در امر خلاصه‌سازی متون، مرحله‌ی پیش‌پردازش متن ورودی است. اولین گام پیش‌پردازش متن، نرمال‌سازی آن است. نرمال‌سازی به عمل یکدست‌سازی واحدهای متنی به‌طوری که قابل پردازش توسط ماشین باشند، اطلاق می‌گردد. گاهی حروف به‌کار رفته در دو کلمه‌ی یکسان، با یکدیگر متفاوتند این در حالی است که از دید انسان آن دو کلمه یکسان محسوب می‌شوند. این مسأله باعث عدم‌شناسایی یکسان بودن دو کلمه توسط ماشین می‌شود در نتیجه لازم است متن نرمال‌سازی شود تا یک شکل واحد برای کلمات یکسان بدست آید. به‌عنوان مثال «ها» را می‌توان به ۳ شکل، چسبان، جدا بافاصله و جدا با نیم‌فاصله بکار برد. در نتیجه ۳ شکل متفاوت از یک کلمه‌ی جمع بسته شده با «ها» تولید می‌شود. به‌منظور نرمال‌سازی متن می‌بایست یکی از این حالات را به‌عنوان شکل قابل‌قبول درنظر گرفت و تمامی اشکال دیگر آن کلمه به شکل قابل‌قبول تعیین شده، تبدیل شوند.

در گام بعدی می‌بایست متن را به واحدهایی براساس جملات و کلمات تقسیم‌بندی کنیم. در واقع باید مرز جملات و کلمات شناسایی شود. علائمی نظیر نقطه (اگر محصور به عدد نباشد) و"\n" را می‌توان به‌عنوان پایان یک جمله قلمداد کرد. همچنین علائمی نظیر خط‌فاصله، فضای خالی و ویرگول را می‌توان به‌عنوان مرز کلمات در نظر گرفت. در مرحله‌ی بعد، کلمات ریشه‌یابی شده و کلمات ایست[۱۵] حذف می‌گردند. کلمات ایست، کلماتی پرتکرار و بی‌اهمیت از لحاظ بارمعنایی هستند که عدم حذف آن‌ها بواسطه‌ی تعداد تکرار بالایی که دارند، ممکن است سیستم را در شناسایی کلمات پراهمیت متن، دچار اشتباه کند. کلماتی نظیر «است»، «برای»، «شده» در مجموعه‌ی کلمات ایست قرار می‌گیرند.

۳-۱- تولید بردار ویژگی‌ها

در این پژوهش برای آموزش خلاصه‌ساز، ۴ ویژگی معرفی می‌شود. این ۴ ویژگی تشکیل یک بردار ویژگی را می‌دهند. هر جمله از متن دارای یک بردار ویژگی است. ویژگی‌ها شامل میزان تکرار کلمات، میزان تشابه جملات با عنوان سند، موقعیت قرارگیری جمله در سند و امتیاز مربوط به برچسب‌گذاری اجزای کلام جمله است.

۳-۱-۱- ویژگی میزان تکرار کلمات

برای اندازه‌گیری میزان تکرار کلمات از فراوانی وزنی TF/IDF بهره برده شده است. در این شیوه به کلمات یک وزن براساس فراوانی آنها در سند داده می‌شود. در واقع این سیستم وزن‌دهی نشان می‌دهد که به چه میزان یک کلمه در یک سند مهم است. میزان تکرار کلمه در یک سند با TF(f,d) نشان داده می‌شود و وزن نهایی با استفاده IDF بدست می‌آید. IDF به معنای عکس فراوانی سند است که نشان‌دهنده‌ی میزان فراوانی کلمه موردنظر در اسناد دیگر است، در واقع به این سوال پاسخ می‌دهد که آیا کلمه موردنظر، در تمامی اسناد متداول است یا خیر. عکس فراوانی سند با لگاریتم‌گیری از نتیجه‌ی تقسیم تعداد کل اسناد بر تعداد اسنادی که کلمه‌ی موردنظر در آن‌ها ظاهر شده، بدست می‌آید. در نهایت با ضرب TF در IDF میزان وزن اختصاص داده شده به کلمات براساس تکرار آنها بدست می‌آید. در رابطه‌ی (۱) روش محاسبه‌ی IDF و در رابطه‌ی (۲) نیز روش محاسبه‌ی معیار TF/IDF نشان داده شده است.

$$IDF(t, D) = \log(\frac{D}{d \in D : t \in d}) \quad (۱)$$

در رابطه‌ی (۱)، D به کلیه‌ی اسناد موجود و t به کلمه موردنظر اشاره دارد. $d \in D : t \in d$ تعداد اسنادی که کلمه t در آن موجود است را نشان می‌دهد.

$$TF/IDF\ (t.d.D): TF(t, d) \times IDF(t, D) \quad (۲)$$

در رابطه‌ی (۲)، منظور از TF(t,d)، تعداد دفعاتی است که کلمه‌ی t در سند d تکرار شده است. در روش پیشنهادی، میانگین مقادیر معیار TF/IDF کلمات یک جمله، بیانگر مقدار ویژگی میزان تکرار کلمات آن جمله خواهد بود، در واقع طول جملات بر حسب تعداد کلمات آن در محاسبه مقدار ویژگی تکرار کلمات یک جمله در نظر گرفته می‌شود. روش محاسبه ویژگی میزان تکرار کلمات یک جمله در رابطه (۳) نشان داده شده است.

$$Sentence\ TF/IDF\ Feature: \frac{\sum_{i=1}^{n} TF/IDF(w_i.d,D)}{n} \quad (۳)$$

در رابطه‌ی (۳)، w_i، کلمه iام در جمله S و n تعداد کل کلمات موجود در جمله‌ی S را نشان می‌دهد.

۳-۱-۲- ویژگی میزان تشابه جملات با عنوان سند

نسبت تعداد کلمات مشترک در یک جمله از سند به تعداد کل کلمات عنوان را امتیاز جمله از نظر میزان تشابه آن با عنوان، در نظر می‌گیریم. البته میزان تشابه با

عنوان بعد از انجام عملیات پیش‌پردازش جمله و عنوان، امتیازدهی می‌گردد. رابطه‌ی (۴) روش محاسبه‌ی ویژگی تشابه با عنوان را نشان می‌دهد. در این رابطه مقدار میزان اشتراک جمله مورد نظر بر تعداد کلمات عنوان نرمال می‌گردد، این عمل سبب می‌شود که درک مناسب‌تری از میزان تشابهات ایجاد گردد، چرا که اگر سند دارای یک عنوان طولانی باشد مطمئناً شمارش تعداد تشابهات به تنهایی محاسبه‌ی درستی از معیار نخواهد بود.

$$\text{TitleSimilarityFeature} : \frac{|S \cap T|}{|T|} \quad (۴)$$

در رابطه‌ی (۴)، $|S \cap T|$ نشان دهنده‌ی تعداد کلمات مشترک بین عنوان و جمله‌ی موردنظر است و $|T|$ تعداد کلمات عنوان را نشان می‌دهد.

۳-۱-۳- ویژگی موقعیت جمله

در متون خبری فارسی عموماً جملات ابتدایی و انتهایی متن حاوی بار معنایی بیشتری نسبت به جملات میانی هستند. از آنجایی که متون مورد استفاده در سامانه‌ی پیشنهادی، متون خبری است لذا از این قاعده به عنوان یکی از معیارهای امتیازدهی به جملات سند بهره می‌بریم. در واقع اگر جمله‌ی موردنظر در ابتدای متن و یا در انتهای متن ظاهر شود، میزان اهمیت آن جمله نسبت به مابقی جملات بیشتر خواهد بود لذا به ویژگی موقعیت جمله‌ی مربوط به جملات ابتدایی و انتهایی سند مقدار عددی ۱ اختصاص داده می‌شود و ویژگی موقعیت جمله در سایر جملات به صفر مقداردهی شده است. هدف از این ویژگی برجسته کردن میزان اهمیت جملات ابتدایی و انتهایی سند در فرآیند امتیازدهی کلی به یک جمله است.

۳-۱-۴- ویژگی برچسب اجزای کلام

برچسب‌زنی اجزای کلام به عمل دسته‌بندی کلمات یک متن براساس ماهیت آنها از نظر اجزای گرامری زبان است. بنابراین هر کلمه را می‌توان در یکی از دسته‌های فعل، اسم، صفت، قید و غیره قرار داد. اسامی، یکی از مهمترین کلمات جمله در انتقال بار معنایی آن محسوب می‌شوند. صفات نیز به‌عنوان یکی از اجزای کلام می‌تواند حاوی اطلاعات مفیدی از نظر معنای جمله باشد [۱۸، ۱۹]. در این پژوهش امتیاز مربوط به اجزای کلام، از مجموع تعداد اسامی و صفات بکاررفته در یک جمله، تقسیم بر تعداد کل کلمات آن جمله بدست می‌آید. رابطه‌ی (۵) نحوه محاسبه‌ی این ویژگی را نشان می‌دهد.

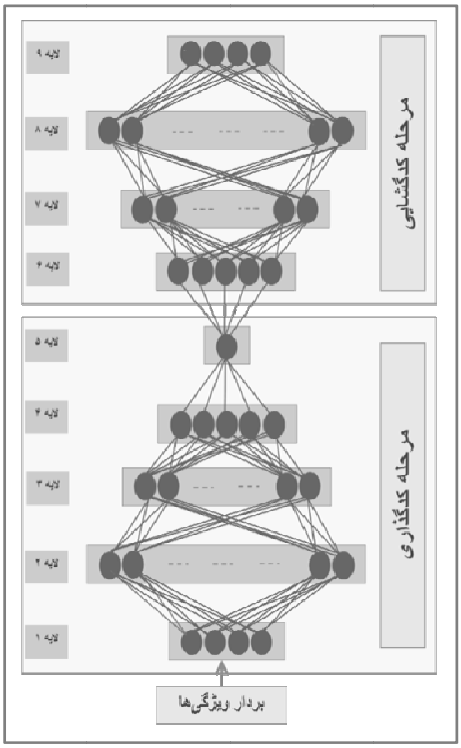
$$\text{POSScoreFeature} : (S_{|N|} + S_{|Adj|}) / |S| \quad (۵)$$

در رابطه‌ی (۵)، $S_{|N|}$ تعداد اسامی در جمله S، $S_{|Adj|}$ تعداد صفات در جمله‌ی S و $|S|$ تعداد کل کلمات موجود در جمله‌ی S را نشان می‌دهد.

۳-۲- استفاده از یادگیری عمیق در امتیازدهی به جملات و تولید متن خلاصه

پس از مرحله‌ی پیش‌پردازش متن و تولید بردار ویژگی‌ها می‌بایست به جملات امتیازی مبنی بر میزان اهمیت آنها اختصاص داده شود. روش پیشنهادی ما برای امتیازدهی به جملات، استفاده از یادگیری عمیق و به کمک یک شبکه‌ی عصبی است. در این پژوهش از یک شبکه‌ی Autoencoder که به انضمام لایه‌ی ورودی شامل ۹ لایه است، بهره برده‌ایم. در Autoencoderها لایه‌ی ورودی مشابه لایه‌ی خروجی است و هدف بازسازی مقادیر گره‌های ورودی در لایه‌ی خروجی است. در نتیجه داده‌های ما بدون برچسب هستند. عملکرد این نوع شبکه به دو بخش کدگذاری و کدگشایی تقسیم‌بندی می‌شود. در مرحله‌ی کدگذاری، شبکه سعی در کدگذاری داده‌ها و تولید ویژگی‌های جدید از ویژگی‌های ورودی دارد، پس از آن و در مرحله کدگشایی، شبکه به بازسازی داده‌های ورودی از روی ویژگی‌های جدید تولید شده در پایان مرحله‌ی کدگذاری، می‌پردازد. شبکه‌های Autoencoder در آموزش‌های بدون ناظر یعنی زمانی که داده‌های آموزشی ورودی فاقد برچسب هدف هستند، کاربرد دارند.

تصمیم‌گیری در مورد تعداد لایه‌ها و گره‌های مخفی در شبکه عملی تجربی بوده و با آزمایشات مکرر می‌بایست در مورد بهترین حالت تصمیم‌گیری نمود. در شبکه‌ی پیشنهادی ما لایه‌ی ورودی و یا همان لایه اول شامل ۴ نورون است که به‌عنوان ورودی، بردار ویژگی ۴ عضوی استخراج شده‌ی مربوط به ویژگی‌های هر جمله به آن تزریق می‌شود. در لایه‌های دوم، سوم و چهارم به‌ترتیب ۱۵، ۱۰ و ۵ نورون داریم و لایه‌ی پنجم، جایی که مرحله‌ی کدگذاری به پایان می‌رسد، شبکه دارای یک نورون است که مقدار آن امتیازی است که شبکه به جمله‌ی ورودی می‌دهد. بعد از آن، مرحله‌ی بازسازی ورودی و یا کدگشایی آغاز می‌شود. بنابراین لایه‌های ششم، هفتم، هشتم و نهم به ترتیب شامل ۵، ۱۰، ۱۵ و ۴ نورون هستند. شکل ۱ شمایی از شبکه‌ی پیشنهادی را نشان می‌دهد.



شکل ۱- شبکه عمیق پیشنهادی

برای آموزش و آزمایش سیستم خلاصه‌ساز پیشنهادی از پیکره‌ی پاسخ [۲۰] استفاده شده است. پیکره‌ی «پاسخ» اولین پیکره‌ی متنی برای ارزیابی خلاصه‌سازی تک‌سندی و چندسندی است که توسط آزمایشگاه فناوری وب دانشگاه فردوسی مشهد و با همکاری سازمان فناوری اطلاعات ایران تولید گردیده است. «پاسخ» یک پایگاه داده متشکل از حجم زیادی از اسناد خبری فارسی در موضوعات متنوع است. همچنین این پیکره دربردارنده‌ی خلاصه‌های تولید شده

توسط عوامل انسانی به اَشکال تک‌سندی، چندسندی، استخراجی و چکیده‌ای است. تیم تولیدکننده پیکره «پاسخ» برای ساخت متن خلاصه توسط عامل انسانی، از ۱۰ دانشجوی آموزش دیده، کمک گرفته است. به‌منظور جلوگیری از تاثیر سلایق و گرایشات شخصی در تولید خلاصه، عمل خلاصه‌سازی هر سند به ۵ نفر محول گشته است. پیکره «پاسخ» در بخش چندسندی دارای ۵۰ عنوان کلی خبری است که هر عنوان شامل ۲۰ سند است. در مجموع پیکره «پاسخ» در بخش چندسندی حاوی ۱۰۰۰ سند است که از این بین ۸۰۰ سند در مرحله آموزش و ۲۰۰ سند در مرحله آزمایش سامانه‌ی پیشنهادی مورد استفاده قرار گرفته است.

به‌منظور آموزش شبکه از ۸۰۰ سند متنی خبری در ۴۰ دامنه‌ی موضوعی متفاوت موجود در پیکره «پاسخ»، بهره گرفته شده است. مجموعاً تعداد ۱۳۷۹۵ جمله برای آموزش شبکه پیش‌پردازش شد و بردار ویژگی‌های مختص به هر جمله استخراج گردید.پس از آموزش شبکه و تنظیم پارامترهای مربوط به آن، در فاز ایجاد متن خلاصه، از شبکه‌ی آموزش دیده شده برای امتیازدهی به جملات استفاده می‌شود. سپس جملات براساس امتیاز اختصاص داده به آن‌ها مرتب شده و با توجه به نرخ فشرده‌سازی متن خلاصه که پیش فرض آن ۱۰ درصد است، جملاتی که بالاترین امتیاز را دارند برای تشکیل متن خلاصه انتخاب می‌شوند.

۴- ارزیابی روش پیشنهادی

برای ارزیابی روش پیشنهادی، با توجه به مطالعات و جستجوهای انجام شده، سیستم خلاصه‌ساز چندسندی قابل دسترسی دیگری که از پیکره‌ی پاسخ برای آموزش و ارزیابی استفاده کرده باشد، یافت نشد. یکی از شروط لازم برای مقایسه‌ی دو سامانه‌ی خلاصه‌سازی خودکار، اشتراک در داده‌های آزمایشی است. به دلیل عدم‌دسترسی به سیستم خلاصه‌ساز چندسندی فارسی دیگری که بتوان داده‌های آزمایشی برگرفته از پیکره «پاسخ» و مشابه با سامانه پیشنهادی را به آن اعمال نمود لذا ارزیابی‌های انجام شده در این مقاله محدود به مقایسه‌ی خروجی سیستم خلاصه‌ساز پیشنهادی با خلاصه‌های انسانی موجود در پیکره‌ی پاسخ، می‌شود.

سامانه‌ی پیشنهادی برروی ۱۰ مجموعه سند موجود در پیکره «پاسخ» که هر یک دارای ۲۰ سند خبری است، مورد ارزیابی و آزمایش قرار گرفت. تعداد ۲۴۹۳ جمله‌ی آزمایشی برای تولید بردار ویژگی پردازش گشت و شبکه‌ی پیشنهادی برای امتیازدهی به این داده‌های آزمایشی، به کار گماشته شد و خلاصه‌های متناظر با هر مجموعه سند تولید گشت.

یک روش ارزیابی استفاده شده در این مقاله، مقایسه‌ی جمله به جمله‌ی متن خروجی سامانه با خلاصه‌ی انسانی متناظر با آن است، جملاتی که عیناً مشابه یکدیگرند در امتیازدهی به عملکرد سامانه‌ی خلاصه‌ساز شرکت داده می‌شوند. معیارهای ارزیابی Precision، Recall و F-Score برای متن خلاصه محاسبه شده است. به‌عنوان مثال، برای یک مجموعه ۲۰ سندی که شامل ۱۸۰ جمله است، یک خلاصه‌ی ۱۸ جمله‌ای توسط خلاصه‌ساز تولید شد، خلاصه‌ی انسانی متناظر با این مجموعه سند شامل ۱۵جمله است. تعداد جملاتی که عیناً در هردو خلاصه‌ی تولید شده توسط سیستم و خلاصه‌ی تولید شده توسط انسان، ظاهر شده ۷ جمله است.

همچنین سیستم خلاصه‌ساز پیشنهادی را در ۵ حالت مختلف از نظر تعداد و نوع ویژگی‌های استخراج شده برای هر جمله، مورد ارزیابی قرار دادیم. بدین‌منظور از ۳ ویژگی دیگر برای جملات استفاده شده است که این ۳ ویژگی شامل تعداد کلمات ایست در جمله، وجود ضمیر در ابتدای جمله و طول جمله است [۲۱]. معمولاً جملاتی که حاوی تعداد کلمات ایست بالایی هستند، کلمات پراهمیت کمتری را شامل می‌شوند. در نتیجه نسبت تعداد کلمات ایست یک جمله به کل تعداد کلمات آن جمله به‌عنوان ویژگی تعداد کلمات ایست در جمله در نظر گرفته شده است. به‌طور کلی جملاتی که آغازگر آن‌ها یک ضمیر است به نوعی حاوی توضیحات در رابطه با جملات ماقبل خود هستند، در واقع این نوع جملات وابسته به جملات دیگر بوده و لحاظ کردن آن‌ها در متن خلاصه ممکن است میزان خوانایی آن متن را کاهش دهد. نسبتی از تاثیر مکان قرارگیری ضمیر در جمله، تعداد ضمایر در جمله و تعداد کلمات آن جمله بیانگر مقدار ویژگی تاثیر ضمیر در جمله است. ویژگی دیگر تاثیر طول جمله است، جملات طولانی و همچنین جملات بسیار کوتاه، جملات مناسبی برای انتخاب در متن خلاصه نیستند لذا حد آستانه‌ای برای طول جملات در نظر گرفته می‌شود و جملات نسبت به آن حد آستانه امتیازدهی می‌شوند. در هر یک از حالات سعی شده بهترین مجموعه ممکن از ویژگی‌ها انتخاب شود. جدول ۱ در بردارنده‌ی ویژگی‌های انتخابی در هر یک از این حالات است.

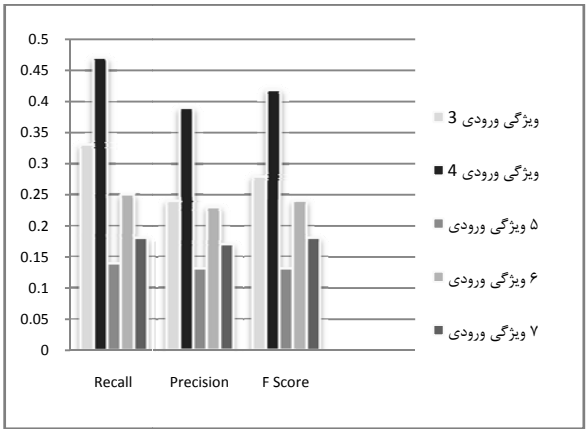
جدول ۱- حالات مختلف از نظر تعداد ویژگی‌های استخراج شده

تعدادویژگی‌ها / ویژگی‌ها	۳	۴	۵	۶	۷
میزان تشابه با عنوان	*	*	*	*	*
مکان رخداد جمله	*	*	*	*	*
TF/IDF	*	*	*	*	*
برچسب اجزای کلام		*	*	*	*
تعداد کلمات ایست در جمله			*	*	*
وجود ضمیر در ابتدای جمله				*	*
تاثیر طول جمله					*

نتایج حاصل از مقایسه‌ی این ۵ حالت در شکل ۲ نشان داده شده است. همانطور که قابل مشاهده است، زمانی که شبکه از ۴ ویژگی برای ارزیابی میزان اهمیت جمله استفاده نموده، نتایج بهتری را نسبت به ۴ حالت دیگر کسب کرده است. می‌توان یکی از دلایل رخداد این امر را مسأله‌ی تنک[16] بودن داده‌های آموزشی دانست. در واقع ممکن است افزایش و یا کاهش تعداد ویژگی‌ها و ترکیبات آن‌ها به‌دلیل عدم‌گستردگی حالات به وجود آمده در سطح مجموعه داده‌های در دسترس باعث گردد، اطلاعاتی که در رابطه با میزان اهمیت یک جمله بدست می‌آید، قابل تعمیم بر روی داده‌های آزمایشی نباشد. دلیل دیگر را می‌توان عدم‌هماهنگی معماری شبکه پیشنهادی با تغییر در تعداد ویژگی‌های ورودی دانست. در واقع با توجه به مساله تنک بودن داده‌ها و معماری شبکه، سامانه‌ی پیشنهادی در حالتی که از ۴ ویژگی شامل میزان تکرار کلمات، میزان تشابه جملات با عنوان سند، موقعیت قرارگیری جمله در سند و امتیاز مربوط به برچسب‌گذاری اجزای کلام استفاده می‌کند، بهترین عملکرد را از خود نشان داده است.

جدول ۲ مقادیر معیارهای ارزیابی مربوط به مقایسه‌ی جمله‌ای در حالتی که شبکه از ۴ ویژگی استفاده کرده است را نشان می‌دهد. با توجه به ماهیت چندسندی خلاصه‌ساز پیشنهادی، ممکن است در اسناد مختلف یک مجموعه، جملاتی وجود داشته باشند که در عین حالی‌که حامل بارمعنایی مشابه‌ای هستند، اما از کلمات و شیوه‌ی نگارش متفاوتی بهره برده‌اند. در ارزیابی‌های صورت گرفته بر روی خروجی خلاصه‌ساز، به جملاتی برخوردیم که از نظر مفهومی مشابه جملاتی در خلاصه‌ی تولید شده توسط انسان بودند، اما از آنجایی که معیار ارزیابی ما جملات کاملاً یکسان بود لذا جملات با مفهوم مشترک را در ارزیابی شرکت ندادیم هرچند که با احتساب اینگونه جملات می‌توان دقت خلاصه‌ساز را بالاتر از مقادیر موجود در جدول ۲ در نظر گرفت. با توجه به جدول ۲ می‌توان اینگونه

برداشت کرد که خروجی سامانه پیشنهادی، در قیاس جمله‌ای و در حالت میانگین نزدیک به ۵۰ درصد مشابه خلاصه‌ی انسانی موجود در پیکره است.



شکل ۲- ارزیابی خلاصه‌ساز با تعداد ویژگی‌های ورودی مختلف

جدول ۲- نتایج حاصل از ارزیابی جمله‌ای

معیارهای ارزیابی	Recall	Precision	F Score
مقادیر بدست آمده از مقایسه‌ی جمله‌ای	۰/۴۶۶۷	۰/۳۸۸۹	۰/۴۲۴۳

یک بسته‌ی نرم‌افزاری استاندارد برای ارزیابی سیستم‌های خلاصه‌سازی خودکار متون و ترجمه‌ی ماشینی تحت عنوان ROUGE موجود است. این نرم‌فزار به مقایسه‌ی خلاصه‌های تولید شده توسط سیستم‌های مختلف با یکدیگر و همچنین خلاصه‌های تولید شده توسط انسان می‌پردازد [۲۲]. در این پژوهش، توسط نرم‌افزار ROUGE، خلاصه‌ی تولید شده توسط سیستم خلاصه‌ساز پیشنهادی را با خلاصه‌های انسانی موجود در پیکره پاسخ و براساس unigramها و bigramهای مشابه مورد ارزیابی قرار دادیم. در واقع میزان کلمه‌ها و جفت کلمه‌های مشابه مابین خلاصه‌های تولید شده توسط سامانه و خلاصه‌های ایده‌آل مدنظر قرار می‌گیرند. لازم به‌ذکر است که این ارزیابی‌ها با حذف کلمات ایست و نادیده گرفتن آن‌ها انجام شده است. نتایج بدست آمده از این ارزیابی در جدول ۳ قابل مشاهده است.

جدول ۳- نتایج حاصل از ارزیابی توسط نرم‌افزار ROUGE

معیارهای ارزیابی	Recall	Precision	F Score
مقایسه براساس unigramهای مشابه	۰/۶۸۵۰	۰/۳۷۲۶	۰/۴۸۲۷
مقایسه براساس bigramهای مشابه	۰/۵۱۲۷	۰/۲۷۲۸	۰/۳۵۶۲

۵- نتیجه‌گیری

در این مقاله طراحی یک سیستم خلاصه‌ساز استخراجی چندسندی متون فارسی، با بهره‌گیری از پیکره‌ی پاسخ تشریح گشت. ۴ ویژگی تاثیرگذار در تصمیم‌گیری در رابطه با میزان اهمیت یک جمله از متن، برای تمامی جملات موجود در پیکره‌ی آموزشی استخراج شد. برای امتیازدهی به جملات براساس ویژگی‌های استخراج شده، از روشی تحت عنوان یادگیری عمیق استفاده کردیم. در این روش با استفاده از یک شبکه‌ی عصبی مصنوعی به‌نام Autoencoder و تزریق بردار ویژگی‌ها به‌عنوان ورودی به آن، به آموزش سیستم برای امتیازدهی به جملات پرداختیم. در فاز نهایی برای تولید متن خلاصه، جملات براساس امتیازهای بدست آمده، مرتب می‌شوند و جملاتی که بالاترین امتیاز را دارا هستند، تشکیل متن خلاصه را می‌دهند.

متن خلاصه‌ی تولید شده توسط سیستم پیشنهادی با خلاصه تولید شده توسط انسان مورد مقایسه و ارزیابی قرار گرفت. این ارزیابی به ۳ شکل، ارزیابی جمله‌ای، ارزیابی براساس unigramهای مشابه و ارزیابی براساس bigramهای مشابه صورت گرفت. نتایج حاصل از این ارزیابی‌ها نشان از عملکرد خوب سامانه‌ی پیشنهادی داشته و توانایی آن را در تولید یک خلاصه‌ی قابل‌قبول با بهره‌گیری از تعداد محدودی ویژگی ورودی را بازگو می‌نماید. در واقع یکی از مزایای استفاده از یادگیری عمیق در این سامانه، کاهش صرف وقت و هزینه در تولید و طراحی ویژگی‌ها و معیارهای اهمیت متن است، زیرا شبکه‌ی پیشنهادی به لطف استفاده از یادگیری عمیق قابلیت تولید ویژگی‌های جدید از ویژگی‌های محدود ورودی را دارد.

مراجع

[1] D. Das, and A. Martins, "A Survey on Automatic Text Summarization," Literature Survey for the Language and Statistics II Course at Carnegie Mellon University, 2007.

[2] D. Radev, and T. Allison, "MEAD – A platform for multidocument multilingual text summarization," Proceedings ofthe 4th International Conference on Language Resources and Evaluation (LREC), pp. 699-702, 2004.

[3] A. Thiago, and P. Salgueiro, "GistSumm: A Summarization Tool Based on a New Extractive Method," Proceedings of 6th International Conference on Computational, pp. 210-218, 2012.

[4] M. Hassel, and N. Mazdak, "FarsiSum - A Persian text summarizer," Proceedings of the Workshop On Computational Approaches to Arabic Script-Based Languages, pp. 82-84, 2004.

[۵] ز. کریمی، و م. شمس‌فرد، "خلاصه‌سازی متون فارسی،" در مجموعه مقالات یازدهمین کنفرانس سالانه انجمن کامپیوتر ایران، ص ۱۲۹۴-۱۲۸۶، ۱۳۸۴.

[6] M. A. Honarpisheh, Gh. Ghassem-Sani, and G. Mirroshandel, "A Multi-Document Multi-Lingual Automatic Summarization System," Proceedings of the Joint Conference on Natural Language Processing, pp. 733-738, 2008.

[7] F. Kiyoumarsi, and F. Rhimi Esfahani, "Optimizing Persian Text Summarization Based on Fuzzy Logic Approach," Proceedings of theInternational Conference on Intelligent Building and Management, pp. 264-269, 2011.

[8] G. Hinton, O. Simon, and T. Yee-Whye, "A Fast Learning Algorithm for Deep Belief Nets," Neural Computation, vol. 18, pp. 1527-1554, 2006.

[9] L. Arnold, S. Rebecchi, S. Chevallier, and H. Paugam-Moisy, "An Introduction to Deep Learning," Proceedings of theEuropean Symposium on Artificial Neural Networks-Computational Intelligence and Machine Learning, pp. 477-488, 2011.

[10] Y. Bengio, "Learning Deep Architectures for AI," Foundations and Trends in Machine Learning, vol. 2, no. 1, pp. 1-127, 2009.

[11] G. Hinton, L. Deng, D. Yu, G. Dahl, and A. Mohamed, "Deep Neural Networks for Acoustic Modeling in Speech Recognition," IEEE Signal Processing Magazine, vol. 29, no. 6, pp. 82-97, 2012.

[12] R. Collobert, and J. Weston, "A Unified Architecture for Natural Language Processing: Deep Neural Networks with Multitask Learning," Proceedings of theInternational Conference on Machine Learning, pp. 160-167, 2008.

[13] R. Collobert, J. Weston, L. Bottou, M. Karlen, M. Kavukcuoglu, and P. Kuksa, "Natural Language Processing (almost) from Scratch," Journal of the Machine Learning Research, vol. 12, pp. 2493-2537, 2011.

[14] E. Arısoy, T. Sainath, B. Kingsbury, and B. Ramabhadran, "Deep Neural Network Language Models," Proceedings of the NAACL-HLT, pp. 20-28, 2012.

[15] Y. Liu, S. Zhong, and W. Li, "Query-Oriented Multi-Document Summarization via Unsupervised Deep Learning," Proceedings of the 26th AAAI Conference on Artificial Intelligence, pp. 1699-1705, 2012.

[16] A. Fischer, and Ch. Igel, "An Introduction to Restricted Boltzmann Machines," Proceedings of the 17th Iberoamerican Congress on Pattern Recognition, pp. 14-36, 2012.

[17] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, and P. Manzagol, "Stacked DenoisingAutoencoders: Learning Useful Representations in a Deep Network with a Local Denoising Criterion," Journal of the Machine Learning Research, vol. 11, no.11, pp. 3371-3408, 2010.

[18] N. Prabhakar, and N.Chandra, "Automatic Text Summarizaion Based on Pragmatic Analysis," International Journal of the Scientific and Research Publications, vol. 2, Issue 5, pp. 1-4, 2012.

[19] R. Mihalcea, and P. Tarau, "TextRank: Bringing Order into Texts,"Proceedings of the Conference on Empirical Methods in Natural Language Processing, pp. 404-411, 2004.

[20] B. Behmadi Moghaddas, M. Kahani, S. A. Toosi, A. Pourmasoumi, and A. Estiri, "Pasokh: A standard corpus for the evaluation of Persian text summarizers," Proceedingsof the Computer and Knowledge Engineering (ICCKE), pp. 471-475, 2013.

[۲۱] الف. پورمعصومی، م. کاهانی، الف. طوسی، الف. استیری، و ه. قائمی "ایجاز: یک سامانه‌ی عملیاتی برای خلاصه‌سازی تک‌سندی متون خبری فارسی،" دوفصلنامه‌ی پردازش علائم و داده‌ها، شماره ۱، پیاپی ۲۱، ص۳۳-۴۸، ۱۳۹۳.

[22] Ch. Lin, "Rouge: A package for automatic evaluation of summaries," Proceedings of the ACL workshop on Text Summarization Branches Out, pp. 74-81, 2004.

شیما محرابی فارغ‌التحصیل ازموسسه آموزش عالی طبرستان در مقطع کارشناسی و دانشگاه گیلان در مقطع کارشناسی‌ارشد در رشته مهندسی کامپیوتر گرایش نرم‌افزار. زمینه‌های تحقیقاتی مورد علاقه یادگیری ماشین، یادگیری عمیق، پردازش زبان‌های طبیعی، خلاصه‌سازی خودکار متون هستند.

آدرس پست‌الکترونیکی ایشان عبارت است از:

shima.mehrabi85@gmail.com

حمیدرضا احمدی‌فر فارغ‌التحصیل از دانشگاه شهید بهشتی در مقاطع کارشناسی و دکتری و دانشگاه صنعتی امیرکبیر در مقطع کارشناسی‌ارشد در رشته مهندسی کامپیوتر با گرایش معماری سیستم‌های کامپیوتری. از سال ۱۳۸۲ عضو هیات علمی گروه مهندسی کامپیوتر دانشگاه گیلان بوده و زمینه‌های مورد علاقه حساب کامپیوتری، سیستم‌های توزیع شده، رایانش ابری و بکارگیری شبکه‌های عصبی در حل مسائل مختلف هستند.

آدرس پست‌الکترونیکی ایشان عبارت است از:

ahmadifar@guilan.ac.ir

سید ابوالقاسم میرروشندل فارغ‌التحصیل از دانشکده فنی دانشگاه تهران در مقطع کارشناسی در رشته مهندسی کامپیوتر با گرایش نرم‌افزار و دانشگاه صنعتی شریف در مقاطع کارشناسی‌ارشد و دکتری در رشته مهندسی کامپیوتر گرایش هوش مصنوعی. از سال ۱۳۹۱ عضو هیات علمی گروه مهندسی کامپیوتر دانشگاه گیلان بوده و زمینه‌های مورد علاقه ایشان پردازش زبان‌های طبیعی، داده‌کاوی، یادگیری ماشینی و پردازش تصویر هستند.

آدرس پست‌الکترونیکی ایشان عبارت است از:

mirroshandel@guilan.ac.ir

اطلاعات بررسی مقاله:

تاریخ ارسال: ۱۳۹۵/۰۷/۱۱

تاریخ اصلاح: ۱۳۹۵/۱۰/۰۲

تاریخ قبول شدن: ۱۳۹۵/۱۰/۲۵

نویسنده مرتبط: دکتر حمیدرضا احمدی‌فر، دانشکده فنی، دانشگاه گیلان، رشت، ایران.

[1] Text Summarization
[2] Extractive Summarization
[3] Abstractive Summarization
[4] Support Vector Regression
[5] Deep Learning
[6] Name Entity Recognition
[7] Part Of Speech Tagging
[8] Semantic Role Labeling
[9] Chunking

[10]Language Modeling
[11]Restricted Boltzmann Machine
[12]Feed-Forward Neural Networks
[13]Unsupervised Layer-Wise Pre-Training
[14]Deep Belief Networks
[15]Stop Words
[16]Data Sparseness

علوم رایانش و فناوری اطلاعات
نشریه علمی انجمن کامپیوتر ایران
مجلد ۱۴، شماره ۲، ۱۳۹۵
صفحات ۴۷- ۵۷
مقاله عادی

یک سیستم توصیه‌گر در بستر تجارت اجتماعی برای صنعت گردشگری: مبتنی بر شباهت، جوامع اجتماعی، اعتماد و شهرت

لیلا اسماعیلی سید علیرضا هاشمی گلپایگانی زینب زنگنه مدار

دانشکده مهندسی کامپیوتر و فناوری اطلاعات، دانشگاه صنعتی امیرکبیر، تهران، ایران

چکیده

اینترنت و سرویس‌های مبتنی بر آن، به طور قابل توجهی کسب و کارهای مختلف از جمله صنعت گردشگری را تحت تاثیر قرار داده و تنوع بسیاری در سرویس‌ها و محصولات آن فراهم آورده‌اند. با افزایش چشمگیر تعداد انتخاب‌ها در بسته‌های سفر، هتل‌ها، جاذبه‌های گردشگری و غیره، پیدا کردن آن چه که گردشگر بدان نیاز دارد، بسیار دشوار شده است. به همین دلیل، سیستم‌های توصیه‌گر گردشگری مورد توجه محققان و کسب و کارها قرار گرفته‌اند. جاذبه‌های گردشگری، اغلب دلیل تمایل افراد به سفر و گردشگری هستند. این تحقیق، یک سیستم توصیه‌گر اجتماعی- ترکیبی را در بستر تجارت اجتماعی پیشنهاد می‌دهد که می‌تواند یک فهرست شخصی‌سازی شده از جاذبه‌های گردشگری برای هر گردشگر، مبتنی بر تشابه تمایلات و علایق کاربران، اعتماد، شهرت، روابط و جوامع اجتماعی ایجاد کند. در مقایسه با روش‌های قدیمی پالایش مشارکتی و مبتنی بر محتوی و ترکیبی، مزیت روش پیشنهاد شده جامعیت به‌کارگیری از فاکتورهای مختلف و لحاظ کردن فاکتور اعتماد در منابع توصیه مانند شناسایی رتبه‌دهی‌های برون هشت می‌باشد. نتایج حاصل از آزمایش‌ها برتری روش پیشنهادی نسبت به سایر روش‌های رایج را تایید می‌کند؛ مدل پیشنهادی، می‌تواند در توصیه سایر محصولات و سرویس‌ها در صنعت گردشگری و دیگر کسب و کارهای اجتماعی بکار گرفته شود.

کلمات کلیدی: تجارت اجتماعی، سیستم توصیه‌گر، اعتماد، شباهت، جوامع، شهرت، صنعت گردشگری، روابط اجتماعی.

۱- مقدمه

صنعت گردشگری با داشتن حدود ۱۱٪ تولید ناخالص داخلی[۱] (GDP) جهانی و به‌کارگیری ۲۰۰ میلیون نفر و به خدمت گرفتن ۷۰۰ میلیون گردشگر در سراسر جهان، یکی از بزرگترین بخش‌های تولیدی در جهان محسوب می‌شود، که انتظار می‌رود در سال ۲۰۲۰، این آمار دو برابر شود [۱]. پیشرفت فناوری اطلاعات و ارتباطات، اینترنت و سرویس‌های مبتنی بر آن ضمن فراهم کردن امکان دسترسی کاربر به اطلاعات دقیق‌تر و بیشتر، تعداد انتخاب‌ها را به طور چشمگیری افزایش داده است؛ و این امر موجب دشواری یافتن آن‌چه مصرف‌کنندگان به آن نیاز دارند شده است. با توجه به پژوهش‌ها [۲]، از جمله دلایل عدم‌خرید برخط محصولات گردشگری توسط مشتریان، فقدان خدمات شخصی‌سازی شده، مسائل امنیتی، فقدان تجربه و وقت‌گیر بودن جستجوی آن می‌باشد.

سیستم‌های توصیه‌گر، به کاربران در یافتن گزینه‌های جذاب، مورد نیاز و مناسب از بین گزینه‌های بسیار کمک می‌کنند. هدف اصلی این سیستم‌ها برآورد تمایل کاربر و فراهم کردن پیش‌بینی از اطلاعات موجود است. سه رویکرد پالایش مشارکتی، پالایش مبتنی بر محتوی و ترکیبی از جمله مرسوم‌ترین روش‌های توصیه محسوب می‌شوند [۳].

از سوی دیگر، توسعه سریع رسانه‌های اجتماعی و وب ۲.۰، پتانسیل عظیمی را برای تبدیل تجارت الکترونیکی از یک محیط محصول محور به یک محیط اجتماعی و مشتری محور فراهم کرده است [۴]. در این محیط، دسترسی مشتریان به دانش و تجارب اجتماعی موجب فهم بهتر اهداف خرید برخط و تصمیم‌گیری دقیق‌تر و با آگاهی بیش‌تر می‌شود [۵]. چرا که برای اغلب مردم، خرید، یک تجربه اجتماعی است و به همین دلیل پیش از خرید، علاقه‌مند به آگاهی از نظر دوستان و آشنایاشان هستند. همچنین Gretzel و Yoo [۶] عنوان کرده‌اند که سه چهارم از مسافران، بررسی‌های مصرف‌کنندگان برخط را به عنوان یک منبع اطلاعاتی

هنگام برنامه‌ریزی سفرهای خود در نظر می‌گیرند. از این رو، تجارت‌الکترونیکی با پذیرش قابلیت‌ها، عملکردها و مشخصه‌های وب ۲.۰ به منظور مشارکت مشتری و تشویق روابط مشتریان، با یک تحول جدید روبرو شده است که آن را تجارت اجتماعی می‌نامند [۷] [۸] و ارزش اقتصادی بیشتری را برای کسب و کارها داراست [۹].

با توجه به بستر فراهم شده در تجارت اجتماعی، یک رویکرد جدید در سیستم‌های توصیه‌گر مطرح می‌شود که کمتر مورد توجه قرار گرفته است و آن استفاده از روابط اجتماعی افراد به عنوان یک منبع اطلاعاتی اضافی می‌باشد. براساس اصل هوموفیلی [۱۰] در حوزه شبکه‌های اجتماعی، شباهت موجب ایجاد ارتباط می‌شود؛ به بیان دیگر، کاربران با افرادی که با آن‌ها در ارتباط هستند، صفات مشترکی دارند [۳۶]. از سوی دیگر چگالی ارتباط کاربران در کل شبکه حاصل، یکسان نمی‌باشد؛ این تفاوت موجب تعریف ویژگی اجتماع در یک ساختار شبکه‌ای می‌شود. اجتماعات بخش‌هایی از شبکه هستند که دارای چگالی ارتباطی زیاد در داخل و چگالی ارتباطی کم با خارج از خود می‌باشند. گاهی جوامع می‌توانند برخی از ویژگی‌های افراد را بدون بررسی اطلاعات فردی هر کاربر آشکار سازند [۳۷]. اغلب سیستم‌های توصیه‌گر تجاری و گردشگری از اطلاعات جمعیت‌شناختی کاربران استفاده می‌کنند؛ حال آن که افراد به دلیل شباهت با سایر افرادی که در تعامل هستند، تحت تاثیر نظر دوستان، افراد خبره و افراد با علایق مشترک قرار می‌گیرند و این امر می‌تواند توسعه یک سیستم توصیه‌گر اجتماعی مبتنی بر روابط افراد و جوامع اجتماعی را معقول و عملی کند.

دو دسته کلی برای وب‌سایت‌های تجارت اجتماعی وجود دارد: یک دسته مبتنی بر سایت‌های تجارت الکترونیکی است که ابزارها و مفاهیم وب ۲.۰ را به کار گرفته تا یک سایت تجارت اجتماعی را توسعه دهد (مانند www.amazon.com)؛ و دسته دیگر براساس یک پلتفرم وب ۲.۰ ساخته شده‌اند و ویژگی‌های تجارت الکترونیکی را نیز اضافه کرده‌اند (مانند www.facebook.com/Starbucks) [۱۱]. در دسته اول، به قابلیت‌های اجتماعی مانند اشتراک‌گذاری مطالب، ارتباط کاربران با یکدیگر و .... کمتر پرداخته شده؛ در مقابل، دسته دوم قابلیت‌های خرید/فروش را کمتر مورد توجه قرار داده است؛ به گونه‌ای که فاقد تاریخچه خرید کاربر و قیمت‌گذاری‌ها می‌باشند. از این رو، سیستم‌های توصیه‌گر، در وب‌سایت‌های تجارت اجتماعی تنها یک گروه از قابلیت‌ها (فاکتورهای اجتماعی یا تاریخچه خرید کاربر) را درنظر گرفته‌اند.

به منظور پرداختن به این مسائل، در این تحقیق یک وب‌سایت ارائه‌دهنده خدمات گردشگری که تواماً دارای قابلیت‌های شبکه اجتماعی و تجارت الکترونیکی می‌باشد در نظر گرفته شده است و یک روش توصیه اجتماعی- ترکیبی مبتنی بر شباهت، اعتماد، شهرت و جوامع اجتماعی جهت افزایش دقت و اعتماد در پیش‌بینی توصیه ارائه می‌شود. این توصیه‌گر به پیشنهاد جاذبه‌های گردشگری به کاربر جهت سفر می‌پردازد؛ بسیاری از محققان اظهار داشته‌اند که جاذبه‌های گردشگری، اغلب دلیل تمایل افراد به سفر و گردشگری هستند [۱۲] [۱۳]. فاکتورهای تعاملی و روابط انسانی (مانند اعتماد [۱۴] [۱۵]، شهرت [۱۶] [۱۷]، روابط اجتماعی [۱۸]، جوامع اجتماعی [۱۹]) تاکنون در سیستم‌های توصیه‌گر گردشگری و به طور خاص توصیه جاذبه‌های گردشگری مورد استفاده قرار نگرفته است و در زمینه‌های کاربردی مختلف دیگر نیز به طور جداگانه به کار رفته‌اند و همزمان استفاده نشده‌اند.

روش پیشنهادی اجازه می‌دهد تا جاذبه‌های گردشگری مناسب و قابل اعتماد برای هر کاربر گردشگر با استفاده از هوش جمعی از شبکه اجتماعی کاربران شناسایی شود. مدل پیشنهادی می‌تواند به‌طور عملی علاوه بر بکار رفتن در وب‌سایت‌های اجتماعی گردشگری در سایر وب‌سایت‌ها و پلتفرم‌های تجارت اجتماعی نیز به کار رود.

ادامه مقاله به این صورت سازماندهی شده است. بخش ۲ به مرور ادبیات مرتبط با موضوع می‌پردازد. در بخش ۳ مدل توصیه‌گر اجتماعی- ترکیبی بحث می‌شود. مجموعه داده، تنظیمات و استراتژی توصیه در بخش ۴ توضیح داده می‌شود. بخش ۵ به بحث در خصوص نتایج حاصل از آزمایش‌ها و ارزیابی اختصاص یافته است. نتیجه‌گیری تحقیق و ارائه پیشنهادهایی برای تحقیقات آتی در بخش آخر آمده است.

۲- کارهای مرتبط

۲-۱- سیستم‌های توصیه‌گر جاذبه گردشگری

سیستم‌های توصیه‌گر، به کاربران در یافتن گزینه‌های جذاب، مورد نیاز و مناسب از بین گزینه‌های بسیار کمک می‌کنند. این سیستم‌ها برای حل مشکل سربار اطلاعاتی، به وجود آمده [۲۰] و موجب رشد فروش در تجارت الکترونیکی شده‌اند. هدف اصلی این سیستم‌ها برآورد تمایل کاربر و فراهم کردن فهرست اقلام پیش‌بینی شده از اطلاعات مناسب است [۳].

روش‌های توصیه‌گر مختلفی توسعه یافته‌اند که مرسوم‌ترین آن‌ها عبارتند از پالایش مبتنی بر محتوا، پالایش مشارکتی و روش‌های توصیه ترکیبی [۲۱]. ایده اصلی در روش اول، انتخاب و پیشنهاد قلم‌هایی است که به آنچه کاربر هدف در گذشته خریداری کرده است، شبیه باشد. الگوریتم مبتنی بر محتوا برای انتخاب چنین قلم‌هایی، شباهت بین قلم‌ها را با استفاده از تحلیل محتوای اطلاعات متنی آن‌ها اندازه‌گیری می‌کند؛ با اینحال، مشکل اصلی این روش زمانی است که اطلاعات کافی و مناسب جمع‌آوری و فراهم نشود ،بنابراین سیستم توصیه‌گر با شکست مواجه می‌شود [۲۲]. پالایش مشارکتی یکی از موفق‌ترین الگوریتم‌های توصیه در دامنه‌های تجاری بوده است. این روش کاربران مشابه را شناسایی کرده (مثلا براساس ماتریس رتبه‌دهی کاربر- قلم داده) و تمایلات آن‌ها را برای ساخت توصیه‌ها تحلیل می‌کند. شروع سرد و قلم داده جدید، دو مشکل اصلی این روش هستند. مطالعات بسیاری برای ترکیب موازی یا متوالی سیستم‌های پالایش مشارکتی و پالایش مبتنی بر محتوا به منظور بهبود سیستم‌های توصیه‌گر انجام شده است [۳].

اکتساب اطلاعات برای فراهم کردن توصیه‌های شخصی شده متناسب با تمایل کاربر به دو روش صریح و ضمنی انجام می‌شود [۳۱]. در روش ضمنی رفتار کاربر به منظور اطلاع از تمایلاتش جمع‌آوری می‌شود. زمانی که تغییری در رفتار شناسایی شود، داده مربوط به تمایل کاربر نیز به طور همزمان تغییر می‌کند [۳۲]. روش صریح، تعاملات و بازخوردهای کاربر را پالایش و تحلیل می‌کند تا بتوان ویژگی‌های کاربر و تمایلاتش را تعیین کرد [۳۳].

سیستم‌های توصیه‌گر جاذبه گردشگری مبتنی بر روش‌های پالایش مشارکتی [۲۳] [۲۶] و ترکیبی [۲۴] [۲۵] هستند. همچنین عمدتاً با استفاده همزمان از روش‌های ضمنی و صریح، تمایلات کاربر هدف را شناسایی می‌کنند [۲۴] [۲۶]. Schiaffino و Amandi [۲۵] از اطلاعات جمعیت‌شناختی استفاده کردند تا در روش پالایش مشارکتی، کاربران مشابه را شناسایی کنند. آنها با استفاده از شبکه‌های معنایی، عامل‌ها و ترکیب روش پالایش مشارکتی و پالایش مبتنی بر محتوا به توصیه تورها و جاذبه‌های گردشگری پرداختند. معیارهای توصیه آنها نوع جاذبه، قیمت، موقعیت و زمان سفر است. Huang and Bian [۲۶] ایز از ترکیب روش AHP با شبکه‌های بیزین برای توصیه جاذبه‌های گردشگری استفاده کردند. معیارهای توصیه در این پژوهش نوع جاذبه، فاصله و قیمت است. در روش پیشنهادی .Carcia-Cerspo et al [۲۶] که مبتنی بر پالایش مشارکتی است، علایق، تمایلات کاربر و رتبه‌دهی وی به جاذبه‌های گردشگری به‌صورت صریح در

ابتدا دریافت می‌شود و بعد از آن سایر اطلاعات ضمنی می‌تواند از شبکه اجتماعی که کاربر عضو آن است و رفتار و یاستخراج شود. معیار توصیه در این تحقیق موقعیت، زمان، آب و هوا است. Yang و Marques [۲۳] ضمن کسب اطلاعات به روش ضمنی، روشی مبتنی بر پالایش مشارکتی جهت توصیه جاذبه‌های گردشگری پیشنهاد کرده‌اند.

در مقاله فعلی، ما قصد داریم تا یک روش توصیه‌گر ترکیبی- اجتماعی را مبتنی بر تئوری گراف و با ترکیب روش‌های پالایش مشارکتی و مبتنی بر محتوا ارائه دهیم. این مدل ضمن اکتساب ضمنی و صریح اطلاعات با به‌کارگیری فاکتورهای اعتماد، شهرت، جوامع اجتماعی حاصل از روابط و شباهت کاربران با یک‌دیگر، اقلام مناسب (جاذبه‌های گردشگری) را به کاربر توصیه می‌کند. این روش براساس معیارهایی متفاوت با روش‌های پیشین فهرست توصیه‌ها را ایجاد می‌کند. جزییات مدل در بخش ۳ آمده است.

۲-۲- اعتماد و شهرت در تجارت اجتماعی

جوامع برخط یکی از ویژگی‌های تجارت اجتماعی است که به کاربران اجازه می‌دهد تا به راحتی تمایلات شخصی خود را بیان کنند. همچنین افراد می‌توانند مشخص کنند به کدام کاربران اعتماد دارند یا به کدام محصول/ سرویس علاقه‌مند هستند. در بسیاری از معاملات برخط، خریدار از فروشنده جدید یا از کالای جدید فروشنده‌ای که قبلا از آن خریدی داشته است، شناخت کافی ندارد. این مساله باعث می‌شود خرید کالا با ریسک همراه باشد. اگر عاملی وجود داشته باشد که تضمین‌کننده اعتبار فروشنده باشد، ریسک خرید از وی تا حد زیادی پایین می‌آید [۳۴] [۳۵]. اعتماد و شهرت از جمله‌ی عواملی هستند که می‌توان با استفاده از آن‌ها ریسک خرید را کاهش داد [۱۷] [۲۷]. ریسک کلی یک تراکنش، تابعی از متغیرهای اعتماد مانند هزینه تراکنش، تاریخچه تراکنش‌ها و جبران خسارت است [۲۸].

- **اعتماد:** یک کمیت ذهنی و تعیین‌کننده میزان انتظارات یک فرد از اعمال دیگران است که این کمیت بر نوع رفتار آن فرد در هنگام تعامل با دیگران تأثیرگذار است [۳۵].
- **شهرت:** یک کمیت عمومی و اجتماعی است که براساس نوع رفتار هر فرد در تعاملات قبلی خود با سایر افراد جامعه محاسبه می‌شود [۳۵].

بنابراین تفاوت اصلی بین سیستم‌های اعتماد و شهرت را می‌توان بدین صورت توضیح داد: سیستم‌های اعتماد امتیازی را تولید می‌کنند که بازتاب‌کننده دید شخصی هر فرد از میزان اطمینان به فرد دیگر است. در حالی که سیستم‌های شهرت، قابلیت اطمینان هر فرد را از دیدگاه کل اجتماع محاسبه می‌کند. علاوه بر این، ورودی سیستم‌های اعتماد تنها معیارهای ذهنی و کلی است، در حالی‌که اطلاعات موجود در مورد تراکنش‌ها (مانند رتبه‌ها و نظرات) ورودی سیستم‌های شهرت هستند [۲۷]. امروزه روش‌های متعددی برای محاسبه اعتماد و شهرت ارائه شده است که بعضی از آن‌ها در سیستم‌های تجاری استفاده می‌شوند ولی بعضی هنوز در حد ایده‌های پیشنهادی هستند [۱۷] [۲۷].

در این تحقیق، ما اعتماد یک فرد در شبکه را ساخته شده یا فراهم شده از شهرت وی در نظر می‌گیریم.

۲-۳- جوامع اجتماعی

مجموعه داده‌های زیادی می‌توانند مبتنی بر ساختار شبکه توصیف شوند که در توصیف مبتنی بر تئوری گراف، گره‌ها بازنمای موجودیت‌ها (مانند افراد) و یال‌ها بازنمای روابط بین گره‌ها (مانند رابطه دوستی یا همکلاسی) می‌باشد. یک ویژگی مشترک برای همه این شبکه‌ها، ساختار انجمن یا جامعه است که به گروهی از گره‌ها با چگالی اتصال زیاد اشاره دارد. این گروه از گره‌ها دارای ارتباطات تنک و کم با سایر گروه‌ها هستند. کاوش جوامع با تمرکز بر کشف و توصیف چنین ساختارهای شبکه‌ای، در چند سال اخیر بسیار مورد توجه قرار گرفته است [۲۹]. روش‌ها و رویکردهای بسیاری برای کشف جوامع وجود دارد که می‌توان به روش‌های مبتنی بر مرکزیت، روش‌های بهبود پیمانه‌ای، روش‌های محلی، روش‌های بخش‌بندی طیفی و غیره اشاره کرد؛ [۳۰] مرور کاملی بر این روش‌ها دارد.

افراد دارای ویژگی‌های مختلف هستند که به واسطه آن در طبقات توصیفی- کیفیتی بسیار متنوعی قرار می‌گیرند؛ و می‌توان برخی از ویژگی‌ها را در برخی از طبقات به صورت بارزتری دید. به عنوان مثال زنان عاطفی و افراد تحصیل کرده آزادمنش هستند.چنین ویژگی‌های ذاتی موجب نادیده گرفتن تنوع بسیار طبقات توصیفی در جوامع اجتماعی می‌شود. از این رو افراد به طور قابل توجهی با افرادی در ارتباط هستند که شبیه خودشان می‌باشند؛ در نتیجه برخی از صفات در بخش‌هایی از اجتماع محلی و متمرکز می‌شود. این اصل هوموفیلی نام دارد و به موجب آن ارتباط بین افراد مشابه، با نرخ بیشتری در مقایسه با افراد نامشابه شکل می‌گیرد [۱۰].

از اصل هوموفیلی و همچنین ساختارهای جوامع می‌توان چنین نتیجه گرفت: جوامع اجتماعی که دارای ارتباطات قوی گره‌ها در داخل و ارتباطات ضعیف با سایر جوامع هستند، متشکل از گره‌هایی می‌باشند که این گره‌ها شباهت‌های قابل‌قبولی با یکدیگر دارند. این مقاله با در نظر گرفتن چنین اصلی سعی در شناسایی جوامع به منظور ارائه مدل توصیه‌گر خود دارد.

۳- مدل سیستم توصیه‌گر ترکیبی- اجتماعی

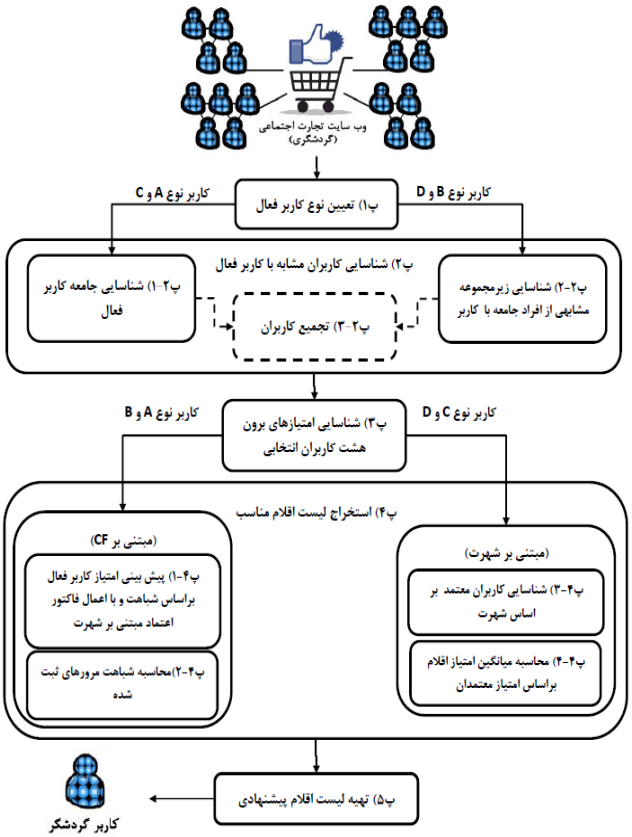
در صنعت گردشگری، تصمیم گردشگران برای سفر تحت تاثیر جاذبه‌های گردشگری است. گردشگران برای انتخاب مقصد سفر، تمایل به سوال و دریافت پیشنهاد از دوستان نزدیک، افرادی که با آنها هم علاقه‌اند یا گردشگران حرفه‌ای دارند. گردشگران حرفه‌ای حداقل یک بار به مقصد مورد نظر رفته‌اند و تجارب سفر خود را می‌توانند از طریق ثبت نظر یا review به دیگران منتقل کنند. با این حال، دوستان نزدیک ممکن است تجربه کافی یا علاقه مشترک نداشته باشند. همچنین ممکن است توصیه‌های گردشگران حرفه‌ای که آشنا نیستند غیرقابل باور باشد. به هر حال زمانی که مقاصد سفر متنوع هستند، منابع مشورتی نیز متفاوت خواهند بود؛ و در نهایت با تجمیع نظر همه منابع گردشگر تصمیم می‌گیرد. بنابراین یک توصیه‌گر جاذبه گردشگری موثر بهتر است این موارد را به‌طور مناسبی ترکیب کند. در این مطالعه، یک سیستم توصیه‌گر اجتماعی- ترکیبی برای جاذبه‌های گردشگری پیشنهاد شده است که به طور موثری از روابط افراد و نظرات گردشگران به جاذبه‌های گردشگری استفاده می‌کند تا یک فهرست توصیه شخصی‌سازی شده برای هر کاربر فعال مبتنی بر اعتماد، شهرت و شباهت ایجاد نماید. کاربر فعال، کاربری است که سیستم توصیه‌گر فهرستی از جاذبه‌های گردشگری را به وی توصیه می‌کند.

معماری سیستم توصیه‌گر پیشنهادی در شکل ۱ آمده است. پنج مولفه برای تحلیل داده‌های موجود در بستر تجارت اجتماعی حوزه گردشگری توسعه یافته است. اهداف هر یک از مولفه‌های مدل در ادامه توضیح داده می‌شود:

- مولفه تعیین نوع کاربر فعال، براساس اطلاعات صریح و ضمنی متفاوتی که از کاربران و جاذبه‌های گردشگری در سیستم است (پیوست الف)، کاربر فعال را به یکی از انواع جدول ۱ تخصیص می‌دهد.

- مولفه شناسایی کاربران مشابه با کاربر فعال، زیرمجموعه‌ای از کاربران را براساس اجتماعی که وی در آن قرار دارد یا کاربرانی که شباهت جمعیت-شناختی دارند استخراج می‌کند.
- مولفه شناسایی امتیازهای برون هشت، نظراتی که توسط کاربران با دانش کم یا کاربران جعلی برای جاذبه‌های گردشگری ثبت شده است را تعیین می‌کند.
- مولفه استخراج فهرست اقلام مناسب، رتبه هر جاذبه گردشگری را برای هر کاربر فعال با توجه به نوع وی محاسبه می‌کند.
- مولفه تهیه فهرست اقلام پیشنهادی، فهرست منتخبی از جاذبه‌های گردشگری مناسب قابل ارائه به کاربر فعال را براساس رتبه‌های آن‌ها ایجاد می‌کند.

برای هر کاربر (گردشگر) بازدیدکننده از وب‌سایت تجارت اجتماعی، سیستم می‌تواند فهرستی از جاذبه‌های گردشگری را مبتنی بر نظر کاربران قابل اعتماد و نظرات مشابه با کاربر فعال فراهم نماید. در ادامه پردازش‌های اصلی سیستم توصیه‌گر پیشنهادی توضیح داده می‌شود.



شکل ۱- مدل سیستم توصیه‌گر ترکیبی- اجتماعی پیشنهادی (کمان‌ها و مستطیل خط چین، الزامی نبودن را مشخص می‌کند)

۳-۱- تعیین نوع کاربر فعال

مولفه تعیین نوع کاربر فعال (u_a) براساس وجود یا عدم‌وجود رابطه بین کاربر فعال و دیگر اعضای بستر تجارت اجتماعی، همچنین تعداد نظرات ثبت شده توسط کاربر فعال برای جاذبه‌های گردشگری، نوع کاربر را تعیین می‌کند. در نتیجه سیستم می‌تواند در ادامه مناسب‌ترین منابع اطلاعاتی را برای روش توصیه استفاده نماید.

جدول ۱- انواع کاربر فعال با توجه به ارتباطات و نظرات

نظر / ارتباط	کاربر دارای ارتباط است	کاربر دارای ارتباط نیست
کاربر نظر ثبت شده دارد	نوع A	نوع B
کاربر نظر ثبت شده ندارد	نوع C	نوع D

۳-۲- شناسایی کاربران مشابه با کاربر فعال

افراد با تمایل‌ها یا رفتارهای مشابه به اقلام یکسانی علاقه‌مند هستند، حتی اگر همدیگر را نشناسند [۳۸] [۳۹]. مبتنی بر فعالیت‌های کاربر در یک زمینه اجتماعی خاص، می‌توان کاربرانی که در یک سطح از شباهت هستند را شناسایی کرد. میزان علاقه‌مندی گردشگر فعال برای یک جاذبه گردشگری را می‌توان بر اساس گروهی از گردشگران که تمایلات مشابه با وی دارند پیش‌بینی کرد. با توجه به نوع کاربر، کاربران مشابه با وی از دو روش شناسایی می‌شوند: ۱) مبتنی بر اصلی هوموفیلی و هم اجتماع بودن کاربران و ۲) مبتنی برتشابه ویژگی‌های جمعیت‌شناختی کاربران.

اگر کاربر نوع A یا C باشد می‌توان زیرمجموعه‌ای از گراف روابط کاربران را که کاربر فعال بدان تعلق دارد را با استفاده از یکی از روش‌های شناسایی و کشف اجتماعات استخراج کرد (پ ۲-۱، شکل ۱). افرادیکه در یک اجتماع هستند براساس اصل هوموفیلی دارای تمایلات و ویژگی‌های مشترک قابل قبولی می‌باشند. در رابطه ۱، U مجموعه کاربران سیستم تجارت اجتماعی است؛ و c_{u_i} اجتماع یا گروهی است که کاربر i به آن تعلق دارد. UC_i مجموعه کاربران هم‌اجتماع با کاربر i را نشان می‌دهد، به‌طوری‌که هر کاربر j که هم‌گروه با کاربر i باشد در این مجموعه قرار می‌گیرد.

$$UC_i = \left\{ UC_i \subseteq U, u_j \in UC_i \mid c_{u_j} = c_{u_i} \right\} \quad (۱)$$

اگر کاربر فعال با دیگر کاربران رابطه نداشته باشد (نوع B و D)، در این صورت براساس اطلاعات جمعیت‌شناختی وی،گروهی از کاربران به عنوان کاربران مشابه باکاربر فعال مشخص می‌شوند (پ۲-۲، شکل ۱). در رابطه ۲، F_{u_i} مجموعه ویژگی‌های جمعیت‌شناختی کاربر i را نشان می‌دهد که شامل n ویژگی مختلف است. UD_i مجموعه کاربرانی است که دارای ویژگی‌های مشترک جمعیت‌شناختی با کاربر i می‌باشند.

$$F_{u_i} = \{f_1, \ldots, f_n\}$$
$$UD_i = \left\{ UD_i \subseteq U, u_j \in UD_i \mid \exists F_1 = F_{u_i} \cap F_{u_j}, \mid F_1 \mid = m, m \le n \right\} \quad (۲)$$

به‌طور کلی کاربر در یکی از چهار گروه A، B، C و D قرار می‌گیرد و براساس رابطه ۱ یا ۲ کاربران مشابه با وی مشخص می‌شوند. شرایطی وجود دارد که ممکن است اندازه (جمعیت) گروه کاربران مشابه با کاربر فعال برای تصمیم‌گیری و پیشنهاد کافی نباشد. در چنین شرایطی از حالت‌های جایگزین استفاده می‌شود. رابطه ۳ نحوه تعیین نهایی گروه کاربران مشابه با کاربر فعال $(US_i)i$ را در حالات مختلف نشان می‌دهد. براساس رابطه ۳، اگر تعداد کاربران مجموعه UC_i به سرحد مورد نظر، w، برسد، این مجموعه به‌عنوان گروه نهایی کاربر فعال انتخاب می‌شود. در صورتی‌که تعداد کاربران مجموعه UC_i از سرحد مورد نظرکمتر باشد، از

مجموعه UD_i کاربرانی که دارای نظر ثبت شده می‌باشند به تصادف انتخاب می‌شوند و به جمعیت گروه کاربر فعال افزوده می‌شوند. برای کاربران نوع B و D که قطعا امکان تشکیل مجموعه UC_i را ندارند، مجموعه UD_i تشکیل می‌شود. درصورتی‌که اندازه مجموعه UD_i از سرحد تعریف شده، w، کمتر باشد، تعداد ویژگی‌های جمعیت‌شناختی مشترک، m، کاهش می‌یابد تا جمعیت موردنظر حاصل شود. بنابراین مجموعه کاربران انتخابی US_i مستخرج از پ۲ در شکل ۱، مبتنی بر رابطه ۳ خواهد بود.

$$US_i = \begin{cases} UC_i & n_{UC} \geq w \\ UD_i & n_{UD} \geq w \\ UC_i \cup \{Ud_i \subseteq UD_i\} & n_{UC} < w \\ UD_i, m\_new < m\_old & n_{UD} < w \end{cases} \quad (۳)$$

۳-۳- شناسایی امتیازهای برون‌هشت کاربران

در سیستم‌های نظردهی اجتماعی به دلایل مختلفی ممکن است نظرات فاقد ارزش و اعتبار باشند: کاربر جعلی است، کاربر فاقد دانش لازم است یا کاربر بنا بر دلایل مختلفی نظر غیرواقعی ثبت می‌کند. چنین کاربران و چنین نظراتی می‌بایست در منبع اطلاعاتی توصیه‌گر شناسایی شوند و برخورد مناسبی با آنها انجام گیرد. بنابراین ابتدا ماتریس کاربر- جاذبه گردشگری UPoA به ازای هر نوع جاذبه CP تشکیل می‌شود. هر جاذبه گردشگری (P) براساس ماهیتش به گروه مشخصی (CP) تعلق دارد. هر درایه ماتریس UPoA برابر با امتیازی است که کاربر i به جاذبه j داده است. به ازای هر جاذبه گردشگری در این ماتریس، با استفاده از روش شناسایی برون‌هشت آماری (توزیع نرمال) [۲۱] امتیاز برون‌هشت شناسایی می‌شود. در نهایت، ماتریس UO ساخته می‌شود که هر درایه آن تعداد امتیازهای برون‌هشت کاربر i در جاذبه‌های گروه k است.

$$CP = \{cp_1, ..., cp_k\} \quad P = \{p_1, ..., p_j\}$$
$$UPoA_k = \left[UP_{u_i p_j} \right], i \leq n_u, j \leq n_p, k \leq n_{cp}, cat_{p_j} = k \quad (۴)$$
$$UO = \left[UO_{u_i cp_k} \right], i \leq n_u, k \leq n_{cp}$$

$$R\_Re = \left[\text{Re}_{u_i cp_k} * fall\_factor_{u_i cp_k} \right]$$
$$\text{Re}_{u_i cp_k} = \frac{num\_of\_reviews\_for\_u_i\_to\_cp_k}{total\_num\_of\_reviews\_for\_cp_k} \quad (۵)$$
$$fall\_factor_{u_i cp_k} = \sigma^{UO_{u_i cp_k}}, 0 \leq \sigma \leq 1$$

تعداد امتیازهای برون‌هشت هر کاربر در شهرت وی تاثیر خواهد داشت. شهرت کاربر در هر گروه از جاذبه‌های گردشگری، $\text{Re}_{u_i cp_k}$، به‌صورت جداگانه محاسبه می‌شود و مقدار آن برابر است با نسبت تعداد نظرهای ثبت شده کاربر i در مورد جاذبه‌های گردشگری گروه k، به کل تعداد نظرات آن گروه از جاذبه‌های گردشگری. زمانی که کاربر امتیازهای برون‌هشت داشته باشد، با یک فاکتور کاهش، fall_factor شهرت وی کاهش می‌یابد. ماتریس R_Re ماتریس شهرت واقعی هر کاربر در هر گروه از جاذبه‌های گردشگری است (رابطه ۵).

۴-۳- استخراج لیست اقلام مناسب

کاربران نوع A و B: این دو گروه، کاربرانی هستند که نظرات گذشته آن‌ها در سیستم موجود می‌باشد. روش استخراج لیست اقلام برای این کاربران مبتنی بر پالایش مشارکتی است؛ اما به دلیل آن‌که نظر کاربر تنها شامل یک امتیاز عددی نیست و برداری از پاسخ به سوالات را نیز شامل می‌شود دارای شرایطی می‌باشد. ابتدا امتیاز کاربر فعال u_a به جاذبه‌های گردشگری در هر گروه که برای آن تاکنون نظری ثبت نکرده است با در نظر گرفتن معیار شهرت تخمین زده می‌شود (P_Ra(a, j)). سپس شباهت هر پاسخ برای جاذبه گردشگری i و j، $sim(r_{p_i}, r_{p_j})$، به عنوان وزنی برای امتیازهای پیش‌بینی شده در مرحله قبل لحاظ می‌شود؛ در نهایت P_P(a, i)، نظر نهایی کاربر فعال a را برای جاذبه گردشگری i پیش‌بینی می‌کند (رابطه ۶).

$$R\_Ra = \left[Ra_{u_i p_j} * R\_\text{Re}_{ik} \right], cat_{p_j} = k$$
$$\overline{R\_Ra_i} = \frac{1}{P_i} \sum_{j \in P_i} R\_Ra_{ij}, \forall i \in U$$
$$Similarity(a, i) = \frac{\sum_{j \in P_i \cap P_a} (R\_Ra_{aj} - \overline{R\_Ra_a})(R\_Ra_{ij} - \overline{R\_Ra_i})}{\sqrt{\sum_{j \in P_i \cap P_a} (R\_Ra_{aj} - \overline{R\_Ra_a})^2} \sqrt{\sum_{j \in P_i \cap P_a} (R\_Ra_{ij} - \overline{R\_Ra_i})^2}} \quad (۶)$$
$$P\_Ra(a, j) = \overline{R\_Ra_i} + \frac{\sum_{i \in U} simlarity(a, i)(R\_Ra_{ij} - \overline{R\_Ra_i})}{\sum_{i \in U} | simlarity(a, i) |}$$
$$sim(r_{p_i}, r_{p_j}) = \cos(\overrightarrow{r_{p_i}}, \overrightarrow{r_{p_j}})$$
$$P\_P(a, j) = \frac{\sum_{all\_similar\_reviews,N} sim(r_{p_j}, N) * P\_Ra(a, N)}{\sum_{all\_similar\_reviews,N} (| sim(r_{p_j}, N) |)}$$

کاربران نوع C و D: چنانچه کاربر از نوع C و D باشد، سابقه‌ای از نظرات وی موجود نیست؛ بنابراین استخراج لیست اقلام مناسب، مبتنی بر سابقه نظرات دیگر کاربران خواهد بود. بر این اساس جاذبه‌های گردشگری در هر گروه با متوسط امتیاز بیشتر در بین کاربران معتمد TU، شناسایی می‌شود. کاربران معتمد کاربرانی هستند که معیار شهرت آن‌ها R_Re_{ik}، از مقدار سرحد مشخص شده α، بیش‌تر باشد (رابطه ۷).

$$TU = \{u_i \in U \mid R\_\text{Re}_{ik} \geq \alpha\}$$
$$\overline{R\_Ra_{p_j}} = \frac{\sum_{i \in TU} R\_Ra_{ij}}{| TU |} \quad (۷)$$

۵-۳- تهیه لیست اقلام پیشنهادی

امتیاز نهایی پیش‌بینی شده برای جاذبه‌های گردشگری در گروه کاربران A و B از مرحله پ۴-۱ و پ۴-۲ (شکل ۱) بدست می‌آید. با مرتب‌سازی این امتیازها، امکان انتخاب n- بالاترین قلم با امتیاز بیشتر برای سیستم وجود دارد. مقدار n با توجه به شرایط می‌تواند متغیر باشد. در صورتی‌که کاربر از گروه C و D شناسایی شود،

n- بالاترین قلم استخراج شده از مرحله پ۴-۳ و پ۴-۴ در شکل ۱، که دارای امتیاز بیشتری از سایر اقلام باشند انتخاب و به کاربر پیشنهاد می‌شوند.

۴- آزمایش‌ها

در این بخش، یک مطالعه تجربی براساس مدل توصیه‌گر اجتماعی- ترکیبی پیشنهادی انجام می‌شود. به منظور ارزیابی نتایج حاصل از روش پیشنهادی، کارایی آن با سایر رویکردهای رایج و قدیمی توصیه‌گری نیز مقایسه می‌شود.

۴-۱- مجموعه داده

آزمایش‌های ما بر روی مجموعه داده‌ای جمع‌آوری شده از یک آژانس گردشگری بین‌المللی انجام شد، که به دلیل حفظ محرمانگی امکان انتشار آن وجود ندارد. جزئیات مجموعه داده در ادامه آمده است.

- پروفایل کاربران (Users): شامل اطلاعات جمعیت‌شناختی ۵۷۸۴۷ کاربر از جمله جنسیت، زبان، کشور و غیره (تنها ۱۴۸۷ نفر در نظردهی مشارکت داشته‌اند) است.
- روابط دوستی کاربران (Friendships): شامل ۲۱۱۹۶۳ رابطه دوستی صریح می‌باشد، بنابراین شمار بسیاری از کاربران شبکه فاقد روابط دوستی هستند.
- اطلاعات جاذبه‌های گردشگری (Attractions): شامل تعداد ۲۷۸۱ جاذبه گردشگری در ۲۷ دسته مختلف
- نظرات کاربران (Reviews): شامل ۳۰۲۷ نظر که توسط کاربران برای جاذبه‌های گردشگری ثبت شده است. آژانس مسافربری به منظور غنای بیشتر و جمع‌آوری نظرات، طی برنامه‌های تبلیغاتی کاربران و گردشگران را به ثبت نظر تشویق نمود. با این حال، در این میان برخی از جاذبه‌های گردشگری فاقد نظر هستند.

۴-۲- استراتژی‌های توصیه

در این تحقیق، رویکرد پیشنهادی (SociHeybrid Rec) با دو رویکرد اصلی دیگر به منظور ارزیابی کارایی مقایسه می‌شود، با توجه به روش پیشنهادی و مجموعه دادگان سایر روش‌ها قابل مقایسه با روش پیشنهادی نبودند. دو رویکرد مورد استفاده در این آزمایش‌ها عبارتند از:

- مبتنی بر محتوا (Rec M1): در این استراتژی، از روش معرفی شده در [۴۳] استفاده می‌شود. به‌طور خلاصه با استفاده از مشخصات اقلامی که قبلا توسط کاربر فعال امتیازدهی شده‌اند، یک پروفایل از علاقه‌مندی‌های وی ایجاد و اقلام جدید مشابه و مطابق با علاقه‌مندی کاربر به او پیشنهاد می‌شود.
- پالایش مشارکتی (Rec M2): روش پالایش مشارکتی معرفی شده توسط [۴۴] می‌باشد. کاربرانی که امتیازهای آن‌ها به اقلام، مشابه با کاربر فعال است شناسایی و با استفاده از امتیازهای آنها، امتیاز کاربر فعال به اقلام پیش‌بینی می‌شود.

۵- نتایج و ارزیابی

به منظور ارزیابی و مقایسه کارایی استراتژی‌های مختلف توصیه، ما به طور تصادفی ۳۰۲۷ رکورد مجموعه نظرات کاربران را مبتنی بر قانون پرتو [۴۰] به ۸۰٪ مجموعه یادگیری و ۲۰٪ مجموعه تست تقسیم کردیم. آزمایش‌ها ۱۰ بار انجام شد؛ در ادامه میانگین نتایج ارزیابی به ازای مجموعه تست آمده است.

۵-۱- ارزیابی امتیازهای پیش‌بینی شده

مقایسه امتیاز واقعی ثبت شده توسط کاربران برای جاذبه‌های گردشگری با امتیاز تخمینی پیش‌بینی شده توسط استراتژی‌های مختلف، به عنوان معیاری برای ارزیابی به شمار می‌رود. این مقایسه می‌تواند به دو صورت انجام شود: اول مقایسه الگوی رفتاری نمودارهای امتیازها برای استراتژی‌های مختلف و دوم مقایسه میزان خطای امتیاز پیش‌بینی شده با امتیاز واقعی (معیار MSE [۴۱]). MSE میانگین مربعات خطا را محاسبه می‌کند.

با توجه به تعدد آزمایش‌ها برخی یا بخشی از نمودارها ارائه می‌شود. شکل ۲، رفتار نمودار امتیازهای استراتژی‌های توصیه را در مقایسه با رفتار نمودار امتیازهای واقعی در یکی از آزمایش‌ها نشان می‌دهد. دو نمودار به ازای ۵۰ جاذبه گردشگری است که به تصادف انتخاب شده‌اند. در شکل ۲ محور افقی کاربران و محور عمودی میانگین امتیازهای هر کاربربه جاذبه‌های گردشگری می‌باشد که در بازه [۰، ۵] قرار دارد. صرف‌نظر از اختلاف امتیازها در روش‌های توصیه Rec M1، Rec M2 و SociHybrid Rec با امتیاز واقعی، رفتار نمودار امتیازهای پیش‌بینی شده با استفاده از استراتژی SociHybrid Rec بسیار مشابه با رفتار امتیازهای واقعی است. نقاط بیشینه و کمینه نمودارهای امتیازهای واقعی و مبتنی بر استراتژی SociHybrid Rec بسیار به یکدیگر نزدیک است.

مقدار معیار MSE به ازای هر سه استراتژی در جدول ۲ آمده است. هر چه مقدار MSE کمتر باشد روش پیش‌بینی یا تخمین‌گر کارایی و دقت بیشتری دارد. براساس نتایج، روش پیشنهادی امتیازهای جاذبه‌های گردشگری را با شباهت بیشتر و اختلاف کمتری نسبت به امتیازهای واقعی و در مقایسه با دیگر روش‌ها تخمین می‌زند.

جدول ۲- نتایج معیار MSE به ازای استراتژی‌های مختلف توصیه

روش پیشنهادی	پالایش مشارکتی	مبتنی بر محتوا
0.78	1.6	2.22

شکل ۲- مقایسه رفتار امتیازهای جاذبه‌های گردشگری، به ازای ۵۰ جاذبه گردشگری به تصادف انتخاب شده

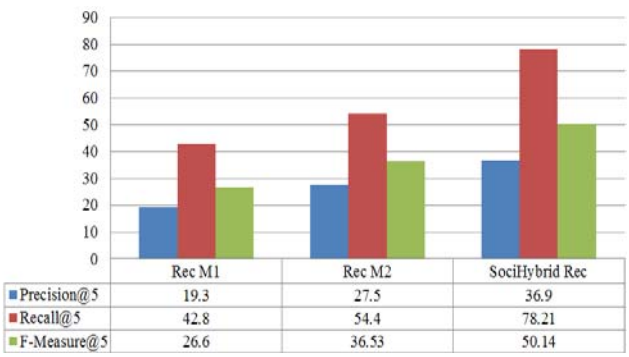
۵-۲- ارزیابی توصیه‌ها

مقایسه و ارزیابی استراتژی‌های توصیه مبتنی بر سه معیار رایج دقت، بازخوانی و F-Measure انجام می‌شود [۳] [۲۱]. معیار دقت، مشخص‌کننده نسبت تعداد توصیه‌های مطلوب به تعداد کل توصیه‌های ارائه شده به کاربر فعال است. معیار بازخوانی، نسبت تعداد توصیه‌های مطلوب به تعداد اقلام مطلوب کاربر می‌باشد. F-Measure نیز میانگین وزن‌دار معیار دقت و بازخوانی است که از نسبت حاصل-ضرب دو معیار دقت و بازخوانی به حاصل جمع این دو معیار بدست می‌آید. مقدار هر سه معیار در بازه [۰، ۱] است و هر چه به ۱ نزدیک‌تر باشد، مطلوب‌تر است. شکل ۳ نتایج این سه معیار را برای سه استراتژی توصیه، به صورت درصد نشان می‌دهد.

بعد از پیش‌بینی امتیازهای کاربر فعال به اقلامی که تاکنون امتیازدهی نکرده است، به روش n- بالاترین، امتیازها به ترتیب نزولی مرتب شده و n جاذبه گردشگری با امتیاز بیشتر به کاربر فعال توصیه می‌شود. طبق شکل ۳، سه معیار ارزیابی برتری روش پیشنهادی را نسبت به روش پالایش مشارکتی (Rec M2) و مبتنی بر محتوا (Rec M1) در حالتی که n=5 است را نشان می‌دهند. روش پالایش مشارکتی (Rec M2) نیز نسبت به روش مبتنی بر محتوا (Rec M1) نتایج بهتری را تولید کرده است.

اگر چه روش پیشنهادی نسبت به دو روش دیگر از نظر محاسباتی پیچیده‌تر است اما برتری نتایج ارزیابی، تاثیر مثبت افراد مشابه با کاربر فعال مبتنی بر اصل هوموفیلی، شناسایی امتیازهای برون‌هشت و بکارگیری اعتماد و شهرت در سیستم توصیه‌گر را مشخص می‌کند. زمانی که خرید به دلایل مختلف همراه با ریسک است، توصیه‌گری مبتنی بر روش پیشنهادی قابل اعتمادتر بوده و بر تصمیم کاربر برای خرید تاثیرگذارتر است.

بیشتر بودن معیار بازخوانی نسبت به معیار دقت، به دلیل تعداد کمتر جاذبه‌های گردشگری نظردهی شده توسط کاربر (اقلام مطلوب) به تعداد جاذبه‌های توصیه شده به وی می‌باشد. براساس مجموعه داده، کاربران به‌طور میانگین برای ۲ جاذبه گردشگری نظر ثبت کرده‌اند، اما فهرست پیشنهادی جاذبه‌های گردشگری شامل ۵ جاذبه گردشگری است (@5). از این‌رو در همه استراتژی‌ها معیار بازخوانی بیشتر از معیار دقت می‌باشد. همچنین معیار دقت کمتر ۴۰٪ است، در نگاه اول این مقدار به نظر کم است ولی این مورد نیز قابل توجیه می‌باشد. از آنجا که به طور میانگین تعداد جاذبه‌های گردشگری نظر داده شده توسط کاربران ۲ است، در بهترین حالت اگر این ۲ جاذبه عضو مجموعه اقلام توصیه شده به وی باشند، مقدار دقت ۴۰٪ خواهد بود. بنابراین مقدار ۳۵.۱٪ برای معیار دقت در روش پیشنهادی در مقایسه با بیشترین مقدار ممکن ۴۰٪ مطلوب است. از این‌رو، مقادیر بدست آمده برای معیارهای ارزیابی برتری روش پیشنهادی نسبت به روش‌های پالایش مشارکتی و مبتنی بر محتوا را نشان می‌دهد.

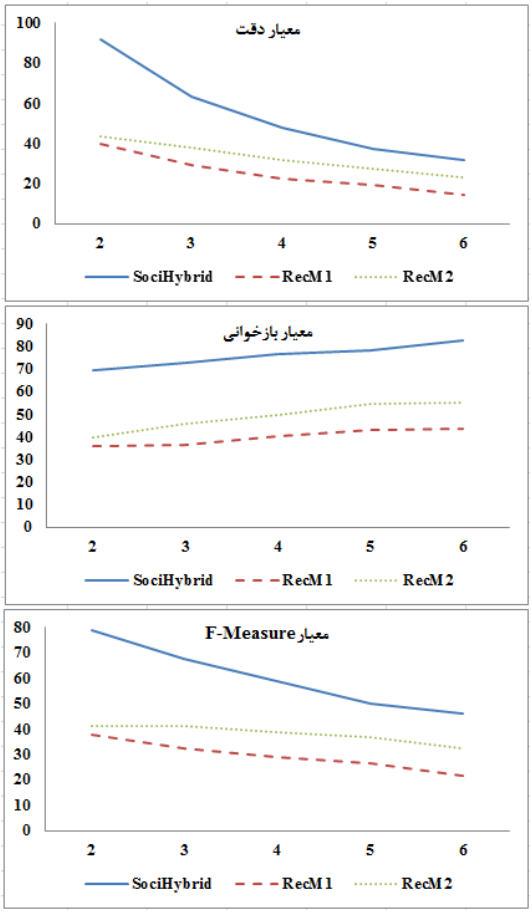


شکل ۳- مقایسه معیارهای ارزیابی برای سه استراتژی توصیه

نتایج معیارهای ارزیابی به ازای فهرست پیشنهادی (n) با تعداد ۲ الی ۶ جاذبه گردشگری نیز در جدول ۳ و شکل ۴ آمده است. به دلیل انتخاب اقلام فهرست پیشنهادی به روش n- بالاترین و وجود اقلام موردنظر کاربر در بالای فهرست پیشنهاد شده به وی، تغییرات معیار بازخوانی در سه روش نسبت به تغییرات معیار دقت کمتر است، و معیار دقت با کاهش اندازه فهرست پیشنهادی افزایش چشم گیری دارد. علت این امر نیز آن است که تعداد اقلام یا جاذبه‌های مطلوب کاربران به‌طور متوسط ۲ مورد می‌باشد. بنابراین با افزایش اندازه فهرست پیشنهادی، معیار دقت کاهش می‌یابد و از مقدار مشخصی نمی‌تواند بیشتر شود.

جدول ۳- نتایج معیارهای ارزیابی برای سه استراتژی توصیه به ازای فهرست پیشنهادی با اندازه‌های مختلف

#N	Method	Precision	Recall	F-Measure
@2	SociHybrid Rec	91.64	69.35	78.95
	Rec M1	39.7	35.8	37.65
	Rec M2	43.2	39.6	41.32
@3	SociHybrid Rec	63.51	72.8	67.84
	Rec M1	29.2	36.3	32.37
	Rec M2	37.87	45.5	41.34
@4	SociHybrid Rec	47.65	76.56	58.74
	Rec M1	22.34	40.4	28.77
	Rec M2	31.6	49.34	38.53
@5	SociHybrid Rec	36.9	78.21	50.14
	Rec M1	19.3	42.8	26.6
	Rec M2	27.5	54.4	36.53
@6	SociHybrid Rec	31.81	82.43	45.91
	Rec M1	14.5	43.21	21.73
	Rec M2	22.7	55.3	32.19



شکل ۴- مقایسه معیارهای ارزیابی برای سه استراتژی توصیه به ازای فهرست پیشنهادی با اندازه‌های مختلف

۶- نتیجه‌گیری

در خرید برخط، مردم تحت تاثیر پیشنهادها و راهنمایی‌های افراد مشابه، خریداران خبره، کارشناس یا دوستان صمیمی خود هستند. با این حال هنوز بسیاری از پلتفرم‌های شبکه اجتماعی مانند توئیتر و فیسبوک، همچنین پلتفرم‌های تجارت الکترونیکی مانند آمازون و یاهو به صورت مستقل عمل می‌کنند. امروزه شبکه‌های اجتماعی فراتر از کاربردشان در جهت برقراری ارتباط افراد با یک‌دیگر می‌باشند؛ تجارت الکترونیکی تحت تاثیر قابلیت‌ها و پتانسیل اطلاعاتی موجود در این شبکه‌ها با نسل جدیدی روبرو شده است. تجارت اجتماعی به‌طور فزاینده‌ای از نظر عملی و علمی مورد توجه قرار گرفته و کسب و کارها، سیاست‌گذاری‌ها و سرمایه‌گذاری‌های خود را به این سمت متمایل کرده‌اند. از این رو توسعه سیستم‌های توصیه‌گر مانند آمازون و eBay که مبتنی بر تاریخچه خرید فردی، تجمیع نظرات اعضا و بازخوردها می‌باشد [۴۲]، دیگر اثرگذار و کافی نخواهد بود. این‌ها ارتباطات بین اعضا و قدرت تاثیر اجتماعی را در نظر نمی‌گیرند. به‌منظور در نظر گرفتن اثرات اجتماعی افراد و ایجاد توازن بین عوامل مختلف، در این مقاله یک سیستم توصیه‌گر اجتماعی- ترکیبی پیشنهاد شد که شباهت افراد در جوامع حاصل از ساختار اجتماعی، همراه با تشابه نظرات افراد، به علاوه اعتماد حاصل از شهرت را به منظور ارائه توصیه به کاربران در بستر تجارت اجتماعی به کار می‌گیرد. نتایج آزمایش‌ها بر روی مجموعه داده‌ای در صنعت گردشگری و با هدف توصیه جاذبه‌های گردشگری، کارایی روش توصیه پیشنهادی ما را نسبت به سایر رویکردهای اصلی نشان می‌دهد. چارچوب توصیه‌گری پیشنهاد شده می‌تواند به‌طور موثری توسط فروشندگان الکترونیکی در بستر تجارت اجتماعی به منظور ترغیب کاربران به خرید استفاده شود.

۶-۱- نوآوری تحقیق

نوآوری‌ها و کاربردهای این مقاله در ادامه خلاصه می‌شود:

- از دیدگاه نوآوری سیستمی، همچنان که توصیه اقلام در تجارت الکترونیکی روبه گسترش است، طراحی سیستم‌های توصیه‌گر اجتماعی هنوز به‌عنوان یک مسئله باقیمانده‌اند و از طرفی توسعه روزافزون تجارت اجتماعی و فناوری‌های اینترنت اشیاء رسیدگی و حل این مسئله را بیشتر از قبل ضروری نموده است.
- از منظر روش‌شناسی، ما نه تنها شباهت افراد از لحاظ ویژگی‌های جمعیت‌شناختی را در نظر گرفته‌ایم بلکه شباهت حاصل از هوموفیلی و هم‌اجتماع بودن افراد را نیز که حاصل از ساختار اجتماعی می‌باشد لحاظ کرده‌ایم. به‌علاوه اعتماد مبتنی بر شهرت کاربران را نیز در توصیه‌ها نقش داده‌ایم. بنابراین منبع اطلاعاتی توصیه، افراد مشابه با کاربر و نظرات قابل اعتماد هستند. از سوی دیگر در بیش‌تر سیستم‌های تجارت الکترونیکی نظر کاربر به یک قلم داده ترکیبی از متن، یک رتبه کلی یا چند رتبه برای جوانب مختلف کالا می‌باشد. بدلیل زمان‌بر بودن پردازش متن، عموما توضیحات نظرات کاربر در سیستم‌های توصیه‌گر استفاده نمی‌شود و توصیه صرفا براساس رتبه‌ها صورت می‌گیرد. حال آن‌که در توصیه‌گر پیشنهادی، هر نظر علاوه بر یک امتیاز کلی، دارای برداری از مقادیر اسمی می‌باشد که در پیش‌بینی رتبه کاربر فعال به اقلام پیشنهادی لحاظ شده است.
- از منظر کارایی نیز دقت، بازخوانی و F-Measure روش پیشنهادی در مقایسه با دیگر روش‌ها بهتر است. بنابراین روش ما اطلاعات قابل اعتماد مرتبط با سلیقه و خواست کاربران را بازیابی می‌کند. همچنین انتخاب زیرمجموعه‌ای از کاربران مشابه با کاربر هدف، کاهش بار محاسباتی را همراه دارد.
- از دیدگاه کاربردی نیز با توجه به گسترش روزافزون تجارت اجتماعی، مکانیزم پیشنهادی می‌تواند در سایر زمینه‌های تجاری مانند فروشگاه‌های الکترونیکی و صنایع مختلف دیگر استفاده شود. در مواردی که به دلیل قیمت بالای کالا/خدمات، ریسک خرید برای کاربر زیاد است روش پیشنهادی می‌تواند توصیه‌ها قابل اعتمادتری را نسبت به سایر روش‌ها ارائه کند.

۶-۲- محدودیت‌ها

چندین محدودیت در این تحقیق وجود دارد که عبارتند از:

- اگر چه خدمات تجارت اجتماعی افزایش یافته است و سایت‌های تجارت‌الکترونیکی تا حدودی به این امکانات مجهز شده‌اند اما بیشتر به عنوان اعتبار ظاهری به‌کار می‌رود؛ و صاحبان پلتفرم‌های تجارت اجتماعی به دلیل حفظ محرمانگی، اطلاعات خود را به اشتراک نمی‌گذارند. مجموعه داده ما نیز از یک وب‌سایت نه چندان مشهور گردشگری بدست آمد که تعداد نظرات ثبت شده و کاربران نظردهنده در آن با توجه به تعداد کل کاربران بسیار محدود بود. بنابراین چنانچه کسب وکارها بتوانند کاربران را تشویق به مشارکت در ارائه نظرات نمایند، این محدودیت از بین رفته و داده غنی‌تری را می‌توان برای سیستم توصیه‌گراستفاده نمود.
- امکان دریافت بازخود از کاربران در تحقیق حاضر وجود نداشت. در صورتیکه امکان دریافت بازخورد از کاربران وجود داشت، ارزیابی توصیه برای کاربران نوع C و D نیز انجام می‌گرفت؛ با توجه به ویژگی‌های مدل، در این صورت نتایج ارزیابی مدل پیشنهادی در مقایسه با رویکردهای دیگر تفاوت بیشتری داشت و تعداد اقلام مطلوب فعلی کاربران که از مجموعه داده قابل استخراج است، بر روی مقدار معیارهای ارزیابی محدودیت ایجاد نمی‌کرد.
- مدل پیشنهادی بر روی مجموعه داده کم حجمی اجرا شد، از آنجا که در واقعیت و در بستر تجارت اجتماعی تعداد کاربران و اقلام بیشتر است و تراکنش‌های بیشتری اتفاق می‌افتد، می‌بایست آزمایش‌هایی بر روی حجم بیشتر داده نیز انجام شود و روش پیشنهادی برای کاربرد در حجم بالا بهینه شود و از پردازش‌های موازی بهره گرفته شود.

۶-۳- مطالعات آینده

برخی از جهت‌گیری‌ها برای مطالعات آینده به شرح زیر می‌باشد:

- تقویت ساختار روابط دوستی کاربران از طریق استخراج روابط اجتماعی از دیگر شبکه‌های اجتماعی به منظور شناسایی کاربران مشابه
- تعمیم روش پیشنهادی جهت ارائه یک سیستم توصیه‌گر آگاه به زمینه
- بهبود و ارتقای روش پیشنهادی به منظور رفع مشکل تنکی داده؛ روش‌های پالایش مشارکتی نسبت به تنکی داده حساس هستند و در چنین شرایطی نتایج خوبی حاصل نمی‌شود. با توجه به آنکه روش پیشنهادی از پالایش مشارکتی استفاده می‌کند، این مشکل در سیستم بروز خواهد کرد.
- بررسی تاثیر دیگر روش‌های کشف اجتماع بر نتایج روش پیشنهادی و استفاده از دیگر روابط به‌منظور محاسبه اعتماد، شهرت و شباهت؛ بدین‌منظور آزمایش‌های بیشتری مورد نیاز است تا بهترین روابط و روش کشف اجتماع برای سیستم توصیه‌گر پیشنهادی تعیین شود.

پیوست الف

در وب‌سایت تجارت اجتماعی گردشگری، ضمن دسترسی به اطلاعات ارتباطی کاربران، اطلاعات سفر و خرید تورهای کاربران نیز موجود می‌باشد. با توجه به آن‌که جاذبه گردشگری مهمترین محرک و انگیزه برای سفر شناخته شده است، سیستم به توصیه جاذبه گردشگری می‌پردازد تا از این طریق کاربر اقدام به خرید تورهای مرتبط نماید. اطلاعات صریح و ضمنی متفاوتی از کاربران و جاذبه‌های گردشگری در سیستم موجود می‌باشد که برخی از اطلاعات در چارچوب توصیه‌گر پیشنهادی استفاده می‌شود و در جدول ۴ آمده است. اطلاعات ضمنی نیازمند به محاسبه و پردازش براساس سایر اطلاعات هستند. همچنین هر کاربر می‌تواند یک نظر یا review به ازای هر جاذبه گردشگری در سیستم ثبت نماید که شامل یک امتیاز کلی، مشخصات کاربر (جدول ۴) و سوالاتی در رابطه با جاذبه گردشگری با پاسخ‌های مشخص می‌باشد. عناصر اطلاعاتی یک نظر در جدول ۵ مشخص شده است.

جدول ۴- اطلاعات کاربر و جاذبه گردشگری در سیستم گردشگری اجتماعی (■ در سیستم توصیه‌گر استفاده می‌شود و □ استفاده نمی‌شود)

مشخصات کاربر (وضعیت استفاده: عنوان: نوع داده: وضعیت)	مشخصات جاذبه گردشگری (وضعیت استفاده: عنوان: نوع داده: وضعیت)
■: جنسیت: اسمی: صریح	■: رتبه: نسبی: ضمنی
■: زبان: اسمی: صریح	■: نوع جاذبه: اسمی: صریح
■: کشور: اسمی: صریح	□: محل: اسمی: صریح
□: بازه سنی: فاصله‌ای: ضمنی	□: وضعیت پرداخت: اسمی: صریح
□: تحصیلات: ترتیبی: صریح	□: امکانات: اسمی: صریح
□: شغل: اسمی: صریح	□: معرفی: متن: صریح
□: وضعیت تاهل: اسمی: صریح	□: شرایط خاص: متن: صریح
□: شهرت: نسبی: ضمنی	

جدول ۵- عناصر اطلاعاتی یک نظر

مشخصات کاربر	جدول ۴
رتبه	نوع داده: ترتیبی؛ وضعیت: صریح
سوال (عنوان سوال: نوع جواب: وضعیت جواب)	هدف سفر: اسمی: صریح همراهان: اسمی: صریح نوع تور: اسمی: صریح فصل سفر: اسمی: صریح روش سفر: اسمی: صریح

مراجع

[1] K. Kabassi, "Personalizing recommendations for tourists," Telematics and Informatics, Elsevier, 27, pp. 51–66, 2010.

[2] K. Wolfe, C. H. C. Hsu, and S. K. Kang, "Buyer characteristics among users of various travel intermediaries," Journal of Travel and Tourism Marketing, vol. 17, no. 2/3, pp. 51–62, 2004.

[3] L. Esmaeili, M. Nasiri, and B. Minaei-Bidgoli, "Personalizing group recommendation to social network users," Web Information Systems and Mining. Springer Berlin Heidelberg, pp. 124-133, 2011.

[4] R. T. Wigand, R. I. Benjamin, and J. Birkland, "Web 2.0 and beyond: implications for electronic commerce," In Proceedings of the 10th International Conference on Electronic Commerce, Austria, ACMPress, New York, 2008.

[5] G. Dennison, S. Bourdage-Braun, and M. Chetuparambil, "Social commerce defined," White paper #23747, IBM Corporation, Research Triangle Park, NC, November 2009.

[6] U. Gretzel, and K. Yoo, "Use and impact of online travel reviews," In: O'Connor, P., Hopken, W., Gretzel, U. (Eds.), Information and Communication Technologies in Tourism. Springer-Verlag, Wien, New York, pp. 35–46, 2008.

[7] Y. A. Kim, and J. Srivastava, "Impact of social influence in e-commerce decision making," In Proceedings of the Ninth International Conference on Electronic Commerce, Minneapolis, MN, ACM Press, New York, pp. 293–302, 2007.

[8] T. Liang, Y. Ho, Y. Li, and E. Turban, "What drives social commerce: the role of social support and relationship quality," International Journal of Electronic Commerce, 16, 2, pp. 69–90, 2011.

[9] S. Parise, and P. J. Guinan, "Marketing using Web 2.0. In R. Sprague (ed.)," Proceedings of the 41st Hawaii International Conference on System Sciences, Hawaii, IEEE Computer Society Press, Washington, DC, 2008.

[10] M. McPherson, L. Smith-Lovin, and J. M. Cook, "Birds of a Feather: Homophily in Social Networks," Annual Review of Sociology. 27, pp. 415–444, 2001.

[11] Z. Huang, and M. Benyoucef, "From e-commerce to social commerce: A close look at design features," Electronic Commerce Research and Applications, Elsevier B. V., 12, pp. 246-259, 2013.

[12] A. Lew, "A framework of tourist attraction research," Annals of Tourism Research, 14, pp. 533–575, 1987.

[13] G. Richards, "Tourism attraction systems-exploring cultural behavior," Annals of Tourism Research, vol. 29, no. 4, pp. 1048–1064, 2002.

[14] C.-C. Chang, and K.-H. Chu, "A Recommender System Combining Social Networks for Tourist Attractions," Fifth International Conference on Computational Intelligence, Communication Systems and Networks, IEEE Computer Society, pp. 42-47, 2013.

[15] D. J. Kim, D. L. Ferrin, and H. R. Rao, "A trust-based consumer decision-making model in electronic commerce: the role of trust, perceived risk, and their antecedents," Decision Support Systems, vol. 44, no. 2, pp. 544–564, 2008.

[16] L. Xiong, and L. Liu, "A Reputation-Based Trust Model for Peer-to-Peer eCommerce Communities," Proceedings of the 2004 IEEE International Symposium on Cluster Computing and the Grid, Washington, DC, USA, 2004.

[17] L. Mui, M. Mohtashemi, and A. Halberstadt, "A Computational Model of Trust and Reputation for

[30] M. Planti´e and M. Crampes, “Survey on Social Community Detection,” Social Media Retrieval, Springer Publishers (Ed.), pp. 65-85, 2013.

[31] U. Hanani, B. Shapira, and P. Shoval, "Information filtering: overview of issues, research and systems," User Modeling and User-Adapted Interaction, vol. 11, no. 3, pp. 203–259, 2001.

[32] G. Adomavicius, and A. Tuzhilin, "Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions," IEEE Transactions on Knowledge and Data Engineering, vol. 17, no. 6, pp. 734–749, 2005.

[33] R. Alton-Scheidl, R. Schumutzer, P. P. Sint, and G. Tscherteu, "Voting and Rating in Web4Groups," Oldenbourg, Vienna, Austria, pp. 13–103, 1997.

[34] L. Esmaeili, M. Mutallebi, Sh. Mardani, S. A. Hashemi G., "Studying the Affecting Factors on Trust in Social Commerce," International Journal of Advanced Studies in Computer Science & Engineering, vol. 4, no. 6, pp. 41-46, 2015.

[35] S. A. Hashemi G., L. Esmaeili, S. Mardani, and S. M. Mutallebi Esfidvajani, "A Survey of Trust in Social Commerce," E-Systems for the 21st Century: Concept, Developments, and Applications, Book chapter, vol. 1, Chapter 1, 2016.

[36] J. Tang, H. Gao, X. Hu, and H. Liu. "Exploiting homophily effect for trust prediction," In Proceedings of the sixth ACM international conference on Web search and data mining, ACM, pp. 53–62, 2013.

[37] H. Zardi, L. B. Romdhane, and Z. Guessoum, "A multi-agent Homophily-based-Approach for community detection in social networks," 2014 IEEE 26th International Conference on Tools with Artificial Intelligence, IEEE Computer Society, pp. 501-505, 2014.

[38] A. Abdul-Rahman, and S. Hailes, "Supporting trust in virtual communities," Proceedings of the Hawaii International Conference on System Sciences, Maui, Hawaii, 4–7 January, 2000.

[39] P. DeMeo, A. Nocera, G. Terracina, and D. Ursino, "Recommendation of similar users, resources and social networks in a social internetworking scenario," Information Sciences, vol. 181, no. 7, pp. 1285–1305, 2011.

[40] Pareto principle, https://en.wikipedia.org/wiki/Pareto_principle, Last access: 12/19/2016.

[41] Mean squared error, https://en.wikipedia.org/wiki/Mean_squared_error, Last access: 12/19/2016.

[42] J. B. Schafer, J. A. Konstan, and J. Riedl, "E-commerce recommendation applications," DataMining and Knowledge Discovery, vol. 5, no. 1–2, pp. 115–153, 2001.

E-businesses," Proceedings of the 35th Annual Hawaii International Conference on System Sciences, Washington DC, USA, 2002.

[18] J. He, and W.W. Chu, "A Social Network-Based Recommender System (SNRS)," Data Mining for Social Network Data, Springer Berlin Heidelberg, vol. 12, pp. 47-74, 2010.

[19] J. Kamahara, T. Asakawa, S. Shimojo, and H. Miyahara, "A Community-based Recommendation System to Reveal Unexpected Interests," Proceedings of the 11th International Multimedia Modelling Conference (MMM’05), IEEE Computer Society, 2005.

[20] J. Ben Schafer, J. Konstan, and J. Riedl, "Recommender systems in e-commerce," Proceedings of the 1st ACM conference on Electronic commerce, New York, USA, 1999.

[21] L. Esmaeili, B. Minaei-Bidgoli, H. Alinejad-Rokny, and M. Nasiri, "Hybrid Recommender System for Joining Virtual Communities," Research Journal of Applied Sciences, Engineering and Technology, vol. 4, Issue 5, pp. 500-509, 2012.

[22] H. J. Ahn, "A new similarity measure for collaborative filtering to alleviate the new user cold-starting problem," Information Sciences, vol. 178, no. 1, pp. 37–51, 2008.

[23] Y. Yang, and N. C. Marques, "User group profile modeling based on user transactional data for personalized systems," Lecture Notes in Computer Science, LNCS 3808, pp. 337–347, 2005.

[24] Y. Huang, and L. Bian, "A Bayesian network and analytic hierarchy process based personalized recommendations for tourist attractions over the Internet," Expert Systems with Applications, vol. 36, no. 1, pp. 933–943, 2009.

[25] S. Schiaffino, and A. Amandi, "Building an expert travel agent as a software agent," Expert Systems with Applications, 36, pp. 1291–1299, 2009.

[26] A. García-Crespo, J. Chamizo, I. Rivera, M. Mencke, R. Colomo-Palacios, and J. M. Gómez-Berbís, "SPETA: Social pervasive e-Tourism advisor," Telematics and Informatics, vol. 26, no. 3, pp. 306–315, 2009.

[27] A. Josang, R. Ismail, and C. Boyd, "A survey of trust and reputation systems for online service provision," Decision Support Systems, vol. 43, no. 2, pp. 618–644, 2007.

[28] D. W. Manchala, "E-commerce trust metrics and models," IEEE journal of Internet Computing, vol. 4, Issue 2, 2000

[29] J. Chen, O. R. Zaiane, and R. Goebel, "Detecting Communities in social networks using Max-Min Modularity," SIAM International Conference On Data Mining - SDM, pp. 978-989, 2009.

[43] P. Lops, M. de Gemmis, and G, Semeraro, "Content-based Recommender Systems: State of the Art and Trends," Recommender Systems Handbook, Springer, Chapter 3, pp. 73-105, 2011.

اطلاعات بررسی مقاله:

تاریخ ارسال: ۱۳۹۵/۱۱/۲۵
تاریخ اصلاح: ۱۳۹۵/۱۲/۱۴
تاریخ قبول شدن: ۱۳۹۵/۱۲/۱۷
نویسنده مرتبط: دکتر سید علیرضا هاشمی گلپایگانی، دانشکده مهندسی کامپیوتر و فناوری اطلاعات، دانشگاه صنعتی امیرکبیر، تهران، ایران.

[1]Gross Domestic Product

لیلا اسماعیلی دانشجوی دکترای مهندسی فناوری اطلاعات در دانشگاه صنعتی امیرکبیر (پلی تکنیک تهران) می‌باشد. زمینه تحقیقاتی وی فرایندکاوی در بستر تجارت اجتماعی است. این زمینه تحقیقاتی با مباحث باز مهندسی فرایندهای کسب و کار، تحلیل داده ها و تحلیل شبکه‌های اجتماعی مرتبط است. وی همچنین از سال ۱۳۸۹ تا کنون تحقیقات متعددی در حوزه سیستم‌های توصیه‌گر به ویژه سیستم‌های توصیه‌گر مبتنی بر تحلیل شبکه‌های اجتماعی انجام داده است. صنعت گردشگری یکی از حوزه‌های کاربردی مورد علاقه ایشان می‌باشد.
آدرس پست‌الکترونیکی ایشان عبارت است از:
Leila.esmaeili@aut.ac.ir

سید علیرضا هاشمی گلپایگانی استادیار و عضو هیات علمی گروه مهندسی فناوری اطلاعات در دانشگاه صنعتی امیرکبیر (پلی تکنیک تهران) می‌باشد. وی تحصیلات خود در مقطع کارشناسی، کارشناسی‌ارشد و دکتری را در رشته مهندسی صنایع و سیستم در دانشگاه صنعتی امیرکبیر به پایان رسانده است. دکتر هاشمی گلپایگانی به عنوان مشاور برای توسعه انواع مختلف سیستم‌های اطلاعاتی سازمانی در بسیاری از مراکز و موسسات دولتی و خصوصی مشغول به کار بوده است. وی از سال ۱۳۸۴ دروس مهندسی سیستم‌های تجارت الکترونیکی، مدیریت زنجیره تامین و بازمهندسی فرایندهای کسب‌وکار را در دانشگاه صنعتی امیرکبیر تدریس می‌نماید. زمینه تحقیقاتی مورد علاقه وی به ویژه در سال‌های اخیر شامل سیستم‌های توصیه‌گر، تحلیل شبکه‌های اجتماعی و کاربرد آن در تجارت الکترونیکی، مدیریت زنجیره تامین، فرایندکاوی و اینترنت اشیا می‌باشد.
آدرس پست‌الکترونیکی ایشان عبارت است از:
sa.hashemi@aut.ac.ir

زینب زنگنه مدار متولد ۱۳۷۰ و دارای مدرک کارشناسی در رشته مهندسی فناوری اطلاعات از دانشگاه صنعتی امیرکبیر تهران است. وی در حال حاضر دانشجوی رشته فناوری و مدیریت اطلاعات در مقطع کارشناسی‌ارشد در دانشگاه تگزاس در دالاس (The University of Texas at Dallas) می‌باشد. زمینه‌های مورد علاقه او شامل سیستم‌های توصیه‌گر، شبکه‌های اجتماعی، تجارت اجتماعی، مدیریت و تحلیل داده‌ها است.
آدرس پست‌الکترونیکی ایشان عبارت است از:
zeinab.zanganehmadar@utdallas.edu

راهنمای تهیه مقالات

نشریه علوم رایانش و فناوری اطلاعات

نسخه نهائی مقالات ارسالی برای چاپ در نشریه "علوم رایانش و فناوری اطلاعات" باید بر طبق اصول مطرح شده در این راهنما تهیه شده باشد. رعایت این اصول در نسخه اولیه نیز قویاً توصیه می‌شود. مقالات به زبان فارسی ارسال گردد.

۱- ساختار مقاله

- عنوان: کوتاه و معرف محتوای مقاله باشد و از ۱۵ کلمه تجاوز نکند.
- نام نویسندگان و مؤسسه محل اشتغال آنان: از ذکر عناوین خودداری شود.
- چکیده فارسی: حاوی تعریف مسأله، روش حل، و نتایج مهم باشد و از ۱۵۰ کلمه تجاوز نکند.
- واژه‌های کلیدی: حداکثر ۱۰ کلمه
- بدنه اصلی مقاله: بدنه اصلی با **"مقدمه"** شروع و با **"نتیجه‌گیری"** خاتمه می‌یابد. بخش‌ها و زیربخش‌های بدنه اصلی باید شماره‌گذاری شوند. شماره **"مقدمه"** یک خواهد بود.
- تشکر و قدردانی (در صورت نیاز).
- مراجع: مراجع به ترتیبی که در متن به آنها رجوع می‌شود آورده شوند. نام مؤلفان مراجع در صورت لزوم در متن بصورت فارسی آورده شود. رجوع به مراجع با ذکر شماره آنها در داخل کروشه ([]) انجام پذیرد.
- پیوست‌ها (در صورت نیاز)
- واژه‌نامه (در صورت نیاز)
- برای مقالات فارسی، عنوان مقاله، نام نویسندگان، مؤسسه محل اشتغال، چکیده، و کلمات کلیدی به زبان انگلیسی نیز در صفحه‌ای جداگانه داده شود.
- بیوگرافی کامل نویسندگان به زبان فارسی به همراه عکس

۲- معادله‌ها، شکل‌ها، جدول‌ها، و عکس‌ها

- معادله‌ها باید با فاصله کافی از بالا و پائین تایپ و به صورت متوالی شماره‌گذاری شوند. شماره معادله در پرانتز در انتهای سمت راست سطر حاوی معادله قرار داده شود. معادلات دستنویس به هیچ شکل قابل قبول نیستند.
- شکل‌ها و جدول‌ها باید دارای شماره و عنوان باشند. در شکل‌ها شماره و عنوان در زیر شکل و در جدول‌ها در بالای شکل قرار می‌گیرد. اعداد و متون روی شکل‌ها و جدول‌ها باید دارای اندازه مناسب و کاملاً خوانا باشند.
- اعداد و کلمات روی شکل‌ها و جدول‌ها در مقالات فارسی به زبان فارسی باشند.
- عکس‌ها سیاه و سفید، برقی، و با کیفیت عالی باشند.

۳- نحوه نگارش مراجع

در لیست مراجع انواع مختلف مرجع‌ها به شکل زیر نوشته شوند:

[۱] ب. مقدم، ا. تقوی، و ن. طاهری، آشنائی با شبکه‌های کامپیوتری، چاپ دوم، انتشارات نصر، تهران، ۱۳۷۵.

[۲] ی. براون، مقدمه‌ای بر شبکه‌های عصبی، ترجمه م. ع. آرام، انتشارات فجر، مشهد، ۱۳۷۰.

[۳] راهنمای کاربران حسابرس، شرکت پردازش رایانه‌ای ایران، تهران، ۱۳۶۵.

[۴] ج. عارف، استنتاج فازی بوسیله شبکه‌های عصبی، پایان‌نامه کارشناسی ارشد هوش مصنوعی، دانشکده مهندسی کامپیوتر، دانشگاه صنعتی شهر، ۱۳۷۴.

[۵] ج. حسینی، و ح. ربانی، "تشخیص چهره انسان در تصویر،" نشریه امیرکبیر، سال هشتم، شماره ۴۲، ص ۱۲۵-۱۴۷، ۱۳۷۷.

[۶] ج. حسینی، و ح. ربانی، "تشخیص چهره انسان در تصویر،" در مجموعه مقالات هفتمین کنفرانس سالانه انجمن کامپیوتر ایران، ص ۲۲۴-۲۳۲، ۱۳۸۰.

[7] M. A. Ahmadi, and M. H. Rahimi, *Fuzzy Set Theory*, New Jersey: Prentice-Hall, 1995.

[8] M. A. Ahmadi, M. H. Rahimi, and A. Fatemi, ″Evidence-Based Recognition of 3D Objects,″ *IEEE Trans. Patt. Anal. and Mach. Intell.*, vol. 12, no. 10, pp. 811-835, 1994.

[9] A. Taheri, ″On-Line Fingerprint Verification,″ *Proc. IEEE Intl Conf. Pattern Recognition*, pp. 752-758, 1992.

[10] M. A. Ahmadi, *On-line Fingerprint Verification*, Ph. D. Dissertation, MIT, Cambridge, MA, 1982.

[11] A. J. Washington, ″The Fingerprint of MalcomX,″ http://www.dermatoglyphics.com, June 2003.

[12] International Biometrics Group, http://www.biometricgroup.com, May 2003.

۴- واژه‌نامه

هر واژه خارجی در واژه‌نامه **در انتهای مقاله** با شماره‌ای مشخص شود و شماره در معادل فارسی آن واژه در متن، بصورت بالانویس آورده شود.

۵- آماده‌سازی مقاله

- مقاله را با نرم‌افزار Word تایپ نمائید.
- متن چکیده به صورت تک ستونی با طول سطر ۱۸ سانتیمتر و متن مقاله به صورت دو ستونی با طول هر ستون ۸۷ میلیمتر و فاصله دو ستون ۶ میلیمتر تایپ شود. حاشیه‌ها از بالا و پائین برابر ۲۰ میلیمتر و از طرفین برابر ۱۵ میلیمتر اختیار شود.
- فاصله عنوان مقاله در صفحه اول از بالای صفحه برابر ۸۵ میلیمتر باشد و عنوان وسط چین شود.
- کلیه عناوین بصورت پررنگ با قلم **"B Nazanin"** تایپ شوند، اندازه قلم عنوان مقاله ۱۸، عناوین سطح اول ۱۴، و عناوین سطح دوم و سوم ۱۲ انتخاب شوند.
- متن چکیده‌ها با قلم **"B Nazanin"** اندازه ۹، متن مقاله با قلم **"B Nazanin"** اندازه ۱۰، و کلمات و متن انگلیسی با قلم **Times New Roman** اندازه ۹ تایپ شوند.
- تمام متن بصورت تک فاصله تایپ شود. اسامی نویسندگان از عنوان مقاله و اسامی نویسندگان از عناوین محل اشتغال نویسندگان دو خط فاصله داشته باشد. بالای هر عنوان یک سطر فاصله قرار داده شود.
- سعی شود تعداد صفحات مقاله از ۳۰ صفحه بیشتر نباشد.

۶- نحوه ارسال مقاله

- ارسال مقاله فقط از طریق ایمیل مجله (**csitjour@gmail.com**) انجام شود.
- **مقاله ارسالی برای نشریه علوم رایانش و فناوری اطلاعات نباید در جای دیگری به چاپ رسیده باشد و یا در زمان بررسی توسط نشریه برای چاپ به نشریه دیگری ارسال گردد.**
- پس از قبول مقاله، نسخه نهائی تصحیح شده مقاله باید در قالب‌های Word و PDF به نشریه ارسال گردد.
- در نسخه نهائی باید بیوگرافی کلیه نویسندگان (به زبان فارسی) و عکس آنها در انتهای مقاله قرار داده شود، همچنین عنوان مقاله، نام نویسندگان، مؤسسه محل اشتغال، چکیده، کلمات کلیدی به زبان انگلسی در فایل جداگانه ارسال شود.

The CSI Journal on Computing Science and Information Technology
Vol. 14, No. 2, 2017

A Recommender System for the Tourism Industry, in the Context of Social Commerce: Based on the Similarity, Social Communities, Trust, and Reputation

Leila Esmaeili Seyyed Alireza Hashemi G. Zeinab Zangeneh Madar

Department of Computer Engineering and Information Technology, Amirkabir University of Technology, Tehran

ABSTRACT

The internet and its services have significantly affected various businesses, including the tourism industry, and provided a wide range of diversity in the products and services. Due to a dramatic increase in the number of available options in travels, hotels, tourist attractions, etc., the process of decision-making has become more difficult for the consumer. As a result, Tourism Recommender Systems (TRS) have attracted the attention of researchers and businesses. Tourist attractions are often the reason why people love to travel. This research represents a social-hybrid recommender system in the context of a social commerce, which can create a personalized list of tourist attractions for each tourist, based on the similarities of the users' desires and interests, trust, reputation, communications, and social communities. The advantage of the proposed method in comparison with the traditional methods like collaborative filtering, content based filtering and hybrid, is the comprehensive usage of various factors and consideration of the trust factor in recommendation resources- as in the identification of the outliers. The results of the tests show the superiority of the presented method compared to other common methods; In addition, the proposed model can be used to recommend other products and services in the tourism industry and also in other social businesses.

Keywords: Social Commerce, Recommender System, Trust, Similarity, Communities, Reputation, Tourism Industry, Social Communications.

The CSI Journal on
Computing Science and
Information Technology
Vol. 14, No. 2, 2017

Persian Multi-Documents Summarization by Deep Learning

Shima Mehrabi **Hamid Reza Ahmadifar** **Seyed Abolghasem Mirroshandel**

Faculty of Engineering, University of Guilan, Rasht, Iran

ABSTRACT

With the increasing amount of the accessible textual information via internet, it seems necessary to have summarization system which is able to generate summary of information for user demands. Since long time ago, summarization has been considered by natural language processing researchers. Today, with improvement in processing power and development of computational tools, efforts to improve the performance of summarization system are continued. In this paper, a novel Persian multi-document summarization system is proposed that works based on a machine learning method called Deep Learning. Mostly Deep Learning uses artificial neural networks in learning process. The result of using deep learning in speech recognition and picture processing are sound promising which convinced natural language processing researchers to apply Deep Learning in NLP tasks. The proposed system ranks the sentences based on some predefined features and by using a deep artificial neural network called Autoencoder. The performance of the system is evaluated in Persian and the result of evaluations demonstrates the effectiveness and success of proposed summarization system.

Keywords: Artificial Neural Networks, Deep Learning, Multi-Documents Summarization, Persian Language Processing.

The CSI Journal on Computing Science and Information Technology
Vol. 14, No. 2, 2017

System Level Multi-Core Thermal Management for Work-Stealing Based Parallel Programs

Hamid Goharjoo[1] **Morteza Moradi**[1,2] **Hamid Noori**[1,2]

[1]Faculty of Engineering, Ferdowsi University of Mashhad, Mashhad, Iran
[2]School of Computer Science, Institute for Research in Fundamental Sciences (IPM), Tehran, Iran

ABSTRACT

In recent years, temperature and power consumption of multi-core processors have become major challenges for designers and users. As processor's temperature rises, cooling costs and power consumption increases and affects processor life time. According to our studies, no thermal management approach has been proposed at the operating system being aware of work-stealing scheduler in parallel programs. In this paper, a dynamic thermal management algorithm has been proposed at the operating system level that manages the processor temperature according to a threshold for work-stealing based parallel programs. Hence, we propose temperature and performance models to predict the future temperature and estimate the amount of program performance changes. Using proposed models, the proposed algorithm determines the appropriate number of active cores and processor frequency such that temperature does not violate the threshold and the performance degradation minimizes. Experiments show that the proposed algorithm has up to 28% higher performance than neighbor-aware algorithm and, unlike this algorithm, never violates the specified temperature threshold.

Keywords: Parallel Execution, Multi-core, Work-stealing, Operating System, Scheduler, Thermal Management.

The CSI Journal on Computing Science and Information Technology
Vol. 14, No. 2, 2017

Prediction of Insurance Customers' Behavior Using a Combination of Data Mining Techniques

Ehsan Mokhtari **Seyed Abolghasem MirRoshandel**

Faculty of Engineering, University of Guilan, Rasht, Iran

ABSTRACT

Today, the segmentation and differentiation of customers based on their behaviors and needs is the most important action of insurance companies. Therefore, these companies widely and purposefully carry out advertising and marketing in all communication environments in order to identify and stimulate their clients. For better effectiveness of this approach, customers are segmented and differentiated based on special criteria and objectives. Clustering is an analytical method for detecting the performance and behavior of clients through their information. This allows companies to make decisions and carry out purposeful advertising toward them through the performance of clients. The main objective of this study is to provide a way to identify and predict the performance and behavior of new customers in choosing the insurance type in order to protect their house against risks by combining the K-medoids method with neural network to identify the cluster of new customers for offering insurance products advertisements. In this regard, due to the excess of characteristics in datasets and their dispersion, first the conceptual patterns have been discovered through K-means and K-medoids techniques and after determining the cluster of customers, their cluster is predicted using these patterns just through demographic information from new customers. The significant feature of this study is the combination of clustering and classification methods in pattern discovery. The conducted experiments show the success of proposed method in the recognition and discovery of customers' needs, behavior, and performance based on which advertising takes place.

Keywords: Marketing, Purposeful Advertising, Clustering, K-Medoids, K-Means, Neural System, Customer Segmentation, Pattern Discovery.

The CSI Journal on Computing Science and Information Technology
Vol. 14, No. 2, 2017

Improving Rule-Based Machine Translation by Using Statistical Syntactic Rules

Hakimeh Fadaei **Heshaam Faili** **Farnaz Ghasemi**

School of Electrical and Computer Engineering, University of Tehran, Tehran, Iran

ABSTRACT

Rule-based machine translation uses a set of linguistic rules in the process of translation. The results of these systems are usually better than the results of statistical models from grammatical and word order perspective, But it has been shown that statistical models are more powerful in selecting proper words and generating more fluent translations. In this paper our goal is to improve the word choice in rule-based machine translation. This is done by a set of lexical syntactic rules based on Tree Adjoining Grammar. These probabilistic rules are statistically extracted from a large parallel corpus. In the proposed system, the input sentence is first reordered by a rule-based system, and them the decoding is carried out monotonically by using dynamic programming. In this system the best translation is chosen based on the extracted rules and the language model score. The experiments on English-Persian translation showed that the proposed method resulted in an improvement of 1.3 in BLEU score in comparison to our baseline rule-based method.

Keywords: Hybrid Machine Translation, Rule-Based Machine Translation, Statistical Rules, Lexical Syntactic Rules, Tree Adjoining Grammar.

The CSI Journal on Computing Science and Information Technology
Vol. 14, No. 2, 2017

Simultaneous Image Classification and Annotation Using Probabilistic Topic Models and LLC Encoding of Visual Words

Seyed Navid Mohammadi Foumani **Ahmad Nickabadi**

Department of Computer Engineering and Information Technology, Amirkabir University of Technology, Tehran, Iran

ABSTRACT

There have been several attempts so far to use topic models like LDA for simultaneous image classification and annotation. Recently, other models based on probabilistic neural networks, e.g. SupDocNADE, have been introduced with promising results in modeling multinomial data. In these models, annotation words along with visual words are embedded in the same framework and are considered as the feature vector. The imbalance between the number of visual words and annotation words may cause some problems. For instance, the contribution of the annotation words to the hidden representation is so smaller than the contribution of the visual words that the network might easily ignore it. To address this issue, we propose a weighting scheme for the annotation words. We also use LLC to generate the visual words in our proposed model. Experimental results show a 5% increase in the F-measure in the proposed annotation words and a 1% improvement in classification accuracy compared to existing models on UIUC_Sports and LabelMe datasets.

Keywords: Image Classification and Annotation, Topic Models, Probabilistic Model, Neural Network, LLC Coding.

The CSI Journal on Computing Science and Information Technology

Vol. 14 **No. 2** **2017**

ABSTRACTS

The CSI Journal on Computing Science and Information Technology

A Semiannual Publication of Computer Society of Iran (CSI)