استفاده از راهکار شبکهٔ عصبی در بازیابی اطلاعات متنی

سارا کلینی[1]

چکیده

با افزایش حجم اطلاعات و با پیشرفت تکنولوژی، استفاده از الگوریتم‌های سنتی جهت بازیابی سریع داده‌ها کافی نبوده و به‌کارگیری راهکارهای نوین را جهت تسریع در بازیابی اطلاعات مربوط طلب می‌کند. در روش‌های سنتی، پردازش اطلاعات، معمولاً به‌صورت ترتیبی صورت می‌گیرد. در روش‌های جدید بازیابی اطلاعات، علاوه بر سرعت بازیابی، درک محتوای مدرک و بازیابی مدرک مربوط حائز اهمیت می‌باشد. به‌کار بردن روش‌های هوش مصنوعی در بازیابی سریع مدارک مربوط، بسیار موفق بوده است. استفاده از شبکه‌های عصبی به‌عنوان یکی از تکنیک‌های هوش مصنوعی، راهکار مناسبی جهت افزایش سرعت بازیابی اطلاعات در حجم انبوه است. شبکه‌های عصبی، بازنمایی مناسب دانش جهت کاربردهای بازیابی اطلاعات را ارائه می‌دهند. گره‌های شبکهٔ عصبی نمایانگر عناصر مربوط در مجموعه مدارک از قبیل کلیدواژه، نویسنده، و... است و از پیوندهای موجود در شبکه جهت انتقال ورودی از لایه‌ای به لایهٔ دیگر استفاده می‌شود که در نهایت به دریافت خروجی شبکه، که همان مدرک بازیابی شده است، می‌انجامد. در این مقاله، به نحوهٔ استفاده از شبکهٔ عصبی خودسازمانده (SOM) در خوشه‌بندی داده‌ها به‌منظور بازیابی اطلاعات متنی پرداخته شده و یک مدل شبکهٔ عصبی خودسازمانده برای بازیابی اطلاعات نمونه از پایگاه اطلاعاتی Medlineپیاده‌سازی گردیده است.

کلیدواژه‌ها

شبکهٔ عصبی، بازیابی اطلاعات، اطلاعات متنی، شبکهٔ عصبی خودسازمانده، الگوریتم خوشه‌بندی، شبکهٔ SOM

۱. کارشناس ارشد مهندسی کامپیوتر و مدیر ادارهٔ فن‌آوری‌های اطلاعاتی مرکز منطقه‌ای اطلاع‌رسانی علوم و فن‌آوری شیراز
koleini@srlst.com

مقدمه

انگیزهٔ پیاده‌سازی شبکهٔ عصبی، توسط رایانه با الگوگرفتن از مغز انسان و فعالیت‌های پیچیدهٔ آن آغاز شد. در توسعهٔ شبکهٔ عصبی، سعی در پیش‌گویی عکس‌العمل مبنی بر محرک‌های دریافتی مطابق با شبکهٔ عصبی انسان است و فرضیه‌های زیر جهت پیاده‌سازی این شبکه‌ها بر طبق شبکهٔ عصبی طبیعی درنظر گرفته شده است:

۱. پردازش داده‌ها در عناصر ساده‌ای به‌عنوان نرون انجام می‌شود؛

۲. سیگنال‌ها (اطلاعات) از طریق اتصالات بین نرون‌ها منتشر می‌شوند؛

۳. به هر اتصال، یک وزن تخصیص می‌یابد که این وزن (که کمیّت عددی است) در سیگنال‌های (اطلاعات) منتشر شده ضرب می‌شود، یا به بیان دیگر، هنگامی که یک سیگنال (اطلاعات) از یک نرون به نرون دیگر در حال حرکت است، تحت‌تأثیر محیط انتشار خود قرار می‌گیرد.

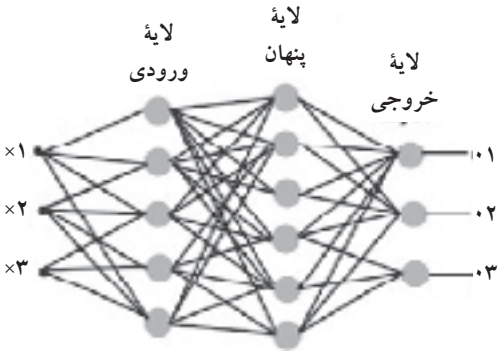
در یادگیری با سرپرستی[2] شبکه‌های عصبی، فرض بر آن است که یک معلم در هنگام عملیات یادگیری حضور دارد و هر الگو برای یادگیری شامل داده‌های ورودی و خروجی است. در زمان یادگیری، مقایسه‌ای میان خروجی محاسبه شده توسط شبکهٔ عصبی با مقدار داده‌های خروجی که موردنظر است، انجام می‌گیرد و اختلاف این دو مقدار به‌عنوان خطا درنظر گرفته می‌شود. چنانچه مقدار خطا از میزان معینی بیشتر باشد، می‌تواند جهت تعیین پارامترهای اصلی شبکه از جمله وزن‌های ارتباطی بین نرون‌ها، مجدداً در فرایند آموزش وارد شود (**۱**: ۱۰۴-۱۱۲).

شبکه‌های عصبی چندلایه، از تعدادی گره و پیوند تشکیل شده است. اطلاعات، از طریق گره‌های ورودی به شبکهٔ عصبی وارد شده و سپس با استفاده از پیوندها به لایه‌های بعدی (پنهان) منتقل گردیده و درنهایت، خروجی شبکه از گره‌های لایهٔ خروجی به‌دست می‌آید (**۲**: ۹۶). در شبکهٔ عصبی، هر گره دارای مقادیر ورودی و خروجی شبکه است. مقادیر خروجی پس از فعال شدن از طریق پیوندها به سایر گره‌ها منتقل می‌شوند. پیوندها وزن‌دهی شده‌اند، بنابراین مقادیری که در طول پیوندها عبور می‌کند برابر با حاصل‌ضرب خروجی گره‌های فرستنده و وزن پیوند می‌باشد. مقدار ورودی یک گره برابر با حاصل‌ضرب خروجی گره‌های فرستنده و وزن پیوند می‌باشد. مقدار ورودی یک گره برابر با حاصل جمع همهٔ وزن‌های ورودی است. شبکه‌های عصبی می‌توانند در لایه‌ها طوری ساخته شوند که تمام داده‌هایی که توسط شبکه دریافت می‌شود در چند مرحله فعال گردد و در اینجا کل یک لایهٔ شبکه، داده را در یک مرحله به لایهٔ بعدی منتقل می‌کند.

جهت انتخاب یک مدل برمبنای شبکهٔ عصبی، نخست نوع شبکه و الگوریتم آن معین می‌گردد، سپس معماری شبکه (تعداد لایه‌های شبکه، تعداد گره‌ها و چگونگی پیوند میان گره‌ها) انتخاب می‌شود و شبکهٔ

2. Supervised learning

عصبی براساس این اطلاعات آموزش می‌بینند. با توجه به نوع الگوریتم یادگیری، وزن پیوندها تغییر یافته و با تنظیم توابع انتقال و گره‌های هر لایه به‌صورت سعی و خطا، خروجی مطلوب به‌دست می‌آید. شکل ۱، نمونه‌ای از شبکهٔ عصبی چندلایه را نشان می‌دهد.



شکل ۱

در این مقاله، از شبکهٔ عصبی خودسازمان‌ده[3] جهت خوشه‌بندی داده‌ها استفاده شده است؛ لذا به معرفی اجمالی این شبکه پرداخته و سپس مروری بر نحوهٔ بازیابی اطلاعات با به‌کارگیری شبکهٔ عصبی خواهیم داشت.

شبکهٔ عصبی خودسازمان‌ده (SOM)

کوهنن[4] از سال ۱۹۸۸، مطالعات عمیقی را دربارهٔ شبکهٔ سادهٔ خودسازمان‌ده به انجام رسانیده و درنهایت موفق به طراحی شبکهٔ پیچیده SOFM [5] شده است (۸: ۱۶۹). شبکهٔ خودسازمان‌ده جهت نگاشت داده‌هایی با ابعاد زیاد به فضایی با ابعاد کمتر به‌کار برده می‌شود. این شبکه، شامل تعدادی گره است که معمولاً به‌صورت شبکهٔ مستطیلی یا شبکهٔ شش ضلعی مرتب می‌شوند و هر نرون به ورودی متصل است.

از ویژگی‌های شبکهٔ SOFM سازمان‌دهی آن است که از توانایی انتخاب همسایگی برنده به‌جای یک گره برنده برخوردار است. در شبکهٔ عصبی خودسازمان‌ده که به‌منظور بازیابی اطلاعات به‌کار گرفته شده است، بردارهای مدارک مشابه در یک فضای معین، گروه‌بندی یا به عبارت دیگر خوشه‌بندی می‌شوند.

انواع آرایش و جای‌گشت خوشه‌ها، ارتباط ویژگی خوشه‌ها را در فضای مدارک منعکس می‌سازد. برای مثال، اندازهٔ خوشه‌ها (تعداد گره‌هایی که به هر خوشه اختصاص می‌یابد) بیانگر توزیع فراوانی مدرک در مجموعه مدارک ورودی است.

از جمله موارد استفادهٔ شبکه‌های عصبی می‌توان به کاربرد آن در زمینهٔ داده‌کاوی[6] اشاره نمود.

معماری شبکهٔ کوهنن شامل دو لایه است:

۱. لایهٔ ورودی
۲. لایهٔ کوهنن (لایهٔ خروجی)

این دو لایه، کاملاً به یکدیگر متصل هستند. هر نرون، لایهٔ ورودی اتصال پیش‌خورد یا مستقیم[7] به تمام نرون‌های لایهٔ خروجی دارد.

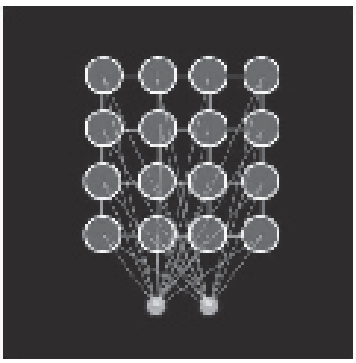
3. Self Organizing Map (SOM)
4. Kohonen
5. Self Organizing Feature Map
6. Data mining
7. Feed forward

شبکهٔ کوهنن در دو مرحله کار می‌کند:

۱. شبکـه، نرون‌هایی را که بردار وزن ارتباطی آنها نزدیکی بیشتری به بردار ورودی جاری داشته باشد، به‌عنوان نرون برنده انتخاب می‌کند.

۲. پـس از انتخـاب همسایگـی بـرنـده، بردارهای متصـل به واحـدهایـی کـه مقـدار خروجـی آنها مثبت است بـه طرف بردار ورودی به حرکت درمی‌آیند.

ورودی به لایهٔ کوهنن یا لایهٔ خروجی می‌تواند با ضرب داخلی میان بردار وزن نرون و بردار ورودی محاسبه گردد. نرون لایهٔ خروجی برنده، نرونی است که بزرگ‌ترین ضرب داخلی را داشته باشد. زاویهٔ میان بردار وزنی نرون برنده و بردار ورودی، کوچک‌تر از زاویه با سایر نرون‌هاست. شکل ۲، نمونه‌ای از شبکهٔ خودسازمان‌ده را نشان می‌دهد.



شکل ۲. شبکهٔ عصبی خودسازمان‌ده

روش دوم جهت انتخاب نرون برنده، انتخاب نرونی است که بردار وزنی آن دارای کوچک‌ترین حد فاصلهٔ اقلیدسی از بردار ورودی باشد. در اینجا، از این روش جهت انتخاب همسایگی نرون برنده استفاده شد. الگوریتم ارائه شده در شکل ۳، جهت آموزش شبکهٔ خودسازمان‌ده به‌کار برده می‌شود.

step0 initialize weights w_{ij}
set topological neighborhood parameters
set learning rate parameter
step 1 while stopping condition is false do step 2-8
step2 for each input vector x , do step 3-5
step3 for each j, compute:
$d(j)=\sum_i (w_{ij}-x_i)^2$

step4 find index J such that d(J) is a minimum
step 5 for all units j whitin a specified neighborhood of J and for all i:
$w_{ij}(new)= w_{ij}(old)+\alpha [x_i-w_{ij}(old)]$
step6 update learning rate
step7 reduce radius of topological neighborhood at specified time
step8 test stopping condition

شکل ۳. الگوریتم یادگیری شبکهٔ خودسازمان‌ده

بازیابی اطلاعات براساس شبکهٔ عصبی

راهکارهای بازیابی، میزان شباهت میان یک «پرس‌وجو»[8] و یک مدرک را بیان می‌کنند. اساس این راهکارها بر اصل ارتباط بیشتر میان «پرس‌وجو» و مدارک استوار است. یک راهکار بازیابی، الگوریتمی است که از پرس‌وجوی Q و مجموعه‌ای از مدارک $D_1, D_2, .., D_n$ استفاده کرده و ضریب شباهت $SC(Q,D_i)$ [9] را برای تمام مدارک محاسبه می‌نماید (۲: ۹۶).

یکی از راهکارهای تأثیرگذار در بازیابی اطلاعات، استفاده از سیستم شبکهٔ عصبی است که در آن مجموعه‌ای از نرون‌ها یا گره‌های شبکه به هنگام پرس‌وجو و در زمان بازیابی مدارک فعال می‌شوند.

در مدل‌های مختلف شبکه‌های عصبی، اطلاعات به‌صورت شبکهٔ وزن‌دار نمایش داده می‌شود. برخلاف روش‌های سنتی پردازش اطلاعات، مدل‌های شبکه‌های عصبی به‌عنوان خود- پردازشگر بدون دخالت برنامهٔ خارجی دیگر در شبکه، عمل می‌نمایند. مطابق خصوصیت شبکهٔ عصبی در زمان فعل و انفعال‌های محلی که به‌طور هم‌زمان میان اجزای شبکه رخ می‌دهد، پردازش داده‌ها انجام می‌شود. در مدل بازیابی اطلاعات سنتی، پردازش خارجی که بر روی ساختار داده‌ها عمل می‌کند، معمولاً به تمامی مجموعه مدارک، دسترسی کلی دارد و پردازش، عمدتاً ترتیبی است.

به نظر می‌رسد که محاسبات شبکه‌های عصبی، در مقایسه با مدل فضای برداری و مدل‌های احتمال، تناسب بهتری با مدل‌های بازیابی سنتی داشته باشد.

از مزایای استفاده از شبکهٔ عصبی در بازیابی سریع اطلاعات در مجموعه مدارک با حجم انبوه، می‌توان به موارد زیر اشاره نمود:

۱. در زمانی که اطلاعات (کلیدواژه) مورد جستجو، دقیقاً در مدارک پیدا نشود با استفاده از شبکهٔ عصبی می‌توان به بازیابی داده‌هایی که از نظر همسایگی به اطلاعات خواسته شده نزدیک‌تر هستند، پرداخت.

۲. می‌توان اطلاعات را با الگوهای مشترک دسته‌بندی نمود.

پیشینهٔ تحقیق

داس کاکس[10] و دیگران، مرور جامعی دربارهٔ کاربرد مدل‌های ارتباطی در بازیابی اطلاعات، انجام داده‌اند. بخش مهمی از تحقیقات پیرامون بازیابی اطلاعات را می‌توان در چارچوب مدل‌های ارتباطی مورد توجه قرار داد . برای مثال، از آنجا که تمام مدل‌های ارتباطی به‌عنوان سیستم‌های رده‌بندی ورودی - به- خروجی مطرحند، خوشه‌بندی مدرک را می‌توان به‌عنوان رده‌بندی فضای مدرک × مدرک در نظر گرفت. ساخت اصطلاح‌نامه به‌عنوان سیستمی هماهنگ با فضای نمایه × نمایه مطرح بوده و جستجو را نیز می‌توان به‌صورت ارتباط پیوند در فضای مدرک × نمایه تلقی نمود (۴: ۲۰۹-۲۶۰) .

کرستانی[11]، در مقالهٔ خود مدلی از شبکه را ارائه می‌دهد که می‌تواند در ایجاد طرح

8. Query
9. Similarity coefficient
10. Doszkocs
11. Crestani

مفهومی و منطقی جهت کاربردهای بازیابی اطلاعات مورد استفاده قرار گیرد. این مدل، دارای خصیصه‌های انعطاف‌پذیر قابل توجهی است که می‌توان آن را به طرق مختلف و مؤثر به‌کار گرفت. این مؤلفان، در سال ۱۹۹۷ شبکهٔ تبدیل[12] را برای بهینه‌سازی پرس‌وجو پیشنهاد کردند. این شبکه شامل شبکهٔ پس انتشار خطا با یک یا چند لایهٔ پنهانی است که در آن، ورودی و خروجی طرحواره هایی از بازنمونه‌ها به‌شمار می‌آید (۵: ۸۴۶-۸۵۰).

گرونفلد[13]، با استفاده از شبکهٔ عصبی هاپفیلد، گره‌هایی را برای مفاهیم «پرس‌وجو» و نیز گره‌هایی را برای «مدارک» درنظر گرفت. در این شبکه، گره‌های مدارکی که بیشتر فعال بودند برای بازیابی به‌عنوان مدرک مربوط انتخاب می‌شدند. گرونفلد با استفاده از این مدل پیوند بین گره‌ها را براساس ماتریس مدرک × کلمه که به‌وسیلهٔ الگوریتم‌های متداول نمایه تعریف می‌شوند، وزن‌دهی نمود (۶: ۱۴۵).

هاتانو[14]، در ارتباط با خوشه‌بندی مؤثر و بازیابی متن و داده‌های ویدئویی مبتنی بر شباهت‌های موجود، نظام سامان‌دهندهٔ اطلاعات را پیشنهاد کرد. به‌جای کلیدواژه‌ها، این نویسندگان از مدل فضای برداری و کدگذاری تصویر براساس تبدیل کسینوسی گسسته به‌منظور استخراج خصایص داده‌ها استفاده نمودند. داده‌ها برحسب شبکهٔ عصبی کوهنن خوشه‌بندی گردیده و نتیجهٔ آن، به یک شکل سه بُعدی نمایش داده می‌شود. هاتانو و همکاران وی معتقدند که سیستم پیشنهادی آنان به‌طور مؤثری به استفادهٔ مجدد از داده‌های تصویری و متنی توزیع یافته، کمک می‌کند (۷: ۲۰۵-۲۱۴).

مقایسهٔ میان الگوریتم‌های کوهنن، که براساس بسامد واژه‌ها استوارند، و الگوریتم‌هایی که براساس اندازه‌گیری سالتن هستند، نشان می‌دهد که الگوریتم‌های کوهنن جهت خوشه‌بندی و تولید نگاشت کلی اسناد مؤثرترند. ماندل[15]، از شبکهٔ عصبی «پس انتشار خطا»[16] برای ساخت مدل کاسی میر[17] به‌منظور تطبیق میان بازنمایی پرس‌وجو و مدرک استفاده نمود. پرس‌وجو و مدرک، هر دو به‌عنوان ورودی شبکه عمل می‌کنند و شبکه به مشابهت آنها در لایهٔ خروجی می‌پردازد. به‌عبارت دیگر، هر چه ضریب شباهت بیشتر باشد، میزان ربط بین مدرک و پرس‌وجو نیز بیشتر خواهد بود (۹: ۱۳-۱۶).

روش پژوهش

برای این پژوهش، از داده‌های متنی موجود در دیسک نوری مدلاین[18] استفاده شد. اطلاعات درخواستی از طریق کلیدواژه‌ها قابل دسترسی می‌باشند. موضوع مورد بحث Radiation در سه مقولهٔ زیر می‌باشد:

1. Radiology
2. Radiotherapy
3. Shielding

12. Transformation network
13. Grunfeld
14. Hatano
15. Mandl
16. Back propagation
17. COGnitive SIMilarity Learning in IR (COSIMIR)
18. Medline

از هر مقوله ۵۰ مدرک و در مجموع ۱۵۰ مدرک مورد بررسی قرار گرفت. در سه مقولهٔ فوق، چهار کلیدواژهٔ Dose، X-ray ، CT ، و Depth در نظر گرفته شد.

هـدف، استفاده از الگوریتم خوشه‌بندی شبکـهٔ عصبـی، جهت دسته‌بنـدی سـه ردهٔ Radiology، Radiotherapy، و Shielding است. بـا اطلاع از این‌کـه در مدارک مورد بحث در زمینهٔ Radiology کلیدواژهٔ Dose بیش از سه کلیدواژهٔ دیگر بوده و در زمینهٔ Shielding ،کلید واژهٔ Depth بیشتر و در زمینهٔ Radiotherapy، کلیدواژه‌های X-ray و Dose بیشتر است، به محاسبهٔ وزن مدارک پرداخته و کدهایی جهت پیش‌پردازش و رمزگذاری مدارک متنی به بردارهای عددی نوشته شد. جهت تعیین بردار وزنی هر مدرک، از فراوانی کلیدواژه در مدرک استفاده گردید. به عبارت دیگر وزن w_{ik} مدرک به‌صورت فراوانی کلیدواژه یا کلمهٔ t_k در مدرک d_i تعریف می‌شود. از فرمول زیر جهت تعیین وزن استفاده شد:

$$W_{ik} = \frac{tf_{ik} \cdot \log(N/n_k)}{\sqrt{\sum_{j=1}^{t} (tf_{ij})^2 \cdot (\log(N/n_j))^2}}$$

در این فرمول، tf_{ik} بسامد کلیدواژهٔ t_k در مدرک d_i و پارامتر N تعداد مدارک را نشان می‌دهد، و n_k بیانگر تعداد مدارکی است که شامل واژه یا کلیدواژهٔ t_k می‌باشند. برای ۱۵۰ مدرک نمونه، پارامترهای فوق، معین شد.

الگوریتم شبکهٔ عصبی جهت خوشه‌بندی مدارک با اختصاص یک گره برای هر خوشه در شبکه، پیاده‌سازی می‌شود. هر گره، اندازهٔ شباهت را میان مدرک موجود و مرکز ثقلی که خوشه به همراه گره است، به‌طور موازی محاسبه می‌کند.

ابتدا، ضریب شباهت میان مدرک ورودی و مرکز ثقل خوشهٔ موجود محاسبه می‌شود. اگر ضریب شباهت S_1 بزرگ‌تر از آستانهٔ S_{1avg} باشد، در آن صورت گره ورودی فعال می‌شود و سپس، یک حلقهٔ بازگشتی جهت اختصاص مدرک ورودی به خوشه ایجاد می‌شود. گره‌هایی که به مدرک نزدیک نباشند، غیرفعال می‌شوند. در مرحلهٔ دوم، تمام گره‌هایی که به‌عنوان گره برنده انتخاب شده بودند، جهت محاسبهٔ ضریب شباهت انتخاب می‌شوند. تفاوت ضریب شباهت S_2 جهت اطمینان یافتن از این که خوشهٔ برنده شده به مدرک ورودی نزدیک است، محاسبه می‌شود. اگر شباهت گره به خوشه زیاد باشد، آنگاه گره به خوشه اضافه شده و مرکز ثقل روزآمد می‌شود، در غیراین‌صورت، خوشهٔ جدیدی برای مدرک ورودی جدید ساخته می‌شود. با استفاده از این داده‌ها، شبکهٔ عصبی هوشمندی که عمل دسته‌بندی داده‌ها را مطابق با یادگیری بدون سرپرستی انجام می‌دهد، پیاده‌سازی می‌شود. فرایند یادگیری شبکهٔ خودسازمان‌ده را می‌توان به‌عنوان یادگیری رقابتی درنظر گرفت. ایدهٔ اصلی یادگیری رقابتی، تنظیم یک خوشهٔ آزمایشی (C) شبکه با بالاترین سطح فعالیت مطابق با ورودی‌های تصادفی انتخاب شده است. سطح فعال خروجی براساس فاصلهٔ اقلیدسی میان بردار وزن خوشه‌ها m_c و ورودی تعیین می‌شود. مدل فضای برداری برای نمـایش

داده‌ها به وسیلهٔ بردارهایی با وزن W تعیین گردیده، به‌طوری که $\sum (W_i)^2 = 1$ باشد. در مدل فضای برداری، بردار مربوط به هر مدرک دارای n مؤلفه به تعداد اصطلاحات موجود در مجموعه مدرک می‌باشد. مؤلفه‌های بردار، دارای وزن‌هایی است که برای هر اصطلاح در مجموعهٔ مدرک محاسبه می‌شود. به اصطلاحات هر مدرک، براساس فراوانی رخداد اصطلاح در کل مجموعهٔ مدرک و تعداد دفعات حضور اصطلاح در یک مدرک خاص، وزن، اختصاص می‌یابد. شباهت میان دو واژه به‌وسیلهٔ محاسبهٔ اندازه کسینوسی بردارها تعیین شد:

(۲)

$$\frac{(w,v)}{\|W\|^2\|V\|^2} = \frac{(w.v)}{|W||V|} = \frac{(w.v)}{(1)(1)} = \sum W_i V_i$$

در این اندازه‌گیری، زاویهٔ بین دو بردار معین گردید. یک بردار برای هر مدرک، براساس کلیدواژه‌های چهارگانه ساخته شد.

به‌طورکلی، چهار مرحلهٔ زیر جهت فرایند یادگیری در نظر گرفته شد:

۱. انتخاب تصادفی ورودی $x(t)$.

۲. محاسبهٔ فاصلهٔ میان بردارهای وزن و بردار ورودی.

۳. تعیین خوشهٔ برنده.

۴. تنظیم بردارهای وزن در همسایگی خوشهٔ برنده.

نتایج

پیاده‌سازی الگوریتم ارائه شده با استفاده از نرم‌افزار MATLAB ویرایش ۶.۲، صورت گرفت. با استفاده از فرمول ۱، بردار وزن مدارک تعیین شد. از آنجاکه الگوریتم شبکهٔ خودسازمان‌ده براساس فاصلهٔ اقلیدسی استوار است، محدودهٔ اندازهٔ متغیرها در اختصاص متغیر به خوشهٔ خاص، بسیار مهم می‌باشد و معمولاً متغیرها نرمال شده، به‌طوری که هر جزء، دارای واریانس واحدی بوده و با داده‌های نرمال شده، عمل یادگیری انجام می‌پذیرد.

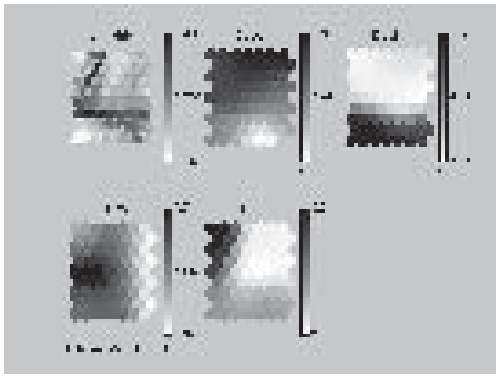
مجموعه داده‌های مورد بررسی، شامل ۱۵۰ نمونه از سه دسته، مربوط به موضوع radiation انتخاب شد که نگاشت آن در شکل ۴ نشان داده شده است. همان‌طور که در این شکل مشاهده می‌شود، محدودهٔ مقدار متغیرهای نرمال شدهٔ هر کلیدواژه در مجموعه مشخص شده است. ماتریس U در این شکل، بیانگر فاصله بین همسایگی‌هاست و ساختار خوشه‌بندی، شبکهٔ خودسازمان‌ده را مشخص می‌کند. برای محاسبهٔ ماتریس U از تمام یا تعدادی از متغیرهای شبکه استفاده می‌شود که در اینجا به علت کم بودن تعداد متغیرها، از تمام مقادیر برای محاسبه استفاده شد. مقادیر بیشتر در این ماتریس بیانگر فاصلهٔ همسایگی بین نگاشت‌هاست و بنابراین محدودهٔ خوشه را بیان می‌کند. معمولاً خوشه‌ها، با مساحت‌های یکنواختی از مقادیر نشان داده شده‌اند. با استفاده از نمودار ستونی، رنگ‌ها مشاهده می‌شود که هر خوشه دارای چه مقداری است.

در شکل ۵، ماتریس U به همراه برچسب آنها نمایش داده می‌شود. هریک از ۱۵۰ نمونهٔ ورودی از سه دستهٔ مربوط به موضوع radiation، دارای برچسب که نشان‌دهندهٔ نوع دسته است، می‌باشد که با استفاده از تابع SOM-autolable که یکی از توابع

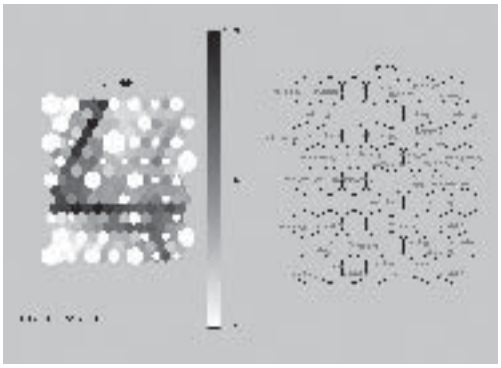
toolbox SOM نرم‌افزار MATLAB است عمل مشخص‌سازی برچسب بر خوشه انجام شد. بهترین خوشه‌ای که با نمونهٔ ورودی مطابقت دارد انتخاب شده و برچسبی به آن اختصاص می‌یابد. این شکل، بیانگر دسته‌بندی خوشه‌هاست.

شکل‌های ۶ و ۷، به نمودار سه بُعدی ماتریس مسافت – که از ماتریس U به‌دست آمده – در فضای سه بُعدی است که محورهای X و Y ابعاد ماتریس و محور Z متوسط فاصله تا خوشه‌های مجاور در نقشه (Map) را نشان می‌دهد. در واقع، شکل ۶، الگویی از شبکهٔ خودسازمان‌ده طراحی شده را با نرون‌های آن در فضای سه بُعدی مشخص ساخته و شکل ۷ شبکهٔ خودسازمان‌ده را به همراه نمایش سه بُعدی داده‌ها نشان می‌دهد.

جهت ارزیابی سیستم، مقادیر دقت بازیافت[19] و صحت بازیافت[20] محاسبه گردید. براساس تعریف، مقدار دقت بازیافت برابر است با نسبت تعداد مدارک مرتبط بازیابی شده به تعداد کل مدارک بازیابی شده و مقدار صحت بازیافت برابر است با تعداد مدارک مرتبط بازیابی شده به تعداد کل مدارک مرتبط در مجموعه (**۳**:۷۰–۷۳). شکل ۸، منحنی صحت بازیافت و دقت بازیافت را نشان می‌دهد. با توجه به شکل و با محاسبهٔ میانگین دقت بازیافت که برابر با مقدار تقریبی ۰/۳۵ است و با توجه به این‌که در سیستم‌هایی که براساس یک مجموعه مدرک استاندارد فعالیت می‌کنند، میانگین دقت را ۰/۲ و ۰/۳ گزارش می‌نمایند (**۲**: ۹۶)، مشخص است که شبکهٔ طراحی شده در بازیابی اطلاعات دارای کارآیی خوبی است. در مقایسه، نمودار قسمت بازیابی و دقت بازیابی به‌دست آمده با نمودارهای مشابه در سیستم‌های بازیابی اطلاعات برمبنای مجموعه مدارک TREC نتیجه می‌شود که رفتار نمودار به‌دست آمده مشابه با رفتار نمودارهای سیستم‌های بهینهٔ بازیابی اطلاعات می‌باشد.



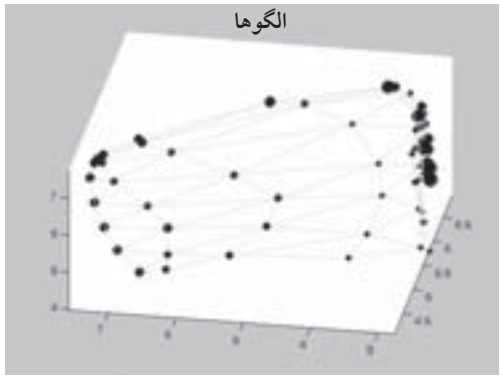
شکل ۶. مقدار متغیرها در هر نگاشت به همراه ماتریس U



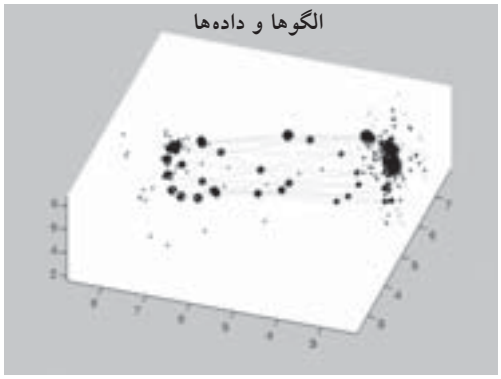
شکل ۵. ماتریس U به همراه نمایش خوشه‌ها

19. Precision

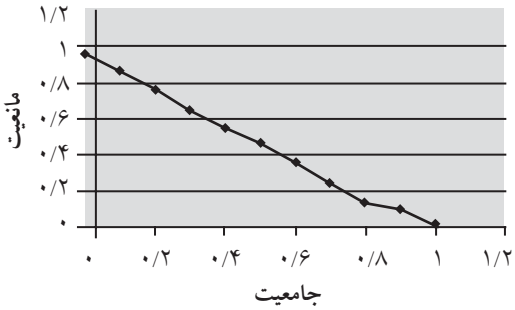
20. Recall



شکل ۶. شبکهٔ خودسازمان‌ده



شکل ۷. شبکهٔ خودسازمان‌ده به همراه داده‌ها



شکل ۸. منحنی صحت بازیافت و دقت بازیافت

بحث و نتیجه‌گیری

در اینجا، با توجه به نوع داده‌های موجود که از نوع متنی می‌باشد به شناخت و بررسی شبکهٔ عصبی مناسب در بازیابی اطلاعات پرداخته شد. شبکه‌های عصبی دیگر نیز بررسی گردید که با توجه به خصوصیت داده‌های متنی و نیز نحوهٔ بازیابی اطلاعات، شبکهٔ خودسازمان‌ده بسیار نزدیک به هدف دیده شد. از ویژگی‌هایی که شبکهٔ خودسازمان‌ده را در بازیابی اطلاعات مؤثر می‌گرداند، می‌توان به موارد زیر اشاره نمود:

۱. در شبکهٔ خودسازمان‌ده گره‌های مشابه در یک فضای معینی گروه‌بندی شده، یعنی داده‌های مشابه به یک گره یا گره‌های همسایه نگاشت می‌گردند.

۲. شبکهٔ خودسازمان‌ده، یک شبکهٔ بدون سرپرستی است. این شبکه، یک راهکار یادگیری را می‌پذیرد که در آن از روابط مشابه میان داده‌ها و خوشه‌ها جهت دسته‌بندی و گروه‌بندی داده‌ها استفاده می‌شود. در شبکه‌های عصبی با سرپرستی، مجموعه داده‌های ورودی و مجموعه داده‌های خروجی (مجموعه مدارکی که باید بازیابی شوند) از قبل معلوم است و شبکهٔ با سرپرست، راهکار یادگیری را می‌پذیرد که در آن دو مجموعهٔ ورودی و خروجی را به یکدیگر متصل می‌سازد.

هدف، شناسایی شبکهٔ عصبی مؤثر جهت بازیابی اطلاعات متنی بود، لذا این شبکه در مقیاسی کوچک، طراحی گردید. در عمل و در مجموعه داده‌های بزرگ، جهت بازیابی کل اطلاعات موجود بایستی مراحل زیر انجام پذیرد:

۱. گروه‌بندی موضوعی مدارک

جهت دسته‌بندی مدارک، معمولاً از سیستم‌های نمایه‌سازی خودکار متن برای تخصیص گروه‌های موضوعی به مدارک متنی استفاده می‌شود. فایدهٔ این دسته‌بندی در آن

است که با اختصاص یک مدرک در یک گروه موضوعی خاص، جستجو محدود شده و سرعت بازیابی اطلاعات بیشتر می‌شود. جهت دسته‌بندی مدارک نیز می‌توان از شبکهٔ عصبی استفاده کرد.

به دلیل این‌که فضای ویژگی‌های مدارک متنی از ابعاد زیادی برخوردار هستند، آموزش شبکهٔ عصبی با اطلاعات خام و ابعاد زیاد، بسیار طولانی و کُند می‌باشد. جهت بهبود این مسئله، پیشنهاد می‌شود که از تکنیک‌های کاهش فضای ویژگی داده‌ها، مانند تکنیک‌های DF[21] و CF-DF[22] و TF × IDF[23] به منظور کاهش ابعاد، جهت دسته‌بندی توسط شبکهٔ عصبی استفاده گردد.

۲. شناسایی ویژگی‌های مدارک متنی شامل مراحل زیر است :

- استخراج کلمات
- حذف واژه‌های غیرمجاز[24]
- در نظر گرفتن ریشهٔ کلمات
- محاسبهٔ بردار وزنی کلمات
- تهیهٔ فرهنگ لغات مناسب

یکی از عوامل بهبود کارآیی شبکه، انتخاب ویژگی[25] می‌باشد که پیش از اعمال الگوریتم‌های یادگیری در مجموعه مدارک بزرگ، در نظر گرفته می‌شود.

۳. محاسبهٔ بردار ویژگی ورودی برای هر مدرک متنی

۴. به‌کارگیری الگوریتم ارائه شده در شکل، جهت یادگیری شبکهٔ عصبی خودسازمان‌ده

با در نظر گرفتن خصوصیات شبکه‌های عصبی، به‌نظر می‌رسد که این تکنیکِ هوش مصنوعی جهت بازیابی اطلاعاتِ مؤثر باشد. در حال حاضر، در مدل‌های کاربردی شبکه‌های عصبی در بازیابی اطلاعات تحقیقاتی صورت گرفته است و در آینده با پیشرفت بیشتر سخت‌افزار و نرم‌افزار، به نظر می‌رسد که به‌سرعت بتوان از شبکه‌های عصبی استفاده‌های مؤثرتری نمود. در آینده با ارزان‌تر شدن سخت‌افزار موردنیاز مدل‌های شبکهٔ عصبی، امکان استفاده از این شبکه، جهت طبقه‌بندی موازی مدارک امکان‌پذیر گشته و بازیابی اطلاعات بسیار سریع‌تر خواهد بود. همچنین، با پیشرفت‌های نرم‌افزاری در ایجاد روش‌های جدید کنترل توابع شبکه، به منظور پیاده‌سازی مدل‌های شبکهٔ عصبی، تحولی مهم در سرعت بازیابی اطلاعات فراهم می‌شود.

---

۲۱. Document Frequency Method: تکنیک گزینش خصیصه‌ها در گروه‌بندی متن است. با استفاده از مجموعه مدارک آموزشی و برچسب‌های گروه‌های متناظر آنها، از روش DF به منظور کاهش اندازهٔ واژگان توسط گزینش اصطلاحات براساس تکنیک رتبه‌بندی اصطلاح استفاده می‌شود.

۲۲. Category Frequency-Document Frequency Method: در این روش، دو مرحلهٔ پردازش وجود دارد. ابتدا، آستانهٔ t بر روی بسامدهای گروه‌بندی اصطلاحات تعریف می‌شود، به‌طوری که اصطلاح تنها در صورتی که بسامد آن گروه کمتر از آستانه باشد، بازیابی می‌شود. در مرحلهٔ دوم، از روش DF برای گزینش اصطلاحات بیشتر استفاده می‌گردد.

۲۳. Term Frequency and Inverse Document Frequency Method: در روش TF×IDF اصطلاحات براساس مقادیر TF×IDF آنها رتبه‌بندی شده و به این صورت اصطلاحاتی که دارای بالاترین مقدار TF×IDF هستند، برای تشکیل مجموعهٔ کاهش خصیصه انتخاب می‌شود.

24. Stop word

25. feature selection

منابع

۱. کلینی، سارا. «شبیه‌سازی نظام اطلاع‌رسانی فارسی کتابخانهٔ منطقه‌ای علوم و تکنولوژی شیراز با استفاده از شبکهٔ عصبی». ***فصلنامهٔ کتاب،*** دورهٔ سیزدهم، ۱ (بهار ۱۳۸۱): ۱۰۴-۱۱۲.

۲. گروسمن، دیوید. آ. ***بازیابی اطلاعات: الگوریتم‌ها و روش‌های اکتشافی.*** ترجمهٔ جعفر مهراد و سارا کلینی. مشهد: کتابخانهٔ رایانه‌ای، ۱۳۸۴.

۳. لانکاستر، اف. ویلفرید. ***نظام‌های بازیابی اطلاعات ویژگی‌ها، آزمون و ارزیابی.*** ترجمهٔ جعفر مهراد. شیراز: نوید، ۱۳۷۹.

4. Doszkocs, T. E. "Connectionist models and information retrieval". ***Annual Review of Information Science & Technology,*** Vol 25 (1990): 209-260.

5. Crestani,F.A. "Model for adaptive information retrieval". ***Transactions on Knowledge and Data Engineering,*** Vol.13, No.5 (Sep.2001): 846-850.

6. Grunfeld,L."Routing retrieval and filtering experiments using PIRCS", Text Retrieval Conference (Gaithersburg, Maryland,USA,1995).[on-line].Available: http://citeseer,ist.psu.edu/156569. html.

7. Hatano, K. "A som-based information organizer for text and video data". Proceeding of the Fifth International Conference on Database System for Advanced Applications (Melbourn, Australia, 1997). [on-line]. Available: http: citeseer.ist.psu.edu/ hatano97sombased.html.

8. Kohonen, T. "Self-organized formation of topologically correct feature maps". ***Biological Cybernetics,*** No.43 (1982): 59-69. Quoted in Fausett, Laurene. ***Fundamentals of neural networks : architectures, algorithms and applications.*** USA: Prentice-Hall, Inc, 1994.

9. Mandl, T. " Efficient preprocessing for information retrieval with neural network". EUFIT 7th European Congress on Intelligent Techniques and soft Computing Aachen, (Germany, 1999). [on-line]. Available: http://www.uni-hildesheim.de/~mandl/publikationen/MANDL_EUFIT99.

تاریخ دریافت: ۱۳۸۵/۹/۲۹