

بهینه‌سازی درخواست کاربر مبتنی بر هوشمندسازی بازیابی اطلاعات بوسیله شبکه عصبی

نوشته: محمدباقر دستغیب*

کلیدواژه‌ها: بازیابی اطلاعات متنی/ شبکه عصبی/ بازیابی اطلاعات هوشمند/ بهینه‌سازی درخواست

چکیده

امروزه استفاده از کامپیوتر برای رده‌بندی و ذخیره اطلاعات مرسوم شده است. با توجه به انبوه اطلاعات موجود در شبکه‌ها از جمله اینترنت، و ساختمان نبودن اطلاعات، نیاز به بازیابی خودکار اطلاعات بیشتر گردیده است. با توجه به تنوع اطلاعات موجود در شبکه و ناهمگن بودن مدارک ایجاد درخواست برای کاربران ساده نیست و در بسیاری از موارد نیازمند اصلاح بوسیله شخص خیره می‌باشد. منظور از هوشمندسازی سیستم بازیابی اطلاعات، سیستمی است که محتوای مدارک و درخواست را درک کند، و با توجه به دانش زمینه، کاربر را در یافتن اطلاعات موردنیاز، راهنمایی نماید. در بسیاری از سیستم‌های تجاری، فهرست‌گذاری موضوعی، انجام شده است. یکی از کاربردهای چنین سیستمی استفاده از فهرست موضوعی و آموزش شبکه عصبی برای بهینه‌سازی درخواست کاربر می‌باشد. اصول این سیستم شباهت جواب درخواست‌های مشابه است، بنابراین سیستم هوشمند بردار درخواست کاربر را طوری تغییر می‌دهد تا با دانش موجود در مجموعه بهترین جواب بدست آید.

مقدمه

سیستم بازیابی یک ابزار محاسباتی است که اطلاعات را به شکلی پیاده‌سازی^۱ می‌کند که بعداً بتواند بطور خودکار بازیابی شوند. سیستم‌های بازیابی اطلاعات غالباً، فقط اطلاعات متنی را ذخیره و بازیابی می‌کنند. ولی این فرآیند به علت حجم بالای اطلاعات (معمولاً از صدها تا هزاران مدارک^۲) و ساختمان نبودن مدارک، کار پیچیده و دشواری است.

مدرك مجموعه‌اي از كلمات و جمله‌ها است که درباره موضوع خاصی به بحث می‌پردازد [۱].

کاربر برای دسترسی به اطلاعات و مدارک موردنظر خود، یک درخواست را مطرح می‌کند، و سیستم تمام مدارک‌های شبیه به این درخواست را بازیابی می‌کند. برای جستجو، باید مدارک برای مقایسه با درخواست، شاخص‌گذاری^۳ شود. شاخص‌گذاری عبارت از استخراج کلمات کلیدی متن و ذخیره کردن آنها با قالب‌بندی^۴ مشخص است.

برای آنکه مدارک و درخواست‌ها را بتوان ذخیره و پردازش کرد، باید روشی جهت پیاده‌سازی اطلاعات مدارک، انتخاب گردد. یکی از روش‌های مرسوم در سیستم‌های بازیابی اطلاعات، روش برداری^۵ است. در این روش مدارک و درخواست بصورت بردار ذخیره می‌گردند. اجزای بردارها، کلمات موجود در شاخص است که بصورت عددی براساس فرمول‌های وزن‌دهی^۶ محاسبه گردیده است [۱، ۲].

مجموعه مدارک در نهایت یک ماتریس به نام ماتریس کلمه-مدرك (شکل ۱) از وزنها را ایجاد خواهد کرد که هر سطر ماتریس یک بردار است. و هر ستون از این ماتریس، وزن یکی از کلمات شاخص را در مدارک موجود در مجموعه نشان می‌دهد. درخواست نیز جهت پردازش در چنین سیستمی، به یک بردار از وزن کلمه‌های شاخص تبدیل می‌گردد تا توسط موتور جستجو^۷، پردازش گردد [۲].

حالت خاصی از مدل برداری، مدل منطقی^۸ است که در آن هر عضو ماتریس یک مقدار منطقی (صفر یا یک) است. در این مدل وجود، و یا عدم وجود کلمه شاخص در مدرک مشخص می‌گردد. این روش

* عضو هیئت علمی کتابخانه منطقه‌ای علوم و تکنولوژی شیراز

نخواهد بود. بايد دقت داشت كه كاربران سيستمهاي بازيايي اطلاعات، هميشه افراد خبره^{۱۱} نيستند بنابراین سيستم بايد بتواند درخواستهاي ضعيف را با جاگزين كردن كلمات كليدي تقويت كند و كاربر را در جهت ساخت درخواست مناسب راهنمايي نمايد.

برخي از سيستمهاي هوشمند بازيايي اطلاعات، سعي بر آن دارند كه محتواي مدرك و درخواست را درك نمايند و يك رابطه ميان درخواست و مدارك بوجود آورند. به عنوان مثال انتظار داريم سيستم هوشمند، درخواستهاي "Clever Man" و "Bright Person" را يكسان بشمارد، و جوابهاي يكسان براي آن استخراج نمايد. اين امر ميسر نخواهد شد، مگر آنكه ميان محتواي كلمات كليدي و مدارك، ارتباط منطقي بوجود آيد. در سيستم فعلي از اين تئوري كه درخواستهاي مشابه داراي جوابهاي مشابه هستند استفاده خواهد شد.

در ادامه ابتدا مباحث مرتبط (بخش ۲) را بررسي خواهيم نمود، سپس يك روش هوشمند مبتني بر شبكه عصبي (بخش ۳) مورد بررسي قرار خواهد گرفت.

مباحث مرتبط

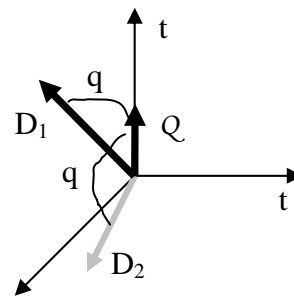
جهت بالا بردن كارآيي سيستمهاي بازيايي اطلاعات، تلاشهاي بسياري انجام شده است، در زمينه ترميم^{۱۲} بردار درخواست، روشهاي مبتني بر سيستمهاي فزي^{۱۳} و همچنين سيستمهاي داراي بازخورد^{۱۴} از انتخاب كاربر مورد استفاده قرار مي‌گيرد. در سيستمهاي فزي، مجموعه قوانين^{۱۵}، به سيستم امکان انتخاب با عدم قطعيت مي‌دهد. در چنين سيستمي قوانين در ابتدا استخراج مي‌گردد و سپس در طول كار سيستم با توجه به بازخورد كه از كاربر گرفته مي‌شود، تغيير خواهد كرد. در اين مقاله صرفاً کاربرد شبكه عصبي^{۱۶} در هوشمندسازي سيستم بازيايي اطلاعات مورد بررسي قرار مي‌گيرد [۵، ۶].

بدليل سهولت محاسبه در بسياري از سيستمهاي تجاري مورد استفاده قرار مي‌گيرد.

$$\begin{matrix} \text{Doc}_1 & \left(\begin{matrix} \text{term}_{11} & \text{term}_{12} & \dots & \text{term}_{1i} \\ \text{term}_{22} & \text{term}_{21} & \dots & \text{term}_{2i} \\ \dots & \dots & \dots & \dots \\ \text{Doc}_{j1} & \text{term}_{j2} & \text{term}_{j3} & \dots & \text{term}_{ji} \end{matrix} \right) \end{matrix}$$

شكل ۱- ماتريس كلمه مدرك در مدل برداري

پس از آماده‌سازي بردار درخواست، سيستم بازيايي اطلاعات با بكارگيري يك معيار مقايسه، بردار درخواست و بردارهاي مدارك را مقايسه مي‌نمايد و نتيجه يك ليست ارزش‌گذاري^{۱۷} شده از مدارك شبیه به درخواست، بصورت نزولي براساس درجه شباهت خواهد بود. معيارهاي مختلفي براي محاسبه شباهت مورد استفاده قرار مي‌گيرد كه ساده‌ترين آنها زاويه ميان دوبردار است، بدین معني كه هرچه زاويه ميان بردارها (شكل ۲) كمتر باشد، بردارها شبیه‌ترند. بنابراین مي‌توان كسينوس زاويه ميان دو بردار را محاسبه نمود و هرچه كسينوس به يك نزديكتر باشد دو مدرك شبیه‌ترند [۳].



شكل ۲- اندازه‌گيري زاويه ميان بردارها

اين روش در بسياري از سيستمهاي موجود کاربرد دارد، عيب اين روش آنست كه بسيار به بردار درخواست وابسته است، به عبارت ديگر اگر بردار درخواست بخوبي بيان نشده باشد، آنگاه جوابهاي سيستم بازيايي اطلاعات از دقت خوبي برخوردار

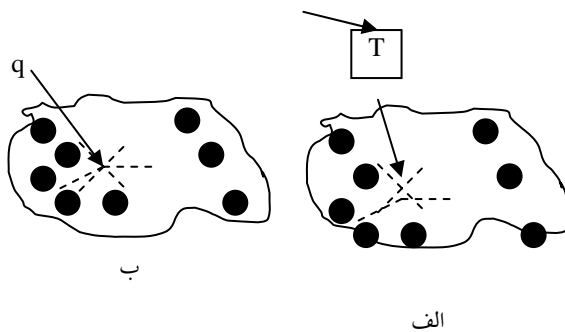
هوشمندسازی سیستم بازیابی اطلاعات

هدف این روش، تصحیح^{۱۷} بردار درخواست کاربر، با توجه دانش محلی^{۱۸} موجود در سیستم بازیابی اطلاعات است. شبکه عصبی را می‌توان یک تابع غیرخطی^{۱۹} دانست که وظیفه آن درونیابی^{۲۰} و یا برونیابی^{۲۱} است. این تابع می‌تواند با توجه به دانشی که در مرحله آموزش^{۲۲} کسب نموده است، خروجی قابل قبولی در دامنه^{۲۳} ورودی مجاز داشته باشد. به عنوان مثال می‌توان یک نقطه در درون و یا بیرون نقاطی که در مرحله آموزش به شبکه داده شده است، محاسبه نمود. این نقطه با توجه به دانش موجود توسط تابع غیرخطی شبکه عصبی تخمین زده می‌شود. با توجه به عدم قطعیت^{۲۴} و ابهام^{۲۵} ذاتی موجود در سیستم‌های بازیابی اطلاعات استفاده از سیستمی که با بهره‌گیری از دانش زمینه بتواند کاربر را در ساخت درخواست مناسب، راهنمایی نماید، ضروری به نظر می‌رسد. در حقیقت این سیستم مانند یک ناظر خبره، بر درخواست‌های رسیده از کاربران نظارت می‌نماید و در صورت نیاز، با تصحیح بردار درخواست، کاربر را در بدست آوردن نتیجه مطلوب راهنمایی می‌نماید [۸، ۱۲، ۱۳].

مطالعات اخیر در زمینه هوشمندسازی بازیابی اطلاعات، به این نتیجه رسیده است که برای بهبود کارایی سیستم بازیابی اطلاعات، احتیاج به تکنیک‌هایی است که محتوای درخواست‌ها و مدارک را درک کنند [۸]. اخیراً محققان تئوری اطلاعات سعی بر این داشتند که رابطه میان مدارک و درخواست‌ها را مشخص کنند [۷، ۹، ۱۰، ۱۱]. هدف این است که درخواست کاربر طوری تطبیق^{۲۶} پیدا کند که اطلاعات مورد درخواست کاربر را در مجموعه محلی مدارک پیاده‌سازی نماید.

پایه و اساس تطبیق درخواست این است که درخواست‌های مشابه دارای مجموعه مدرک‌های مشابه هستند. با استفاده از اطلاعات مدرک‌هایی که با درخواست‌های قبلی مشابه بوده‌اند، می‌توان مدارک مشابه با درخواست‌های جدید را بدست آورد. تغییر شکل درخواست همانند شخص خبره عمل

می‌کند [۱۲، ۱۳]. به عبارت دیگر سیستم ناظر شبکه عصبی حضور شخص خبره را شبیه‌سازی می‌کند. در شکل ۳ مدل کلاسیک (شکل ۳-ب)، با مدل هوشمند (شکل ۳-الف) مقایسه شده است. سیستم هوشمند دارای مدل درخواست T می‌باشد که با توجه به دانش مجموعه، درخواست را بازسازی می‌کند.



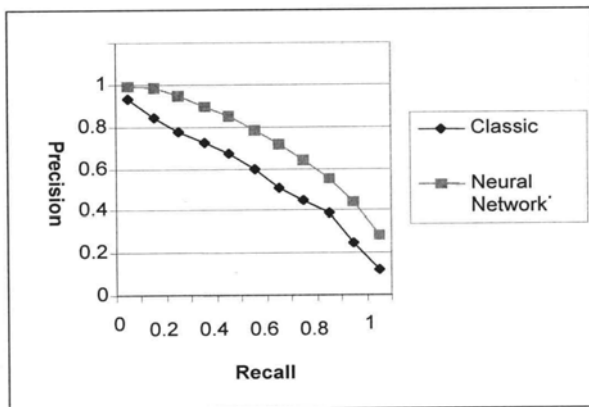
شکل ۳- مقایسه مدل هوشمند و غیرهوشمند برداری

بازیابی اطلاعات

کاربرد این سیستم در مجموعه مدارکی که دسته‌بندی^{۲۷} شده باشند، بهتر نمایان می‌گردد. بدین صورت که مثال‌های آموزشی^{۲۸}، براحتی و بطور خودکار، از شاخه^{۲۹}‌های موجود در مجموعه استخراج می‌گردد. هر مثال آموزشی شامل چند کلمه کلیدی (درخواست) و مجموعه مدارک مرتبط با کلمات کلیدی است.

سیستم هوشمند در دو فاز عملیات بازیابی را انجام می‌دهد. ابتدا مرحله یادگیری و آموزش ماشین^{۳۰} است، در این مرحله باید یک لیست از درخواست‌ها (بردارهای درخواست) و جواب آنها (ماتریس مدارک جواب درخواست) به سیستم داده شود. در این مرحله سیستم شبکه عصبی دانش زمینه‌ای مجموعه را کسب می‌نماید. فاز دوم، فاز بکارگیری و آزمایش سیستم هوشمند است، در این فاز سیستم هوشمند مانند ناظر، درخواست‌های کاربر را پذیرفته و آنها را بهینه‌سازی می‌کند و سپس سیستم کلاسیک مانند قبل، بروی درخواست تغییر یافته، عملیات محاسبه شباهت را انجام می‌دهد.

مي‌توان، مدارك مجموعه را دسته‌بندي نمود و سپس از هر دسته مدارك شبيه، يك نماينده^{۲۵} كه عموميت بيشتري دارد در آموزش شبكه عصبي شركت كند. در شكل ۵ نتيجه آزمايش اين روش بروي مجموعه مدارك CranField مشاهده مي‌گردد، در اين نمودار نتيجه روش كلاسيك با روش هوشمند مقايسه مي‌گردد. اين مجموعه داراي ۱۴۰۰ مدرك و ۲۲۵ مثال آموزشي است. تعداد كلمات كليدي كه در بيش از يك مدرك ظاهر شده‌اند حدود ۴۴۰۰ كلمه مي‌باشد. در عمل براي آموزش شبكه عصبي مي‌توان از فهرست‌هاي موضوعي بيشتري بهره را براي، آموزش شبكه عصبي بدست آورد. بدليل دسته‌بندي اطلاعات در اين فهرستها، بهترين جواب در آموزش سيستم بدست خواهد آمد.



شكل ۵- مقايسه نتيجه بازباني كلاسيك و هوشمند

نتيجه‌گيري

با مشاهده خروجي سيستم هوشمند به اين نتيجه مي‌رسيم كه سيستم هوشمند داراي كارايي بالاترين نسبت به سيستم كلاسيك مي‌باشد. اين نتيجه با نظارت بر درخواست کاربر جواب بهتري را فراهم آورده است. زيرا بردار درخواست با دانش زمينه تطبيق داده شده و بهينه‌سازي مي‌گردد، به عبارت ديگر سيستم هوشمند با درك معنای درخواست، در صورت نياز آن را بهينه‌سازي مي‌نمايد. مدل هوشمند برخي مشكلات مدلهای كلاسيك را حل کرده است:

شكل ۴ سيستم هوشمند را در دو فاز يادگيري و يكارگيري نشان مي‌دهد [۱۴، ۱۵].

همانطور كه مشاهده مي‌گردد سيستم از

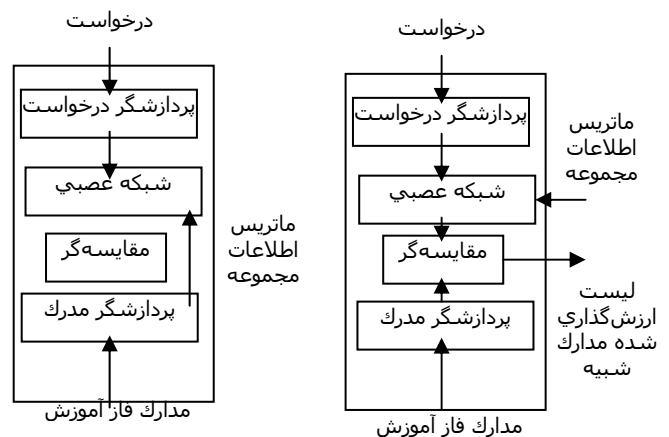
چهار قسمت تشكيل شده است:

۱- پردازشگر درخواست^{۲۱}: در اين قسمت از سيستم، درخواست پردازش مي‌گردد تا به بردار تبديل گردد. در اين مرحله از شاخص و وزن‌دهي استفاده خواهد شد و يا ممكن است براي سرعت بيشتر از مدل منطقي استفاده شود. بنا بر اين خروجي اين مرحله بردار درخواست است.

۲- پردازشگر مدرك^{۲۲}: اين قطعه از سيستم، مدارك را مورد پردازش قرار مي‌دهد و براي هر مدرك يك بردار از وزنها، ايجاد مي‌نمايد بنا بر اين خروجي اين قطعه از سيستم، ماتريس كلمه-مدرك مي‌باشد.

۳- مقايسه‌گر^{۲۳}: اين قطعه از سيستم، بردار درخواست را با تمام بردارهاي مدارك مقايسه مي‌نمايد، و يك ليست ارزش‌گذاري شده از مدارك شبيه را تهيه نموده به کاربر ارايه مي‌نمايد [۱۴، ۱۵].

۴- شبكه عصبي: وظيفه شبكه عصبي تغيير بردار درخواست کاربر با توجه به دانش كسب شده، در مرحله آموزش مي‌باشد. اين بردار به عنوان خروجي اين مرحله به مقايسه‌گر داده مي‌شود.



شكل ۴- سيستم بازباني اطلاعات هوشمند در فاز آموزش و يكارگيري

براي آنكه سيستم هوشمند بتواند بخوبي عموميت بخشي^{۲۴} را در دانش مجموعه ايجاد نمايد، مثالهايي كه جهت آموزش سيستم انتخاب مي‌گردد، بايد از تامامي دامنه مجموعه باشد. براي كارايي بهتر

دیگر در صورتی از سیستم هوشمند استفاده شود، که درخواست دارای مثال‌های مشابهی در زمان آموزش باشد، در غیر اینصورت از سیستم کلاسیک بدون تغییر درخواست، استفاده گردد و بازخورد این درخواست مجدداً سیستم را تعلیم دهد. یعنی با توجه به انتخاب کاربر، می‌توان با مجموعه‌ای از مثال‌های آموزشی سیستم را مجدداً تعلیم داد.

• الزامی ندارد که درخواست ساختاری^{۳۶} مانند مدرک داشته باشد تابع مدل T (شکل ۳) درخواست را پیکربندی^{۳۷} می‌کند، تا شباهت قابل محاسبه و سنجش باشد.

• کاربر ملزم نیست که درخواست خود را به طور کامل، از محتوایی که می‌خواهد بیان کند، تابع مدل T، با استفاده از دانش محیط، درخواست را تغییر شکل خواهد داد، و درخواست را در فضای مدارک قرار خواهد داد.

• یک مدل هوشمند می‌تواند برای محاسبه شباهت استفاده شود، که رابطه میان درخواستها با مدرکهای مشابه را با استفاده از بازخورد کاربر مورد محاسبه قرار دهد. موقعیت مدرک در فضای مدارک، نسبت به تصمیم کاربر تغییر خواهد کرد. باید توجه کرد که مدل هوشمند، در صورتی پاسخ مناسب و صحیح خواه داد که در فاز آموزش با مثال‌های جامع، یادگیری انجام شده باشد در غیر اینصورت ممکن است نتیجه مناسبی حاصل نگردد بنابراین پیشنهاد می‌گردد، که سیستم هوشمند در صورتی مورد استفاده قرار گیرد، که درخواست رسیده دارای تاریخچه‌ای در زمان آموزش باشد، به عبارت

پی‌نوشت‌ها:

1. Implement
- 2.Document
- 3.Query
- 4.Indexing
- 5.Format
- 6.Vector Space Model
- 7.Term Weighting
- 8.Search Engine
- 9.Boolean Model
- 10.Rank
- 11.Expert in domain
- 12.Enhancing
- 13.Fuzzy Logic Systems
- 14.Feedback
- 15.Rule base
- 16.Neural Network
- 17.Enhancing

- 18.Background Knowledge
- 19.Non linear Function
- 20.Interpolation
- 21.Extrapolation
- 22.Learning
- 23.Domain of input data (range of input)
- 24.Uncertainty
- 25.Vagueness
- 26.Adapt
- 27.Clustering
- 28.Learning Samples
- 29.Directroy
- 30.Machine learning
- 31.Query processor
- 32.Document Processor
- 33.Matcher
- 34.Generalization

- 35.Cluster Center
- 36.Structure
- 37.Configure

منابع

- G. Salton. (1989) "*Automatic text processing: the transformation, analysis and retrieval of information by computer*". Addison Wesley.
- G. Salton. And McGill. (1983) "*Introduction to modern information retrieval*", New York, Mc-GrawHill.
- E. Chisholm. (1995) "*New term weighing formulas for vector space method in information retrieval*", New York.
- L. A. Zadeh, (1996) "*Fuzzy Logic=Computing with Words*", IEEE Transactions of Fuzzy Systems, Vol.4, No.2, pp.103-111, May.
- G. Salton and B. Buckley, (1988) "*Term weighting approaches in automatic text retrieval*", IPM.
- K. Sparck, (1972) "*A statistical interpretation of term specificity and its application in retrieval*", Documentation.
- R.K. Belew, (1989) "*Adaptive information retrieval: using a connectionist representation to retrieve and learn about documents*", USA, June.
- W.B. Croft. (1987) "*Approaches to intelligent information retrieval*", IPM.
- R.J. Brachman and D.L. McGuinness,(1988) "*Knowledge representation, Connectionism, Conceptual Retrieval*", ACM SIGIR, France, June.
- K.L. Kwok, (1990) "*Application of neural network to information retrieval*", IEEE, P.623-626, USA.
- J.C. Scoltes, (1991) "*Neural nets and their relevance for information retrieval*", Technical Report, Amsterdam.
- K.J. Schmucher, Fuzzy set, (1990) "*Natural Language Computations, and Risk Analysis*", W.H. Freeman and Company, translated by T.Onisawa, Keigaku Shuppan.
- L.A. Zadeh, (1975) "*The concept of Linguistic Variable and its Application to Approximate Reasoning (Part 1)*", Information Sciences, 8,pp.199-249.
- R.S. Michalski, J.G. Carbonell, T.M. Mitchell (Eds.), (1983) "*Machine Learning: An Artificial Intelligence Approach*", Springer-Verlag.
- R.S. Michalski, J.G. Carbonell, T.M. Mitchell (Eds.), (1986) "*Machine Learning: An Artificial Intelligence Approach*", Vol. II, Morgan Kaufman.