بهينه‌سازي درخواست كاربر مبتني بر هوشمندسازي بازيابي اطلاعات بوسيله شبكه عصبي

نوشته: محمدباقر دستغيب*

كليدواژه‌ها: بازيابي اطلاعات متني/ شبكه عصبي/ بازيابي اطلاعات هوشمند/ بهينه‌سازي درخواست

چكيده

امروزه استفاده از كامپيوتر براي رده‌بندي و ذخيره اطلاعات مرسوم شده است. با توجه به انبوه اطلاعات موجود در شبكه‌ها از جمله اينترنت، و ساختمند نبودن اطلاعات، نياز به بازيابي خودكار اطلاعات بيشتر گرديده است. با توجه به تنوع اطلاعات موجود در شبكه و ناهمگن بودن مدارك ايجاد درخواست براي كاربران ساده نيست و در بسياري از موارد نيازمند اصلاح بوسيله شخص خبره مي‌باشد. منظور از هوشمندسازي سيستم بازيابي اطلاعات، سيستمي است كه محتواي مدرك و درخواست را درك كند. و با توجه به دانش زمينه، كاربر را در يافتن اطلاعات موردنياز، راهنمايي نمايد. در بسياري از سيستمهاي تجاري، فهرست‌گذاري موضوعي، انجام شده است. يكي از كاربردهاي چنين سيستمي استفاده از فهرست موضوعي و آموزش شبكه عصبي براي بهينه‌سازي درخواست كاربر مي‌باشد. اصول اين سيستم شباهت جواب درخواستهاي مشابه است، بنابراين سيستم هوشمند بردار درخواست كاربر را طوري تغيير مي‌دهد تا با دانش موجود در مجموعه بهترين جواب بدست آيد.

مقدمه

سيستم بازيابي يك ابزار محاسباتي است كه اطلاعات را به شكلي پياده‌سازي[1] مي‌كند كه بعداً بتواند بطور خودكار بازيابي شوند. سيستمهاي بازيابي اطلاعات غالباً، فقط اطلاعات متني را ذخيره و بازيابي مي‌كنند. ولي اين فرآيند به علت حجم بالاي اطلاعات (معمولاً از صدها تا هزاران مدرك[2]) و ساختمند نبودن مدارك، كار پيچيده و دشواري است.

مدرك مجموعه‌اي از كلمات و جمله‌ها است كه درباره موضوع خاصي به بحث مي‌پردازد[1] .

كاربر براي دسترسي به اطلاعات و مدارك موردنظر خود، يك درخواست را مطرح مي‌كند، و سيستم تمام مدركهاي شبيه به اين درخواست را بازيابي مي‌كند. براي جستجو، بايد مدارك براي مقايسه با درخواست، شاخص‌گذاري[4] شود. شاخص‌گذاري عبارت از استخراج كلمات كليدي متن و ذخيره كردن آنها با قالب‌بندي[5] مشخص است.

براي آنكه مدارك و درخواستها را بتوان ذخيره و پردازش كرد، بايد روشي جهت پياده‌سازي اطلاعات مدارك، انتخاب گردد. يكي از روشهاي مرسوم در سيستمهاي بازيابي اطلاعات، روش‌برداري[6] است. در اين روش مدرك و درخواست بصورت بردار ذخيره مي‌گردند. اجزاي بردارها، كلمات موجود در شاخص است كه بصورت عددي براساس فرمولهاي وزن‌دهي[7] محاسبه گرديده است[1، 2].

مجموعه مدارك در نهايت يك ماتريس به نام ماتريس كلمه-مدرك (شكل ۱) از وزنها را ايجاد خواهد كرد كه هر سطر ماتريس يك بردار است. و هر ستون از اين ماتريس، وزن يكي از كلمات شاخص را در مدارك موجود در مجموعه نشان مي‌دهد. درخواست نيز جهت پردازش در چنين سيستمي، به يك بردار از وزن كلمه‌هاي شاخص تبديل مي‌گردد تا توسط موتور جستجو[8]، پردازش گردد[2].

حالت خاصي از مدل برداري، مدل منطقي[9] است كه در آن هر عضو ماتريس يك مقدار منطقي (صفر يا يك) است. در اين مدل وجود، و يا عدم وجود كلمه شاخص در مدرك مشخص مي‌گردد. اين روش

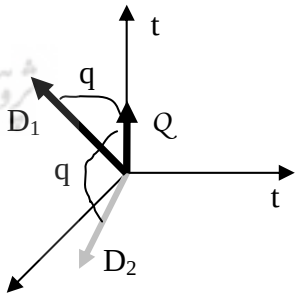
* عضو هيئت علمي كتابخانة منطقه‌اي علوم و تكنولوژي شيراز

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

$$\begin{array}{l} Doc_1 \\ Doc_{21} \\ . \\ Doc_{j1} \end{array} \begin{bmatrix} term_{11} & term_{12} & \dots & term_{1i} \\ term_{22} & term_{11} & \dots & term_{2i} \\ . & . & . & . \\ term_{j2} & term_{j3} & \dots & term_{ji} \end{bmatrix}$$

شكل ١- ماتريس كلمه مدرك در مدل برداري

پس از آماده‌سازي بردار درخواست، سيستم بازيابي اطلاعات با بكارگيري يك معيار مقايسه، بردار درخواست و بردارهاي مدارك را مقايسه مي‌نمايد و نتيجه يك ليست ارزش‌گذاري[١٠] شده از مدارك شبيه 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، بدين معني كه هرچه زاويه ميان بردارها (شكل٢) كمتر باشد، بردارها شبيه‌ترند. بنابراين مي‌توان كسينوس زاويه ميان دو بردار را محاسبه نمود و هرچه كسيسنوس به يك نزديكتر باشد دو مدرك شبيه‌ترند [٣].



شكل٢- اندازه‌گيري زاويه ميان بردارها

اين روش در بسياري از سيستمهاي موجود كاربرد 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 بازيابي اطلاعات از دقت خوبي برخوردار نخواهد بود. بايد دقت داشت كه كاربران سيستمهاي بازيابي اطلاعات، هميشه افراد خبره[١١] نيستند بنابراين سيستم بايد بتواند درخواستهاي ضعيف را با جايگزين كردن كلمات كليدي تقويت كند و كاربر را در جهت ساخت درخواست مناسب راهنمايي نمايد.

برخي از سيستمهاي هوشمند بازيابي 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 ٢) را بررسي خواهيم نمود، سپس يك روش هوشمند مبتني بر شبكه عصبي (بخش ٣) مورد بررسي قرار خواهد گرفت.

مباحث مرتبط

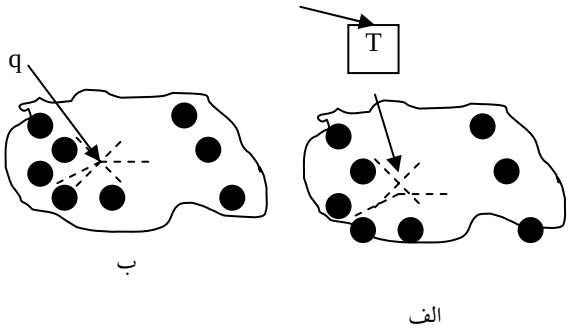
جهت بالا بردن كارآيي سيستمهاي بازيابي اطلاعات، تلاش‌هاي بسياري انجام شده است، در زمينه ترميم[١٢] بردار درخواست، روشهايي مبتني بر سيستمهاي فازي[١٣] و همچنين سيستمهماي داراي بازخورد[١٤] از انتخاب كاربر مورد استفاده قرار مي‌گيرد. در سيستمهاي فازي، مجموعه قوانين[١٥]، به سيستم امكان انتخاب با عدم قطعيت مي‌دهد. در چنين سيستمي قوانين در ابتدا استخراج مي‌گردد و سپس در طول كار سيستم با توجه به بازخوردي كه از كاربر گرفته مي‌شود، تغيير خواهد كرد. در اين مقاله صرفاً كاربرد شبكه عصبي[١٦] در هوشمندسازي سيستم بازيابي اطلاعات مورد بررسي قرار مي‌گيرد[٤، ٥، ٦].

هوشمندسازي سيستم بازيابي اطلاعات

هدف اين روش، تصحيح[17] بردار درخواست كاربر، با توجه دانش محلي[18] موجود در سيستم بازيابي اطلاعات است. شبكه عصبي را مي‌توان يك تابع غيرخطي[19] دانست كه وظيفه آن درونيابي[20] و يا برونيابي[21] است. اين تابع مي‌تواند با توجه به دانشي كه در مرحله آموزش[22] كسب نموده است، خروجي قابل قبولي در دامنه[23] ورودي مجاز داشته باشد. به عنوان مثال مي‌توان يك نقطه در درون و يا بيرون نقاطي كه در مرحله آموزش به شبكه داده شده است، محاسبه نمود. اين نقطه با توجه به دانش موجود توسط تابع غيرخطي شبكه عصبي تخمين زده مي‌شود. با توجه به عدم قطعيت[24] و ابهام[25] ذاتي موجود در سيستمهاي بازيابي اطلاعات استفاده از سيستمي كه با بهره‌گيري از دانش زمينه بتواند كاربر را در ساخت درخواست مناسب، راهنمايي نمايد، ضروري به نظر مي‌رسد. در حقيقت اين سيستم مانند يك ناظر خبره، بر درخواستهاي رسيده از كاربران نظارت مي‌نمايد و در صورت نياز، با تصحيح بردار درخواست، كاربر را در بدست آوردن نتيجه مطلوب راهنمايي مي‌نمايد [12، 13، 8].

مطالعات اخير در زمينه هوشمندسازي بازيابي اطلاعات، به اين نتيجه رسيده است كه براي بهبود كارآيي سيستم بازيابي اطلاعات، احتياج به تكنيكهايي است كه محتواي درخواستها و مدارك را درك كنند[8]. اخيراً محققان تئوري اطلاعات سعي بر اين داشتند كه رابطه ميان مدارك و درخواستها را مشخص كنند [7، 9، 10، 11]. هدف اين است كه درخواست كاربر طوري تطبيق[26] پيدا كند كه اطلاعات مورد درخواست كاربر را در مجموعه محلي مدارك پياده‌سازي نمايد.

پايه و اساس تطبيق درخواست اين است كه درخواستهاي مشابه داراي مجموعه مدركهاي مشابه هستند. با استفاده از اطلاعات مدركهايي كه با درخواستهاي قبلي مشابه بوده‌اند، مي‌توان مدارك مشابه با درخواستهاي جديد را بدست آورد. تغيير شكل درخواست همانند شخص خبره عمل مي‌كند[12، 13]. به عبارت ديگر سيستم ناظر شبكه عصبي حضور شخص خبره را شبيه‌سازي مي‌كند. در شكل ۳ مدل كلاسيك (شكل ۳- ب)، با مدل هوشمند (شكل ۳- الف) مقايسه شده است سيستم هوشمند داراي مبدل درخواست T مي‌باشد كه با توجه به دانش مجموعه، درخواست را بازسازي مي‌كند.



شكل ۳- مقايسه مدل هوشمند و غيرهوشمند بردار

بازيابي اطلاعات

كاربرد اين سيستم در مجموعه مداركي كه دسته‌بندي[27] شده باشند، بهتر نمايان مي‌گردد. بدين صورت كه مثالهاي آموزشي[28]، براحتي و بطور خودكار، از شاخه[29]هاي موجود در مجموعه استخراج مي‌گردد. هر مثال آموزشي شامل چند كلمه كليدي (درخواست) و مجموعه مدارك مرتبط با كلمات كليدي است.

سيستم هوشمند در دو فاز عمليات بازيابي را انجام مي‌دهد. ابتدا مرحله يادگيري و آموزش ماشين[30] است، در اين مرحله بايد يك ليست از درخواستها (بردارهاي درخواست) و جواب آنها (ماتريس مدارك جواب درخواست) به سيستم داده شود. در اين مرحله سيستم شبكه عصبي دانش زمينه‌اي مجموعه را كسب مي‌نمايد. فاز دوم، فاز بكارگيري و آزمايش سيستم هوشمند است، در اين فاز سيستم هوشمند مانند ناظر، درخواستهاي كاربر را پذيرفته و آنها را بهينه‌سازي مي‌كند و سپس سيستم كلاسيك مانند قبل، بروي درخواست تغيير يافته، عمليات محاسبه شباهت را انجام مي‌دهد.

شـكل ٤ سيـستم هوشـمند را در دو فـاز يـادگيري و بكارگيري نشان مي‌دهد [١٤، ١٥].

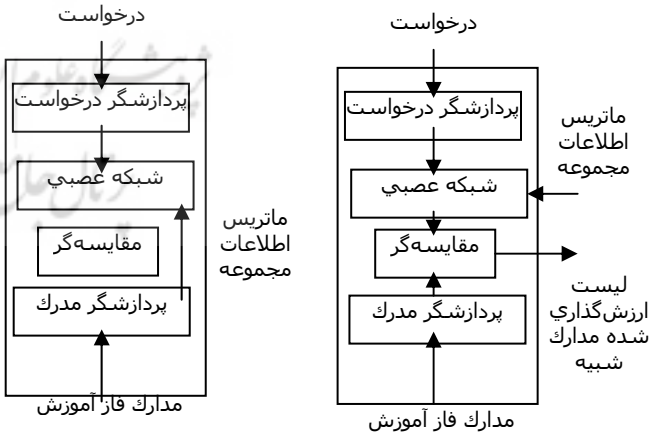
همـانطور كـه مـشاهده مـي‌گـردد سيـستم از چهار قسمت تشكيل شده است:

١- پردازشگر درخواست[٣١]: در اين قسمت از سيستم، درخواست پردازش مي‌گردد تا به بردار تبديل گـردد. در اين مرحله از شاخص و وزن‌دهي استفاده خواهد شد و يا ممكن است براي سرعت بيشتر از مـدل منطقـي اسـتفاده شـود. بنـابراين خروجـي ايـن مرحلـه بـردار درخواست است.

٢- پردازشگر مدرك[٣٢]: اين قطعه از سيستم، مـدارك را مورد پردازش قرار مي‌دهد و براي هر مدرك يك بردار از وزنها، ايجاد مـي‌نمايـد بنـابراين خروجـي ايـن قطعـه از سيستم، ماتريس كلمه-مدرك مي‌باشد.

٣- مقايـسه‌گـر[٣٣]: ايـن قطعـه از سيـستم، بـردار درخواسـت را بـا تمـام بردارهـاي مـدارك مقايـسه مي‌نمايـد، و يـك ليـست ارزش‌گـذاري شـده از مـدارك شبيه را تهيه نموده به كاربر ارايه مي‌نمايد [١٤، ١٥].

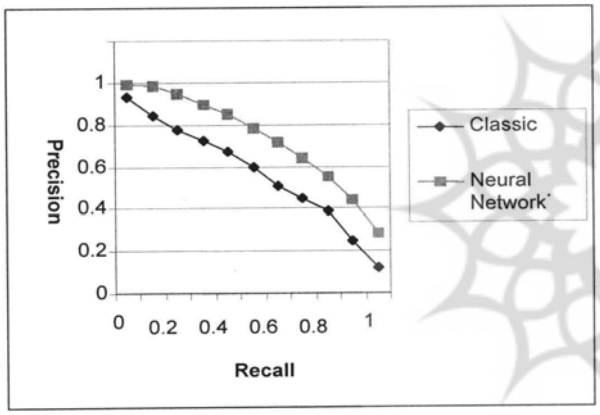
٤- شبكه عـصبي: وظيفـه شـبكه عـصبي تغييـر بـردار درخواست كـاربر بـا توجـه بـه دانـش كـسب شـده، در مرحله آموزش مي‌باشد. اين بردار بـه عنـوان خروجـي اين مرحله به مقايسه‌گر داده مي‌شود.



شكل ٤- سيستم بازيابي اطلاعات هوشمند در فاز آموزش و بكارگيري

بـراي آنكـه سيـستم هوشـمند بتوانـد بخـوبي عموميت بخشي[٣٤] را در دانش مجموعه ايجـاد نمايـد، مثالهايي كه جهت آموزش سيستم انتخاب مي‌گـردد، بايد از تمامي دامنه مجموعه باشد. براي كارآيي بهتـر مي‌توان، مدارك مجموعه را دسته‌بندي نمود و سـپس از هر دسته مدارك شبيه، يك نماينـده[٣٥] كـه عموميـت بيشتري دارد در آموزش شبكه عصبي شركت كند.

در شـكل ٥ نتيجـه آزمـايش ايـن روش بـروي مجموعه مدارك CranField مـشاهده مـي‌گـردد، در ايـن نمودار نتيجه روش كلاسيك با روش هوشمند مقايسه مي‌گردد. اين مجموعه داراي ١٤٠٠ مدرك و ٢٢٥ مثال آموزشي است. تعداد كلمات كليدي كه در بيش از يك مدرك ظاهر شده‌انـد حـدود ٤٤٠٠ كلمـه مـي‌باشـد.در عمـل بـراي آمـوزش شـبكه عـصبي مـي‌تـوان از فهرستهاي موضوعي بيشترين بهره را بـراي، آمـوزش شبكه عصبي بدست آورد. بدليل دسته‌بندي اطلاعات در ايـن فهرسـتها، بهتـرين جـواب در آمـوزش سيـستم بدست خواهد آمد.



شكل ٥- مقايسه نتيجه بازيابي كلاسيك و هوشمند

نتيجه‌گيري

با مشاهده خروجي سيستم هوشمند بـه ايـن نتيجه مي‌رسيم كه سيستم هوشمند داراي كـارآيـي بالاترين نسبت به سيـستم كلاسـيك مـي‌باشـد. ايـن نتيجه بـا نظـارت بـر درخواسـت كـاربر جـواب بهتـري را فراهم آورده است. زيرا بردار درخواست با دانش زمينه تطبيق داده شده و بهينه‌سازي مـي‌گـردد، بـه عبـارت ديگر سيستم هوشمند با درك معنـاي درخواسـت، در صورت نياز آن را بهينه‌سازي مي‌نمايد.

مـدل هوشـمند برخـي مـشكلات مـدلهاي كلاسيك را حل كرده است:

- الزامي ندار كه درخواست ساختاري[36] مانند مدرك داشته باشد تابع مبدل T (شكل ۳) درخواست را پيكربندي[37] مي‌كند، تا شباهت قابل محاسبه و سنجش باشد.
- كاربر ملزم نيست كه درخواست خود را به طور كامل، از محتوايي كه مي‌خواهد بيان كند، تابع مبدلT، با استفاده از دانش محيط، درخواست را تغيير شكل خواهد داد، و درخواست را در فضاي مدارك قرار خواهد داد.
- يك مدل هوشمند مي‌تواند براي محاسبه شباهت استفاده شود، كه رابطه ميان درخواستها با مدركهاي مشابه را با استفاده از بازخورد كاربر مورد محاسبه قرار دهد. موقعيت مدرك در فضاي مدارك، نسبت به تصميم كاربر تغيير خواهد كرد.

بايد توجه كرد كه مدل هوشمند، در صورتي پاسخ مناسب و صحيح خواه داد كه در فاز آموزش با مثالهاي جامع، يادگيري انجام شده باشد در غير اينصورت ممكن است نتيجه مناسبي حاصل نگردد بنابراين پيشنهاد مي‌گردد، كه سيستم هوشمند در صورتي مورد استفاده قرار گيرد، كه درخواست رسيده داراي تاريخچه‌اي در زمان آموزش باشد، به عبارت ديگر در صورتي از سيستم هوشمند استفاده شود، كه درخواست داراي مثال‌هاي مشابي در زمان آموزش باشد، در غير اينصورت از سيستم كلاسيك بدون تغيير درخواست، استفاده گردد و بازخورد اين درخواست مجدداً سيستم را تعليم دهد. يعني با توجه به انتخاب كاربر، مي‌توان با مجموعه‌اي از مثال‌هاي آموزشي سيستم را مجدداً تعليم داد.

پي‌نوشت‌ها:

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.