Evaluation of Most Visited News Websites in Iran Based on Machine Learning

Babak Sohrabi
PhD in Management; Professor in Information Technology Management Group; University of Tehran;
Corresponding Author bsohrabi@ut.ac.ir

Amir Manian
PhD in Management; Associate Professor in Information Technology Management Group; University of Tehran;
amanian@ut.ac.ir

Molood Arman
MS in Information Technology Management;
University of Tehran arman.molood@ut.ac.ir

Received: 15, Feb, 2016 **Accepted: 26, Apr. 2016**

Abstract: Success and effectiveness of websites are largely dependent on their quality. The biggest share of the quality of new concept is that technical aspects of products and services combine with customers usage and understanding. Therefore, websites evaluation based on the maximum usage and perception of the customers is considered an important issue to announce to the related organizations. This ranking and evaluation should be performed in a special activity domain so that the first place of website rank determines among its other competitors. In this article achieving the information of websites is automatic and without the intervention of human so that the instant evaluation could be possible. In this study, one of the multi criteria decision-making methods called TOPSIS is used and the weights of the criteria have been achieved on entropy method in the mentioned method. Eventually, according to the Alexa ranking report, 791 ranking news website have been obtained which have most visitors of Iranian users, but on the other hand, just a numerical rank as a final output of websites evaluation can not be very inconsistent with the purpose of competition between websites, so these numerical ranking from TOPSIS method used as output in machine learing method for separating websites from excellent to very poor in six categories as lables for training dataset in classification, instead of using manual lables achieved from experts and users' opinion. For this classification, machine learning techniques, including artificial neural network and support vector machine were used.

Keywords: Websites Evaluation, News Websites, TOPSIS, Automation, Machine Learning, Neural Network, Support Vector Machine

Iranian Journal of Information Processing and Management

Iranian Research Institute for Science and Technology
ISSN 2251-8223
eISSN 2251-8231
Indexed in SCOPUS, ISC, & LISTA
Vol. 32 | No. 2 | pp: 551-579
Winter 2017

ارزیابی کیفیت پرمشاهده‌ترین وب‌سایت‌های خبری در ایران مبتنی بر روش یادگیری ماشین

بابک سهرابی
دکتری مدیریت؛ استاد؛ دانشگاه تهران؛
گروه مدیریت فناوری اطلاعات؛
پدیدآور رابط bsohrabi@ut.ac.ir

امیر مانیان
دکتری مدیریت؛ دانشیار؛ دانشگاه تهران؛
گروه مدیریت فناوری اطلاعات amanian@ut.ac.ir

مولود آرمان
کارشناسی ارشد مدیریت فناوری اطلاعات؛
دانشگاه تهران arman.molood@ut.ac.ir

دریافت: ۱۳۹٤/۱۱/۲٦ | پذیرش: ۱۳۹۵/۰۲/۰۷ | مقاله برای اصلاح به مدت ۲۷ روز نزد پدیدآوران بوده است.

فصلنامه | علمی پژوهشی
پژوهشگاه علوم و فنّاوری اطلاعات ایران
شاپا (چاپی) ۸۲۲۳-۲۲۵۱
شاپا (الکترونیکی) ۸۲۳۱-۲۲۵۱
نمایه در LISTA، ISC، و SCOPUS
jipm.irandoc.ac.ir
دورهٔ ۳۲ | شمارهٔ ۲ | ص‌ص ۵۵۱-۵۷۹
زمستان ۱۳۹۵

چکیده: موفقیت و کارایی وب‌سایت‌ها تا حد زیادی وابسته به کیفیت آن‌هاست. بزرگ‌ترین سهم از مفهوم جدید کیفیت این است که جنبه‌های فنی محصولات و سرویس‌ها با استفاده و برداشت مشتریان ترکیب می‌شود. بنابراین، ارزیابی وب‌سایت‌ها بر اساس بیشترین استفاده و درک از سمت مشتریان امر مهمی به‌شمار می‌آید تا موفقیت یک وب‌سایت را به سازمان‌های مربوطه اعلام کند. این رتبه‌بندی و ارزیابی بایستی در یک دامنهٔ خاصِ فعالیتی صورت گیرد، تا رتبهٔ وب‌سایت در برابر رقبای آن تعیین گردد. این پژوهش بر آن است که بتوان ارزیابی وب‌سایت‌ها را با به‌دست آوردن اطلاعات آن‌ها به‌صورت خودکار و بدون دخالت نیروی انسانی، به‌صورت لحظه‌ای و بلادرنگ ممکن ساخت. برای این منظور از دو روش یادگیری ماشین، شبکهٔ عصبی و ماشین بردار پشتیبان استفاده شده است. از آنجا که از الگوریتم یادگیری با نظارت در این روش‌ها استفاده کرده‌ایم، نیاز به داده‌های برچسب‌دار وجود داشت که به‌جای برچسب‌دارکردن خروجی‌ها توسط خبرگان و به‌صورت دستی از یکی از روش‌های تصمیم‌گیری چندمعیاره (MCDM) به‌نام «تاپسیس» کمک گرفته شده است. در این روش وزن‌های معیارهای ارزیابی با استفاده از روش آنتروپی به‌صورت خودکار محاسبه گردیده است. در نهایت، رتبه‌بندی ۷۹۱ وب‌سایت خبری که بر اساس گزارش «الکسا» بیشترین بازدیدکننده از طرف کاربران ایرانی را داشته‌اند، با روش «تاپسیس» حاصل شد، ولی از سمت دیگر ارائهٔ یک

رتبهٔ عددی به‌عنوان خروجی نهایی ارزیابی وب‌سایت‌ها چندان با هدف رقابت آن‌ها همخوانی ندارد. به همین خاطر، از این رتبه‌های عددی به‌عنوان خروجی روش‌های یادگیری ماشین به‌جای رتبه‌بندی خبرگان و افراد استفاده گردید تا بتوان وب‌سایت‌ها را در شش طبقهٔ جداگانه، از بسیار عالی تا بسیار ضعیف، برچسب‌دار نمود. در ادامه، از روش‌های یادگیری ماشین ذکرشده جهت طبقه‌بندی و همچنین، پیش‌بینی طبقهٔ یک وب‌سایت جدید استفاده شده است. در نهایت، این پژوهش به این نتیجه رسید که امکان طراحی سیستمی با دقت بالا برای ارزیابی وب‌سایت‌های خبری با روش شبکهٔ عصبی و ماشین بردار پشتیبان وجود دارد، به‌گونه‌ای که این سیستم از عوامل مهم تأثیرگذار بر کیفیت وب‌سایت‌های خبری تشکیل شده و به‌صورت خودکار قابلیت جمع‌آوری اطلاعات آن‌ها وجود داشته باشد.

کلیدواژه‌ها: ارزیابی وب‌سایت، وب‌سایت‌های خبری، روش تاپسیس (TOPSIS)، خودکارسازی، یادگیری ماشین

۱. مقدمه

در سال ۲۰۱۵ بر اساس آمار «استاتستیا»[1] تعداد کاربران اینترنتی فراتر از ۳ بیلیون نفر بوده و تعداد وب‌سایت‌ها به‌صورت روزافزونی در حال رشد و افزایش هستند. مهم‌ترین هدف هر وب‌سایت ارائهٔ اطلاعات است. وب‌سایت‌ها انواع مختلفی دارند، ولی اغلب آن‌ها منابع غنی از اطلاعات بوده و با کمترین هزینه، به‌سرعت و به‌سادگی می‌توانند در فضای جهانی وب، بر روی اینترنت منتشر گردند. بنابراین، در سرتاسر جهان شاهد آن هستیم که با رشد سرعت تغییرات، پویایی محیط، و افزایش رقبا تقریباً هر نوعی از کسب‌وکار برای بهره‌برداری از مزایای اینترنت به سمت آن رفته است.

با نگاهی به پرمشاهده‌ترین وب‌سایت‌ها در آمار وب‌سایت معتبر «الکسا»، بیشترین تکرار در بین صد سایت برتر، هم در جهان و هم در ایران، متعلق به سایت‌های خبری است. از آنجا که رتبه‌بندی وب‌سایت «الکسا» بر اساس تعداد بازدید کاربران وب‌سایت‌ها صورت می‌گیرد، می‌توان به این نکته رسید که این نوع وب‌سایت‌ها از دید عموم به‌شدت مورد استقبال واقع‌شده و توانسته‌اند جایگاه خوبی در میان مردم داشته باشند. بسیاری از نهادها و سازمان های خبری در ایران به‌دنبال درک میزان عمومیت و استقبال میان مخاطبان اخبار آنلاین هستند و بسیار علاقه مندند که جایگاه خود را پیوسته نسبت به سایر رقبای خود ارزیابی نمایند. امروزه، این موضوع برای ارزیابی جایگاه وب‌سایت‌های کسب‌وکاری مشخص تبدیل به یکی از موارد تجاری برای شرکت‌های فعال در حوزهٔ وب گردیده است.

1. Statestia

از این رو، لزوم شناخت مهم‌ترین عوامل و ویژگی‌های تأثیرگذار بر نوع خاصی از وب‌سایت‌ها، جزء ابتدایی‌ترین و در عین حال، مهم‌ترین اقداماتی است که برای ارزیابی وب‌سایت‌ها باید صورت گیرد. همچنین، مدل‌سازی کمّی این ویژگی‌ها با تعیین اهمیت و وزنی که هرکدام از آن‌ها در کیفیت یک وب‌سایت دارند، به نهادهای ذی‌ربط و مالکان وب‌سایت‌ها کمک خواهد کرد که ابتدا بر اساس اولویت و اهمیت مشخص‌شده، وجوه متفاوت وب‌سایت خود را در نظر گرفته و برای بهترکردن جایگاه خود در میان رقبا، در جهت برطرف‌کردن مشکلات و کمبودهای آن اقدام نمایند. بعد از شناسایی این معیارها، می‌توان در راستای امکان طراحی سیستمی برای ارزیابی کیفیت وب‌سایت‌ها اقدام نمود.

دو پرسش اصلی که این پژوهش به‌دنبال یافتن پاسخی برای آن‌هاست، در زیر بیان شده است:

۱. مهم‌ترین عواملی که در ارزیابی خودکار کیفیت وب‌سایت‌های خبری می‌توانند دخیل باشند، چیست؟
۲. آیا می‌توان سیستمی خودکار برای ارزیابی کیفیت وب‌سایت‌های خبری طراحی نمود؟

در ادامه، ابتدا مروری بر ادبیات موضوع داشته و سپس، با بیان روش پژوهش، بر اساس مدل ارائه‌شده شاخص‌ها و نمونه‌ای از اطلاعات کمّی، که به‌صورت خودکار جمع‌آوری شده‌اند، مطرح می‌شود و با طراحی چند سیستم مختلف، خروجی آن‌ها در بخش یافته‌های پژوهش با یکدیگر مقایسه شده و در نهایت، به بیان نتایج و پیشنهادات پرداخته می‌شود.

۲. مروری بر ادبیات موضوع

اینکه یک وب‌سایت تا چه حد باید پاسخگوی انتظارات کاربران و مشتریان باشد، پرسش مهمی است که با کیفیت وب‌سایت ارتباط پیدا می‌کند (زاهدی ۱۳۸۹). از آنجا که تنوع وب‌سایت‌های موجود بسیار زیاد است و ابعاد کیفیت خدمات براساس نوع وب‌سایت‌ها با یکدیگر متفاوت هستند، بنابراین، یکی از بهترین شیوه‌های تعریف کیفیت با درنظرگرفتن نوع وب‌سایت است.

کیفیت، مجموعه‌ای از ویژگی‌ها و مشخصات محصول یا خدمت است که بتواند نیازهای صریح یا ضمنی را برآورده کند (کی.هو، ۱۳۷۹). این نیازها اغلب با توجه به نظر کاربران و مشتریان و برآوردن خواسته‌های آنان درآمیخته است و برخی اعتقاد دارند که جلب رضایت مشتریان و برآوردن خواسته‌های آنان مهم‌ترین عامل کیفیت می‌باشد (زاهدی،۱۳۸۹). از سوی دیگر، ازدیاد سرعت تکامل فناوری‌های وب، منجر به کاهش چرخهٔ عمر وب‌سایت‌ها گردیده و

از آنجاکه پژوهش‌های کیفی به‌شدت زمان‌بر هستند، دیگر زمان چندانی برای پژوهش‌های کیفی و ارتباط مستقیم با کاربران و مشتریان برای ارزیابی کیفی وب‌سایت‌ها باقی نمی‌ماند. علاوه بر این، هر کاربر با توجه به نیاز خود معیارهای متفاوتی برای کیفیت وب‌سایت‌ها داشته و حتی با تغییر زمان، ارزیابی مجدد برای دسته‌ای ثابت از وب‌سایت‌ها نتیجه‌ای متفاوت خواهد داشت (Li and Yamada 2009). در چنین شرایطی در مورد پژوهش‌های کیفی در بارهٔ ارزیابی وب‌سایت‌ها به تعارض بر می‌خوریم و در این جاست که اهمیت ابزارها و سازوکارهای ارزیابی کیفیت وب‌سایت‌ها معلوم می‌شود.

یکی از مهم‌ترین موضوعاتی که امروزه جزء دغدغه‌های عمدهٔ مدیران و بهره‌برداران سایت‌های وب شده، موضوع ارزیابی وب‌سایت‌هاست (پاشازاده و حاجی‌زین‌العابدینی ۱۳۸۸). روش‌های متفاوت زیادی توسط محققان برای ارزیابی وب‌سایت‌ها پیشنهاد شده است، ولی هنوز تحقیقات بر روی آن‌ها ادامه دارد (ولریچ و دیگران،۲۰۱۱). در واقع، مفهوم ارزیابی وب‌سایت اغلب با مفهوم ارزیابی وب در ادبیات تحقیق اشتباه گرفته می‌شود. این امر، بسیار مهم است که دامنه و هدف ارزیابی معلوم گردد: آیا مطالعه بر روی ارزیابی وب است یا ارزیابی وب‌سایت‌هایی از نوع خاص؟ بنابراین، آیا هدف از ارزیابی، طراحی مجدد وب‌سایت است یا رتبه‌بندی وب‌سایت‌ها یا به‌دست‌آوردن آماری از ویژگی‌های آن‌ها؟ در نتیجه، به تمرکزی متفاوت برای روش ارزیابی، بر اساس هدف و پلاتفرم ارزیابی نیاز می‌باشد (Zahran et al. 2014).

علی‌رغم افزایش چشمگیر وب‌سایت‌ها و کاربران اینترنتی آن‌ها، تعداد مقالات آکادمیک، که به تحقیق در وجوه و ویژگی‌های مختلف وب‌سایت‌ها پرداخته باشند، محدود است (Hernandez, Jimenex and Martin 2009). در رابطه با عوامل تأثیرگذار بر کیفیت وب‌سایت‌ها و روش‌های ارزیابی آن‌ها، پژوهش‌هایی بر روی وب‌سایت‌های تجارت الکترونیک، وب‌سایت‌های کتابخانه‌ای و دانشگاهی صورت گرفته است (Younghwa and Serhat and Cengiz 2012; Hsiu-Fen 2010; Kenneth 2006). ولی، تا کنون در بارهٔ معیارهای ارزیابی خاص وب‌سایت‌های خبری، به‌خصوص با توجه به درنظرگرفتن روش ارزیابی خودکار، به‌ندرت پژوهشی انجام گرفته است (کنلوپلوس و کتسیانتیس،۲۰۱۲). در پژوهش «کنلوپلوس» هم تمرکز اصلی بر ارزیابی وب‌سایت‌های خبری یونانی نیست، بلکه گرایش بیشتر به سمت طراحی سیستمی توصیه‌گر برای تطابق اخبار با پروفایل کاربران است. وی در ابتدا بر اساس ویژگی‌هایی، که اغلب آن‌ها را از سایت «الکسا» استخراج کرده، وب‌سایت‌های خبری یونان را ارزیابی می‌کند.

ابزارهای ارزیابی خودکار وب‌سایت‌ها نقش بزرگ‌تر و تأثیرگذارتری را به‌نسبت ابزارهای ارزیابی غیر خودکار وب‌سایت‌ها در رتبه‌دهی و ارزیابی آن‌ها بازی می‌کنند (Dominic et al.

2011). علاوه بر این، در تعریف خودکاربودن ارزیابی، اصولاً دیدگاه نویسندگان فقط به سمت استفاده از ابزارهای آنلاین ارزیابی خودکار وب‌سایت‌ها، مانند استفاده از ابزار وب‌اگزکت[1] (حسن‌زاده و نویدی،۱۳۸۹)، سرویس اعتبارسنجی W3C HTML (Bauer and Scharl 2000) یا ابزار اندازه‌گیری میزان عمومیت جهانی وب‌سایت (Fang 2007; Cho and Adams 2005; Scowen 2007) بوده است. پژوهش‌های انگشت شماری در رابطه با ارزیابی خودکار وب‌سایت‌ها صورت گرفته که تأثیر اغلب معیارها و ویژگی‌های تأثیرگذار بر وب‌سایت‌ها را در نظر داشته‌اند (Li and Yamada 2009).

یکی از موضوعات مهم در زمینهٔ توسعهٔ سیستم‌های اطلاعاتی، یادگیری ماشین است. برای پیش‌بینی موفقیت کسب وکار وب‌سایت‌ها مقالات و پژوهش‌هایی ارائه شده است، ولی غیر از مبحث روش‌های یادگیری برای رتبه‌دهی که مبتنی بر روش‌های یادگیری ماشین هستند و به‌روش‌های مبتنی بر ارتباط نیز معروف‌اند (Derhami et al. 2013) و از مهم‌ترین ویژگی‌های آن نیاز به دسترسی به شبکهٔ ارتباطات وب‌سایت‌ها برای تعیین رتبه می‌باشد، تقریباً از دیگر روش‌های یادگیری ماشین برای مسائل مرتبط با وب‌سایت‌ها، خصوصاً ارزیابی آن‌ها، به‌ندرت استفاده شده است (Zhang et al. 2014; Li and Yamada 2009).

۳. روش پژوهش

در پژوهش حاضر برای تجزیه و تحلیل داده‌های گردآوری‌شده در بخش کمّی، از روش‌های آماری و روش‌های یادگیری ماشین استفاده شده است. همان‌گونه که واضح و مشخص است، باگذشت زمان علم نیز پیشرفت می‌کند. در پی کشف محدودیت های روش‌های قدیمی روش‌های جدید علمی ایجاد می‌شود و از آنجا که روش‌های آماری جزء روش‌های قدیمی داده‌کاوی محسوب می‌شوند، از این قاعدهٔ کلی مستثنی نیستند. داشتن فرض اولیه در مورد داده‌ها یکی از این موارد است.

در جایی پایه و اساس داده‌کاوی به دو مقولهٔ آمار و هوش مصنوعی تقسیم شده است که روش‌های مصنوعی به‌عنوان روش‌های یادگیری ماشین در نظر گرفته می‌شوند. فرق اساسی بین روش‌های آماری و روش‌های یادگیری ماشین بر اساس فرض‌ها و یا طبیعت داده‌هایی است که پردازش می‌شوند. به‌عنوان یک قانون کلی، فرض‌های روش‌های آماری بر این اساس است که توزیع داده‌ها مشخص است و در بیشتر موارد فرض بر این است که توزیع، نرمال است و در

1. webxact

نهایت، درستی یا نادرستی نتایج نهایی به درست‌بودن فرض اولیه وابسته است. در مقابل، روش‌های یادگیری ماشین از هیچ فرض در مورد داده‌ها استفاده نمی‌کنند و همین مورد باعث تفاوت‌هایی بین این دو روش است.

در این پژوهش از روش ماشین بردار پشتیبان و شبکهٔ عصبی مصنوعی به‌عنوان روش یادگیری ماشین و روش «تاپسیس»[1] از روش‌های تصمیم‌گیری چندمعیاره برای داوری در بارهٔ رتبه‌های مجموعه دادهٔ فراهم‌شده استفاده شده است.

مراحل طراحی و اجرای سیستم یادگیری که در این پژوهش به‌کار رفته، به‌شرح زیر است:

۱. انتخاب شاخص‌های مهم ارزیابی کیفیت وب‌سایت؛
۲. جمع‌آوری مجموعهٔ داده؛
۳. رتبه‌دهی بر اساس روش «تاپسیس»؛
۴. تعیین روش یادگیری ماشین؛
   ۴-۱. پیش‌پردازش داده‌ها؛
   ۴-۲. تقسیم داده‌ها؛
   ۴-۳. انتخاب الگوریتم یادگیری؛
   ۴-۴. تعیین پارامترهای مورد نیاز ماشین یادگیری؛
   ۴-۵. آزمایش داده‌ها؛
   ۴-۶. آزمون داده‌ها.

در ادامه به توضیح هر یک از مراحل پرداخته شده است.

موارد شناسایی‌شده از تحقیقات پیشین با درنظرگرفتن نوع استفاده از این معیارها در ارزیابی خودکار استخراج گردیدند و در نهایت، ۱۳ شاخص اصلی برای سنجش حاصل گردید. لیستی از مهم‌ترین عوامل تأثیرگذار در کیفیت وب‌سایت‌ها در زیر ارائه گردیده است:

◇ همگانی‌بودن وب‌سایت که با شاخص رتبهٔ صفحهٔ گوگل، تعداد کاربر مشاهده‌کنندهٔ وب‌سایت، و تعداد صفحات مشاهده‌شده همان کاربر و میزان دلتای تغییرات تعداد کاربران در سه ماه گذشته سنجیده می‌شود.

◇ دسترس‌پذیری وب‌سایت که با شاخص امتیاز موتور جست‌وجو، میزان نمایه‌شدن در موتورهای جست‌وجوی مختلف، و جایگاه وب‌سایت در خروجی query «وب‌سایت‌های خبری» در موتور جست‌وجوی گوگل سنجیده می‌شود.

1. Topsis

۳-۱. انتخاب شاخص‌های مهم ارزیابی کیفیت وب‌سایت

با توجه به اینکه ادبیات پژوهشی ارزیابی کیفیت وب‌سایت‌ها از غنای بالایی برخوردار است، از این رو، برای استخراج شاخص‌ها از روش‌های هدایت‌شدهٔ تلخیصی به‌صورت ترکیبی استفاده شده است. مهم‌ترین فرض در تحلیل محتوا عبارت است از شناسایی واژگانی که بیشترین میزان تکرار و فراوانی را در متون علمی دارند. میزان دفعات تکرار این واژگان در متون مختلف علمی، نشان‌گر میزان اهمیت و توجه پژوهشگران مختلف به آن شاخص‌هاست. در حقیقت، مهم‌ترین گام تحلیل محتوا را می‌توان محاسبهٔ فراوانی عبارات و شاخص‌های مد نظر پژوهشگر در محتوای علمی در نظر گرفت. تحلیل محتوا قادر است با هر داده‌ای از قبیل داده‌های حاصل از سخنرانی، مقاله، کتاب، تصویر، مصاحبه، و مجله ارتباط برقرار کند (Reesi vananie, Manian and Shafia 2013).

مهم‌ترین روش‌های گردآوری اطلاعات در این قسمت از پژوهش، استفاده از روش‌های کتابخانه‌ای است که به‌صورت گسترده از مطالعهٔ کتابخانه‌ای شامل کتب و مقالات فارسی و لاتین، پایان‌نامه‌های داخلی و خارجی، تحقیقات و گزارشات معتبر و مرتبط و همچنین، اینترنت جهت گردآوری مفاهیم علمی و تحلیل محتوای ادبیات موضوعی و شاخص‌های سنجش استفاده خواهد شد.

بخش کیفی این پژوهش از سؤال اول نشأت می‌گیرد، که فرایند شناخت مهم‌ترین عوامل کیفیت وب‌سایت‌ها را فراهم می‌کند، البته با تأکید بر این موضوع که این عوامل امکان استفاده در فرایند گردآوری و ارزیابی خودکار را داشته باشند. بعد از تمامی مطالعات منطبق با موضوعات مطرح در ارزیابی کیفیت وب‌سایت، مهم‌ترین عوامل برای ارزیابی خودکار کیفیت وب‌سایت‌ها شناخته شد. همان‌طور که پیش‌تر هم بیان شد، یکی از روش‌های تحلیل محتوا زمانی که تحقیقات زیادی در ارتباط با موضوع صورت گرفته باشد، بررسی تحقیقات و مطالعات پیشین است تا در نهایت، به اشباع رسیده و حرف تازه‌ای در تحقیقات دیده نشود. موارد شناسایی‌شده از تحقیقات پیشین با درنظرگرفتن نوع استفاده از این معیارها در ارزیابی خودکار استخراج گردیدند و در نهایت، ۱۳ شاخص اصلی برای سنجش حاصل گردید. لیستی از مهم‌ترین عوامل تأثیرگذار در کیفیت وب‌سایت‌ها در زیر ارائه گردیده است:

- ◇ همگانی‌بودن وب‌سایت که با شاخص رتبهٔ صفحهٔ گوگل، تعداد کاربر مشاهده‌کنندهٔ وب‌سایت، و تعداد صفحات مشاهده‌شده همان کاربر و میزان دلتای تغییرات تعداد کاربران در سه ماه گذشته سنجیده می‌شود؛
- ◇ دسترس‌پذیری وب‌سایت که با شاخص امتیاز موتور جست‌وجو، میزان نمایه‌شدن در موتورهای جست‌وجوی مختلف، و جایگاه وب‌سایت در خروجی query «وب‌سایت‌های

خبری» در موتور جست‌وجوی گوگل سنجیده می‌شود؛

- ◇ سرعت که با شاخص میانگین زمان بارگذاری سنجیده می‌شود؛
- ◇ پیمایش که با شاخص تعداد کل لینک‌هایی که به سایت داده شده، محاسبه می‌شود؛
- ◇ نحوهٔ نمایش که با میزان خطای موجود در قالب کد HTML سایت که از وب‌سایت W3C کسب گردیده، سنجیده می‌شود؛
- ◇ محتوای چندرسانه‌ای که با میانگین تعداد فایل‌های تصویری، فلش، ویدئو و صدا در صفحهٔ اصلی وب‌سایت سنجیده می‌شود؛
- ◇ سابقهٔ وب‌سایت که با شاخص میزان عمر دامنه سنجیده می‌شود؛
- ◇ ضریبی از ارزیابی‌های مورد قبول جهانی که در اینجا رتبهٔ ترافیک سایت «الکسا» در نظر گرفته شده که بر اساس میزان مشاهدهٔ کاربران وب‌سایت‌ها رتبه‌بندی می‌شوند.

نمایی از این عوامل را در شکل ۱ می‌توان مشاهده نمود.

شکل ۱. عوامل مؤثر در کیفیت وب‌سایت

۳-۲. جمع‌آوری مجموعهٔ داده

روش‌های غیرکتابخانه‌ای (استفاده از خزشگر): برای جمع‌آوری اطلاعات وب‌سایت‌ها و اطلاعات تاریخچه‌ای آن‌ها، در این پژوهش از روش‌های مبتنی بر خزش استفاده می‌شود. فرایند خزش دو نوع است: نوع اول، خزش و جمع‌آوری تمام داده‌ها از یک سرویس خاص، مانند کاری که موتورهای جست‌وجو انجام می‌دهند؛ نوع دوم، خزش هدفمند. در این روش بسته به اطلاعات مورد نیاز، روبات مورد نظر طراحی‌شده و در سرویس‌های مختلف به خزش و جمع‌آوری اطلاعات می‌پردازد. استفاده از نوع اول خزش در تحلیل و ارزیابی بسیار سودمند بوده، ولی دارای هزینهٔ نگهداری بسیار بالایی است. از نوع دوم خزش نیز بیشتر برای پرکردن خلأهای اطلاعاتی و کسب اطلاعاتی خاص استفاده می‌شود. خزش، بسته به نوع آن دارای الگوریتم‌های مختلف بوده و بایستی برای هر مورد و هر پایگاه داده‌ای به‌صورتی جداگانه طراحی شود. اطلاعات شاخص‌های شناسایی‌شده با کمک خزشگرها از منابع اطلاعاتی موجود در جدول ۱ استخراج گردیده‌است. این اطلاعات مرتبط با ۱۳ شاخص برای ۷۹۱ وب‌سایت خبری، که بر اساس آمار «الکسا» بیشترین بازدید در ایران را داشتند، جمع‌آوری گردیده است.

جدول ۱. منابع جمع‌آوری اطلاعات

نام شاخص	منبع گردآوری اطلاعات
رتبهٔ ترافیک الکسا	سایت الکسا
رتبهٔ صفحهٔ گوگل	سایت گوگل
امتیاز موتور جست‌وجو	سایت استاتس کراپ
تعداد لینک Backward	سایت استاتس کراپ/ سایت الکسا
تعداد کاربران منحصربه‌فرد	سایت الکسا
میزان نرخ تغییر تعداد کاربران منحصربه‌فرد	سایت الکسا
میزان مشاهدات وب‌سایت	سایت الکسا
محتوای مالتی‌مدیا	خود وب‌سایت
عمر دامنه	سایت الکسا/ سایت استاتس کراپ
سرعت بارگذاری صفحه	سایت الکسا/ سایت استاتس کراپ
جایگاه وب‌سایت در جست‌وجوی گوگل	سایت گوگل
میانگین تعداد نمایه در موتورهای جست‌وجو	سایت گوگل/ سایت بینگ/ سایت یاهو
میزان خطای کد HTML	سایت w3c

۳-۳. رتبه‌دهی بر اساس روش «تاپسیس»

بعد از جمع‌آوری اطلاعات شاخص‌ها، داده‌ها را نرمالایز کرده و وزن شاخص‌ها به‌روش آنتروپی اطلاعات به‌دست آمد. این وزن‌ها در جدول ۲ آورده شده است. بر اساس مقادیر شاخص و این اوزان، میزان فاصلهٔ نسبی از ایده‌آل مثبت برای هر وب‌سایت اندازه‌گیری شده است و در نهایت، اطلاعات ۳۰ سایت برتر رتبه‌داده‌شده به وب‌سایت‌های خبری در جدول ۳ آورده شده است.

جدول ۲. وزن معیارها با استفاده از روش آنتروپی

نام شاخص	وزن شاخص از روش آنتروپی
رتبهٔ صفحه گوگل	۰٫۱۴۴۵۲۸
امتیاز موتور جست‌وجو	۰٫۰۶۴۲۶۲
رتبهٔ ترافیک الکسا	۰٫۲۵۸۹۱۵
عمر دامنه	۰٫۱۷۰۵۹۳
میانگین تعداد نمایه در موتورهای جست‌وجو	۰٫۱۷۰۵۹
خطا در کد HTML	۰٫۲۸۸۹۱
لینک‌های backward	۰٫۱۳۶۷۴
تعداد کاربر منحصربه‌فرد	۰٫۱۱۹۱۲
میزان تغییرات در تعداد کاربر	۰٫۰۵۸۴۴
میزان مشاهده صفحات	۰٫۰۴۳۲۹۴
زمان بارگذاری صفحات	۰٫۰۳۱۳۱۴
جایگاه وب‌سایت در جست‌وجو جوی گوگل	۰٫۱۲۷۴۸۵
محتوای چندرسانه‌ای	۰٫۰۸۲۲۹

پژوهشگاه علوم انسانی و مطالعات فرهنگی
پرتال جامع علوم انسانی

جدول ۳. داده‌های خام و نتایج نهایی ۳۰ سایت برتر خبری بر اساس روش «تاپسیس»

رتبه	۱	۲	۳	۴	۵	۶	۷
نام وب‌سایت	Cnn.com	Reuters.com	Cnet.com	Ghatreh.com	Ruvr.ru	Vista.ir	Skysports.com
رتبهٔ صفحه گوگل	۹	۹	۸	۵	۷	۳	۸
امتیاز موتور جست‌وجو	۵۰	۶۰	۶۳	۵۷	۷۶	۴۲	۵۷
رتبهٔ ترافیک الکسا	۱۸۰۱	۳۷۵۷	۴۲۷	۸۰	۲۷۷۷	۶۸	۸۹۲۵
عمر دامنه (ماه)	۲۴۳,۴۵۲۱	۲۴۷,۱۳۴۲	۲۳۸,۰۶۰۳	۸۶,۶۳۰۱۴	۹۵,۳۷۵۳۴	۰	۱۸۶۹۳۷
میانگین تعداد نمایه در موتورهای جست‌وجو	۴۳۶۰۰۰۰	۴۰۰۰۲۸	۲۸۵۰۰۰۰	۲۱۴۳۳۵۲	۱۱۸۰۹۴۷	۱۰۸۸۶۸۱	۸۸۳۳۵۸,۷
خطا در کد HTML	۳۷۵	۲۵۲	۹	۰	۲۵۱	۷۰۹	۲۹۹
لینک‌های backward	۲۶۵۴۷۹	۱۴۹۷۴۸	۱۷۰۴۵۲	۲۹۶۶	۱۴۴۸۹	۱۶۳۸	۱۰۱۵۶
تعداد کاربر منحصربه‌فرد	۱۸۷۷۰	۴۱۱۰	۱۲۲۸۰	۳۵۸	۵۶۰	۳۹۲	۱۲۹۱
میزان تغییرات در تعداد کاربر	۵,۶	۴,۶۵–	۱,۹	۳,۵	۰,۵	۱۹,۵	۱۷,۶۹–
میزان مشاهدهٔ صفحات	۳,۱۹	۲,۳۴	۲,۹۵	۳,۸۶	۲,۶۲	۳,۷۳	۳,۳۸
زمان بارگذاری صفحات (ثانیه)	۲۶۳۷	۳۰۰۷	۳۳۵۱	۲۰۰۰	۲۵۸۷	۱۳۵۷	۲۰۵۸
جایگاه وب‌سایت در جست‌وجو گوگل	–	–	–	۷۹	–	۱۶۴	–
محتوای چندرسانه‌ای	۲۹,۲۵	۸,۵	۱۲,۵	۲,۷۵	۱۷,۵	۴,۵	۱۱,۲۵
میزان فاصله از ایده آل +	۷۵۸,۰۲۰۳	۶۴۴۵,۶۶۴	۲۵۷۹۸,۶۲	۳۷۸۸,۱۷	۵۴۳۴۷,۱۸	۵۴۳۴۷,۱۸	۵۵۹۲۴,۲۵
میزان فاصله از ایده آل -	۷۴۵۶۶,۱۱	۶۸۳۵۰,۲۳	۴۸۵۷,۰۷	۳۶۸۱۸,۴۶	۲۰۴۷۳,۹۵	۱۹۰۷۱,۶	۱۵۲۲۱,۸۴
نزدیکی نسبی به راه‌حل ایده آل مثبت	۰,۰۱۰۰۶۳	۰,۰۸۶۱۷۷	۰,۳۴۵۵۶۸	۰,۵۰۷۸۱۸	۰,۷۲۶۳۶۱	۰,۷۴۵۶۹۸	۰,۷۹۶۱۷۱

رتبه	۸	۹	۱۰	۱۱	۱۲	۱۳	۱۴
نام وب‌سایت	Khabarfarsi.com	Khorasannews.com	Jamnews.com	Jahannews.com	Eqeep.ir	Cbc.ca	Mehrnews.com
رتبهٔ صفحه گوگل	۴	۵	۴	۵	۲	۸	۶
امتیاز موتور جست‌وجو	۳۱	۵۸	۴۷	۴۲	۴۴	۶۳	۷۲
رتبهٔ ترافیک الکسا	۳۴	۳۰۴	۵۶	۷۹	۱۸۶۰	۱۱۲۷۲	۴۲
عمر دامنه (ماه)	۷۲,۱۶۴۳۸	۱۴۵,۷۷۵۳	۰	۸۱,۰۴۱۱	۰	۱۵۸,۵۹۷۳	۱۳۳,۹۳۹۷
میانگین تعداد نمایه در موتورهای جست‌وجو	۷۳۷۶۸۰,۷	۴۸۰۰۱۶,۳	۴۰۶۱۱۴	۳۹۶۰۴۳	۴۰۳۳۵۲,۳	۴۲۰۰۱۴	۳۲۵۰۲۲,۳
خطا در کد HTML	۲۶۹	۰	۹۷	۲۱۸	۱۱	۳۱	۱۰۳
لینک‌های backward	۴۲۱۸	۲۸۷۳	۵۱۵۹	۴۷۷۳	۱۵۲	۶۲۷۱۵	۱۳۸۴۰
تعداد کاربر منحصربه‌فرد	۵۷۶	۱۰۲,۱	۴۰۳	۳۱۸	۱۹,۹	۱۴۵۷	۶۹۰
میزان تغییرات در تعداد کاربر	-۸,۶۲	-۶,۸۴	-۱۸,۷۲	-۱۰,۹۵	-۷,۷۲	۷,۹	-۶,۰۲
میزان مشاهدهٔ صفحات	۶,۳۸	۴,۲۱	۴,۷۱	۵,۴۷	۱,۷۴	۲,۲۱	۳,۶۸
زمان بارگذاری صفحات (ثانیه)	۱۷۸۹	۱۸۸۸	۲۱۹۴	۲۰۵۰	۱۶۱۱	۲۹۸۵	۱۲۶۵
جایگاه وب‌سایت در جست‌وجو گوگل	۱	۱۱۸	۲۲۷	۲۱	–	–	۵۸
محتوای چندرسانه‌ای	۲۶,۲۵	۱۴,۲۵	۱۹,۲۵	۲۳,۷۵	۴,۲۵	۸,۵	۲۳,۷۵
میزان فاصله از ایده آل +	۵۹۴۵۷,۵۷	۶۱۸۹۸,۱۲	۶۶۲۸۷,۵۵	۶۷۵۴۴,۶۶	۶۷۷۱۶,۴۷	۶۷۵۹۷,۰۴	۷۶۳۳۴,۰۶
میزان فاصله از ایده آل -	۱۳۳۱۳,۱۷	۹۲۳۷,۹۵۹	۸۱۷۳,۵۹۲	۸۰۲۵,۹۸۷	۷۹۰۲,۱۸۲	۷۳۹۱,۸۶۵	۷۰۴۸,۰۴۲
نزدیکی نسبی به راه‌حل ایده آل مثبت	۰,۸۲۲۹۹	۰,۸۷۷۶۸۴	۰,۸۹۲۰۵۳	۰,۸۹۴۰۳۶	۰,۸۹۵۳۳۴	۰,۹۰۱۰۸	۰,۹۰۷۲۲۳

رتبه	۱۵	۱۶	۱۷	۱۸	۱۹	۲۰	۲۱
نام وب‌سایت	Youm7.com	Tabnak.ir	Irna.ir	Forexfactory.com	Afkarnews.com	Aljazeera.com	Asriran.com
رتبهٔ صفحه گوگل	۶	۵	۷	۴	۵	۸	۶
امتیاز موتور جست‌وجو	۵۲	۵۰	۵۵	۶۵	۵۲	۵۷	۴۲
رتبهٔ ترافیک الکسا	۱۱۶۴۶	۱۲	۱۱۰	۴۵۷۸	۴۰	۶۴۷۱	۲۵
عمر دامنه (ماه)	۷۲,۹۸۶۳	۰	۰	۱۱۸,۴۲۱۹	۰	۲۰۸,۱۷۵۳	۹۴,۲۹۰۴۱
میانگین تعداد نمایه در موتورهای جست‌وجو	۳۹۶۰۱۳,۷	۳۱۵۶۶۶,۷	۳۱۶۳۴۷	۳۴۶۶۹۴,۷	۳۰۰۳۵۷	۳۳۲۰۱۳,۳	۲۷۱۶۹۴,۳
خطا در کد HTML	۶۷۶	۱۲	۵۵	۰	۳۰۰	۰	۱۲
لینک‌های backward	۱۴۶۳۱	۱۱۸۰۶	۱۲۷۸۳	۳۲۶۲	۵۷۱۹	۳۲۳۶۹	۹۶۷۰
تعداد کاربر منحصربه‌فرد	۲۲۸۱	۱۴۴۱	۳۲۱	۱۲۴۵	۵۷۰	۱۰۰۰	۹۲۴
میزان تغییرات در تعداد کاربر	-۲۱,۱۶	-۰,۴۳	۱۹,۱	۵	-۳,۴۳	-۳۶,۴۱	-۶,۰۳
میزان مشاهدهٔ صفحات	۴,۱۸	۶,۱۶	۳,۳۳	۳,۷	۳,۱	۲,۴	۵
زمان بارگذاری صفحات (ثانیه)	۳۱۶۲	۱۵۹۴	۳۸۱۱	۲۷۰۴	۲۴۶۴	۴۱۲۵	۳۹۴۳
جایگاه وب‌سایت در جست‌وجو گوگل	-	۶	۱۳	-	۳۹۴	-	۱۹
محتوای چندرسانه‌ای	-۰,۵	۵۴,۷۵	۲۵,۵	۰,۷۵	۲۱	۲۰,۵	۰,۲۵
میزان فاصله از ایده آل +	۶۸۹۱۹,۹۱	۶۷۷۷۷,۰۵	۶۹۰۸۰,۷	۶۹۰۶۸,۵۸	۶۸۵۶۸,۳۹	۶۸۸۰۹,۱۱	۶۹۸۳۱,۳۹
میزان فاصله از ایده آل -	۶۹۲۶,۹۸۸	۶۹۲۵,۸۲۹	۶۹۱۰,۶۸	۶۷۲۳,۲۷۳	۶۷۱۴,۶۷۶	۶۲۹۶,۹۱۹	۶۳۴۷,۰۰۷
نزدیکی نسبی به راه‌حل ایده آل مثبت	۰,۹۰۷۲۷۴	۰,۹۰۸۸۷۸	۰,۹۰۹۰۴۵	۰,۹۱۰۷۰۴	۰,۹۱۱۷۱۹	۰,۹۱۶۱۶	۰,۹۱۶۶۸۲

رتبه	۲۲	۲۳	۲۴	۲۵	۲۶	۲۷	۲۸
نام وب‌سایت	Sciencemag.com	Isna.ir	Khabaronline.ir	Javanemrooz.com	Khabaryaab.com	Tnews.ir	Yjc.ir
رتبهٔ صفحه گوگل	۸	۶	۵	۳	۳	۴	۶
امتیاز موتور جست‌وجو	۶۴	۵۷	۵۷	۵۵	۶۸	۶۳	۴۴
رتبهٔ ترافیک الکسا	۱۴۳۵۰	۳۷	۲۹	۱۱۱۵	۳۷۶۲	۹۹	۲۳
عمر دامنه (ماه)	۲۱۲,۲۸۴۹	۰	۰	۵۳,۸۸۴۹۳	۴۷,۷۶۹۸۶	۰	۰
میانگین تعداد نمایه در موتورهای جست‌وجو	۳۵۶۶۹۴,۳	۲۶۱۱۸۰,۷	۲۵۶۳۴۷,۳	۲۴۶۳۴۷,۳	۲۶۸۱۱۰,۳	۲۱۵۶۹۱,۷	۱۹۱۴۸۰,۷
خطا در کد HTML	۰	۵۹	۴۹۹	۳۲	۲۹	۱۵	۲۹
لینک‌های backward	۳۱۹۴۶	۱۶۰۲۳	۹۳۱۵	۹۸۷	۱۷۰	۲۵۶۳	۶۵۸۰۸
تعداد کاربر منحصربه‌فرد	۱۷۹	۷۲۷	۸۵۷	۳۳,۱	۱۰,۷	۲۳۴	۹۴۷
میزان تغییرات در تعداد کاربر	-۲,۱۴	-۰,۳۱	-۲,۳۲	۷	-۵,۲۵	۱۱,۱	-۵
میزان مشاهدهٔ صفحات	۱,۸۵	۴,۱۱	۳,۴۶	۳,۱۵	۲,۱۵	۵,۲۱	۴,۵۱
زمان بارگذاری صفحات (ثانیه)	۳۰۲۱	۱۵۶۶	۲۱۷۵	۲۶۴۸	۱۸۴۳	۱۹۶۷	۱۶۲۱
جایگاه وب‌سایت در جست‌وجو گوگل	-	۵۶	۱۳۹	-	-	۱۳۲	۱۵
محتوای چندرسانه‌ای	۳,۵	۱۶,۲۵	۱۴	۳۴,۷۵	۹,۵	۱۲,۷۵	۲۴,۵
میزان فاصله از ایده آل +	۶۸۴۶۹,۲۵	۷۰۰۰۶,۰۸	۷۰۰۹۳,۰۵	۷۰۲۶۹,۶	۶۹۹۰۵,۳۳	۷۰۷۹۰,۳۴	۷۱۱۶۴,۳
میزان فاصله از ایده آل -	۶۱۸۷,۵۵۹	۶۲۲۸,۶۹۴	۶۱۶۵,۰۴۹	۵۸۴۷,۸۷۴	۵۷۰۹,۲۹۷	۵۶۸۰,۵۳۱	۵۵۱۳,۵۷۲
نزدیکی نسبی به راه‌حل ایده آل مثبت	۰,۹۱۷۱۲	۰,۹۱۸۲۹۶	۰,۹۱۹۱۵۵	۰,۹۲۳۱۷۳	۰,۹۲۴۴۹۵	۰,۹۲۵۷۱۶	۰,۹۲۸۰۹۴

رتبه	۲۹	۳۰
نام وب‌سایت	Jarchi.ir	Irib.ir
رتبهٔ صفحه گوگل	۳	۷
امتیاز موتور جست‌وجو	۶۳	۵۰
رتبهٔ ترافیک الکسا	۱۱۹	۱۲۰
عمر دامنه (ماه)	۰	۰
میانگین تعداد نمایه در موتورهای جست‌وجو	۱۹۸۳۴۶٫۷	۱۷۴۳۶۰٫۳
خطا در کد HTML	۰	۸۸
لینک‌های backward	۱۵۳	۱۴۱۷۲
تعداد کاربر منحصربه‌فرد	۲۸۹	۴۰۹
میزان تغییرات در تعداد کاربر	۳۳	۷٫۲۹-
میزان مشاهدهٔ صفحات	۲٫۳۴	۳٫۸۹
زمان بارگذاری صفحات (ثانیه)	۱۷۱۳	۱۹۹۰
جایگاه وب‌سایت در جست‌وجو گوگل	–	۶۰
محتوای چندرسانه‌ای	۱٫۲۵	۱۴٫۲۵
میزان فاصله از ایده آل +	۷۱۰۸۷٫۴۹	۷۱۴۸۶٫۶۴
میزان فاصله از ایده آل -	۵۴۹۰٫۸۷	۵۲۵۰٫۵۳۷
نزدیکی نسبی به راه‌حل ایده آل مثبت	۰٫۹۲۸۲۹۷	۰٫۹۳۱۵۷۸

با توجه به اثبات کارایی مدل‌های تصمیم گیری چندمعیاره می‌توان چنین فرض کرد که نتایج خروجی حاصل از روش «تاپسیس»، رتبه‌بندی را به دست می‌دهد که به‌عنوان داده‌های یادگیری و آزمون در روش یادگیری ماشین به‌منزلهٔ خروجی مورد انتظار استفاده شود، یعنی بردارهایی از ویژگی کیفیت وب‌سایت‌ها به‌همراه خروجی رتبهٔ آن‌ها حاصل می‌گردد که برای یادگیری و آزمون در روش یادگیری ماشین قرار است از آن‌ها استفاده شود.

۳-٤. استفاده از روش یادگیری ماشین

از دو روش شبکهٔ عصبی و ماشین بردار پشتیبان برای پیش‌بینی مقادیر جدید استفاده گردیده است که کلیت مدل استفاده‌شده در زیر توصیف می‌شود.

به‌صورت کلی، در بررسی‌هایی که به‌عنوان خروجی سیستم ارزیابی کیفیت خودکار وب‌سایت‌ها انجام شد، به این نتیجه رسیدیم که بیان یک رتبه به‌عنوان خروجی نهایی، نه‌تنها ممکن است تأثیر زیاد خوبی بر رقابت سالم در حوزهٔ کسب‌وکار نداشته باشد، بلکه ممکن است با درنظرگرفتن شاخص‌های دیگر یا رقابت تنگاتنگی که رتبه‌های نزدیک با یکدیگر دارند، به جای سیستم رتبه‌دهی عددی از یک الی‌آخر به‌دست‌آمده از روش «تاپسیس» به جای رتبه‌دهی دستی، رتبه‌بندی نهایی را به طیفی از طبقه‌های شش‌گانه «بسیار عالی»، «عالی»، «خوب»، «متوسط»، «ضعیف» و «بسیار ضعیف» نگاشت کنیم. برای این کار، توزیعی نرمال از رتبه‌های نهایی از ۱ تا ۷۹۱ را در نظر گرفته و تلاش شده است تا فاصله بین رتبه‌ها به‌صورت آماری استخراج گردد.

بنابراین، بر اساس توزیع نرمال با میانگین ۳۹۶= μ و ۲۲۸/۴۸= σ بازه‌های طبقه‌های تعیین‌شده بر اساس رتبه‌هایی که در روش «تاپسیس» به‌دست‌آمده، به‌شرح جدول ۴ است.

جدول ٤. محدودهٔ دسته‌بندی طبقات وب‌سایت‌ها

نام طبقه	بازهٔ رتبه
بسیار عالی	۶۲-۱
عالی	۲۲۹-۶۳
خوب	۳۹۶-۲۳۰
متوسط	۵۶۳-۳۹۷
ضعیف	۷۳۰-۵۶۴
بسیار ضعیف	۷۹۱-۷۳۱

نام طبقات، برچسبی به‌عنوان طبقهٔ وب‌سایت‌ها محسوب می‌شود که برای انجام محاسبات به‌ترتیب، عدد ۶ تا ۱ به این برچسب‌ها اطلاق می‌گردد. بنابراین، اعداد ۱ تا ۶ به‌عنوان خروجی در روش یادگیری ماشین مورد استفاده قرار می‌گیرند.

پیش‌پردازش داده‌ها در هر دو روش شبکهٔ عصبی و ماشین بردار پشتیبان مشابه است، بنابراین شرح پیش‌پردازش داده‌ها در این روش‌ها به‌صورت مستقل از روش در ادامه بیان می‌شود. در هر دو روش یادگیری ماشین با روش تحلیل اجزای اصلی، ابعاد بردار ورودی کاهش داده شده و داده‌ها نرمالایز گردیده است.

۳-٤-۱. پیش‌پردازش داده‌ها

معمولاً قبل از آغاز آموزش در روش‌های یادگیری ماشین یک سری پیش‌پردازش بر روی داده‌ها انجام می‌شود، از قبیل نرمالایزکردن داده‌ها و قراردادن آن‌ها در یک فاصلهٔ مشخص مانند بین ۱ و ۱-، که باعث افزایش کارایی روش مورد استفاده می‌شود. در بعضی وضعیت‌ها ابعاد بردار ورودی بسیار بزرگ است، اما اجزای بردارها همبستگی بالایی دارند. مفید است که در این وضعیت ابعاد بردار ورودی را کاهش دهیم. یک رویّهٔ مناسب برای این عمل استفاده از آنالیز اجزای اصلی است. این تکنیک سه تأثیر دارد: ۱) این عمل منجر به متعامدسازی اجزاء بردار ورودی می‌گردد، بنابراین آن‌ها هیچ وابستگی با یکدیگر نخواهند داشت؛ ۲) این کار اجزای متعامد حاصل را مرتب می‌کند و مؤثرترین آن‌ها در اولین اولویت قرار می‌گیرد؛ و ۳) در نتیجه، می‌توان اجزایی را که تغییرات کوچک‌تری در مجموعه داده‌ها ایجاد می‌کنند، حذف کرد.

(Shlens 2006)

در روش تحلیل اجزای اصلی[1] (PCA) پایه‌هایی که بهترین توزیع داده را داشته باشند، استخراج می‌شوند (همان). آنالیز اجزای اصلی برای ناهمبسته‌کردن داده‌ها مفید است و تبدیلی ایجاد می‌نماید که داده‌ها را در جهت‌های متعامد با بیشترین دامنه تغییرات تصویر می‌کند. همچنین، آنالیز اجزای اصلی با تقلیل ابعاد از n به p ($p \leq n$) برای کدکردن فشرده داده‌ها نیز قابل‌استفاده است. یعنی با استفاده از آنالیز اجزای اصلی، بدون ازدست‌رفتن قسمت عمده اطلاعات، اجزای وابسته به تغییرات کم، از داده حذف می‌شود (گشمرد و دیگران، ۱۳۹۰). حذف این عناصر اضافی سرعت آموزش شبکه‌های عصبی و ماشین بردار پشتیبان را افزایش می‌دهد و در صورتی که ابعاد مسئله خیلی بزرگ باشد، کارایی آن‌ها را بهبود می‌بخشد. علاوه بر این، بعد از اعمال این روش، داده‌های ورودی نرمال شده و دارای میانگین صفر و انحراف معیار واحد می‌گردند.

۳-٤-۲. روش شبکه عصبی مصنوعی

در این قسمت از یک شبکه پرسپترون چندلایه استفاده شده است که نشان‌دهندهٔ ارتباطی غیرخطی بین بردار ورودی‌ها و خروجی‌ها می‌باشد. این ارتباط از طریق اتصال نورون‌های هر لایه، در لایه‌های قبلی و بعدی است. خروجی نورون‌ها در ضرایب وزنی ضرب می‌شوند و به‌عنوان ورودی به تابع غیرخطی انتقال داده می‌شود. در مرحله آموزش، به پرسپترون داده‌های آموزشی آموزش داده می‌شود. سپس، وزن‌های شبکه به‌گونه‌ای تنظیم می‌شوند که خطای بین خروجی پیش‌بینی‌شده و هدف، کمینه گردد و یا اینکه تعداد دفعات آموزش به مقدار حداکثر ازپیش‌تعیین‌شده برسد. سپس، به‌منظور سنجش اعتبار آموزش انجام‌شده، تعدادی از داده‌های ورودی مورد آزمون قرار نگرفته به شبکه اعمال می‌شود. این مدل، شامل سه لایهٔ ورودی، مخفی و خروجی است که تعداد سلول‌های هر لایه به‌روش سعی و خطا مشخص می‌گردد. سیگنال‌های ورودی به‌وسیلهٔ ضریب‌های به‌هنجارکننده به مقدار یک نرمالیزه شده و بعد از محاسبات، خروجی به مقدار واقعی برگردانده می‌شود. همچنین، مقادیر اولیه وزن‌ها به‌صورت اتفاقی در نظر گرفته شده‌اند. این شبکه اصولاً بر مبنای الگوریتم پس‌انتشار خطا آموزش می‌بینند. بدین‌ترتیب که خروجی‌های واقعی با خروجی‌های دلخواه مقایسه می‌شوند و وزن‌ها به‌وسیلهٔ الگوریتم پس‌انتشار، به‌صورت تحت نظارت تنظیم می‌گردند تا الگوی مناسب به وجودآید.

در این قسمت، ابتدا مقادیر ویژگی‌های ۱۳گانهٔ هر وب‌سایت نرمالایز شده و مقادیر ورودی

1. principal components analysis

به یک بازهٔ بین [۱ و ۱-] نگاشت گردیده‌اند. در این مرحله از پیش‌پردازش داده‌ها، همان‌طور که بیان شد، از تحلیل اجزای اصلی استفاده گردیده است و در نهایت، همان‌طور که در ادامه بیان می‌گردد، تقسیم داده انجام شده و با شبکهٔ عصبی مصنوعی طراحی‌شده، مجموعه داده‌ها مورد آزمایش و آزمون قرار گرفته‌اند. برای انجام محاسبات یادگیری ماشین از نرم‌افزار Matlab استفاده گردیده است. بعد از نرمالایزکردن داده‌ها و کاهش وابستگی و ابعاد آن‌ها، مرحلهٔ بعدی تقسیم داده‌ها می‌باشد.

۳-٤-۳. تقسیم داده‌ها

تقسیم داده در زمانی که تنها یک مجموعه داده در دسترس است، کاری ضروری است که به‌صورت تقسیم داده‌ها به بخش‌های دوتایی یا سه‌تایی انجام می‌گردد و تعیین نسبت تعداد داده‌ها در مجموعه‌های داده‌ای متفاوت می‌باشد. از آنجا که با وجود بخش داده‌های ارزیابی امکان رسیدن به جوابی بهتر برای پیش‌بینی داده‌ها وجود دارد، باور بیشتر متخصصان بر این است که ۷۰ درصد داده‌ها به‌صورت تصادفی برای آموزش شبکهٔ عصبی، ۱۵ درصد داده‌ها به‌صورت تصادفی برای صحت و اعتبارسنجی و ۱۵ درصد دیگر داده‌ها به‌صورت تصادفی برای آزمایش و آزمون شبکهٔ عصبی مورد استفاده قرار بگیرد (Gallagher 1999). در مطالعهٔ حاضر برای روش شبکهٔ عصبی مصنوعی نیز روند تقسیم‌بندی داده‌ها به‌همین شیوه بوده است.

۳-٤-۳-۱. تعیین معماری شبکه

از یک شبکهٔ پیش‌خور دو لایه به‌عنوان ساختار شبکهٔ عصبی استفاده شده است که در لایهٔ مخفی اولیه از تابع انتقال «سیگموئید»[1] و در لایهٔ خروجی دومی از تابع انتقال خطی استفاده گردیده است. معماری این شبکهٔ عصبی پیش‌خور دولایه در شکل ۳ نشان داده‌شده است.

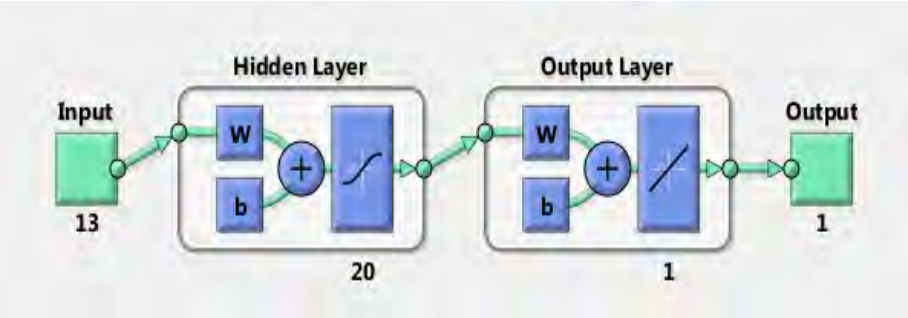
۳-٤-۳-۲. انتخاب روش یادگیری

شبکه با الگوریتم پس‌انتشار «لونبرگ- مارکوآرت»[2] آموزش داده‌شده است.

۳-٤-۳-۳. تعیین پارامترهای شبکهٔ عصبی

به‌صورت آزمون و خطا سعی در به‌دست‌آوردن پارامترهای شبکهٔ عصبی گردید و در نهایت، بدین‌نتیجه رسیدیم که از ۲۰ نورون به‌عنوان نورون‌های مخفی شبکه استفاده نماییم. شبکهٔ عصبی که در نهایت مورد استفاده قرار گرفت، در شکل ۳ نمایش داده شده است.

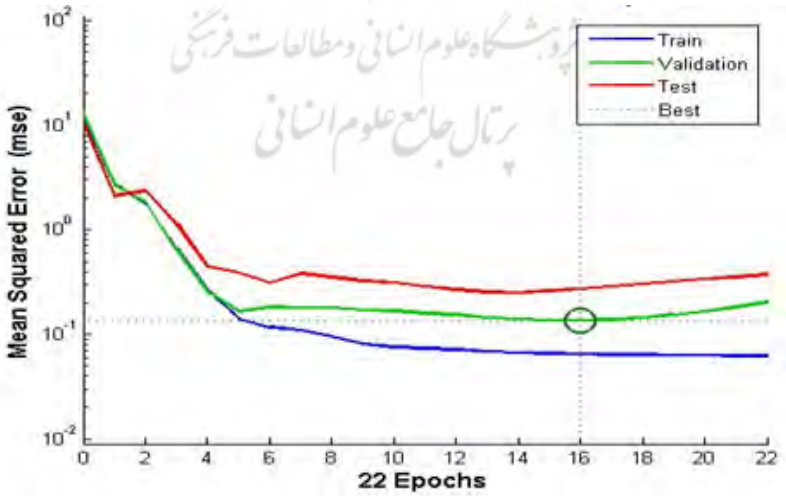
1. Sigmoid
2. Levenberg-Marquardt



شکل ۳. نمایی از شبکهٔ عصبی مورد استفاده در پژوهش

۳-٤-۳-٤. آموزش و آزمون داده‌ها

داده‌ها با استفاده از شبکهٔ عصبی طراحی‌شده آموزش داده شده و مورد آزمون قرارگرفته‌اند. معیار ارزیابی شبکه‌ای که طراحی گردیده، میانگین مربعات خطا (MSE) می‌باشد. این میانگین هرچقدر کوچک باشد، بهتر است. این میزان در شبکهٔ عصبی طراحی شده برابر با ۰/۰۶۲۰ می باشد که خروجی خطای آن در نمودار ۱ نشان داده شده است. همان‌طور که مشاهده می‌شود مقدار میانگین خطا در ابتدا زیاد بوده، ولی با تکرار آموزش و آزمون، میزان میانگین خطا کاهش پیدا کرده است. بعد از تکرار ۱۶، مقدار پیش‌بینی در مجموعه اعتبارسنجی به‌صورت ۶ بار به‌صورت مستمر اشتباه بوده است که به‌همین دلیل آموزش داده‌ها برای جلوگیری از برازش داده‌ها به پایان رسیده است و مقدار میانگین خطا با کمترین مقداری که تا آن لحظه داشته، برابر گردیده است.



نمودار ۱. نمودار میزان میانگین خطا

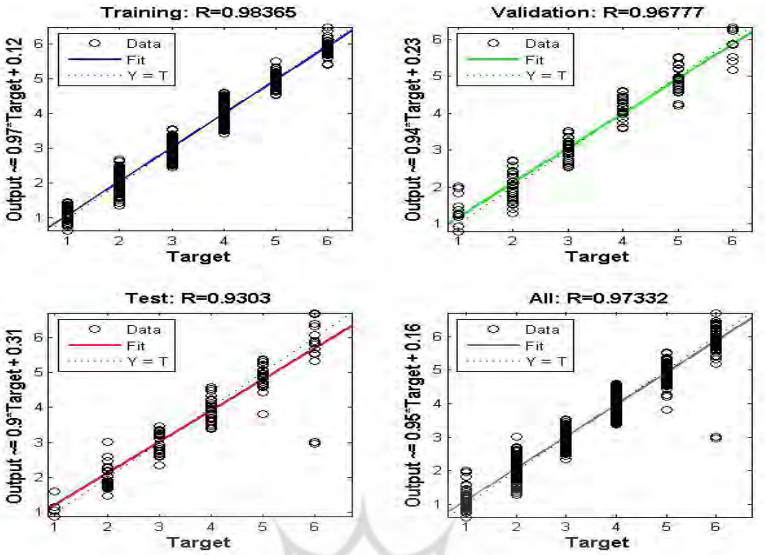
در نمودار ۱ می‌توان ضریب تشخیص سیستم را به‌عنوان معیاری برای اعتبارسنجی سیستم مشاهده نمود که مقدار آن برابر با۰/۹۷۳۳۲ است.

۳-٤-٤. روش ماشین بردار پشتیبان

در این روش که با استفاده از تابع‌های آماده در نرم‌افزار Matlab فراهم‌شده، ابتدا به‌دلیل ازدیاد ابعاد ویژگی‌های مسئله، ابعاد ورودی‌ها کاهش‌یافته و داده‌ها نرمالایز گردیده‌اند و بعد از تقسیم مجموعه داده، با استفاده از روش ماشین بردار پشتیبان یکی در برابر همه، k تا SVM تشکیل می‌شود، که k در اینجا تعداد کلاس‌هاست و در نهایت، میزان پیش‌بینی سیستم سنجیده می‌شود. بعد از پیش‌پردازش بردارهای ورودی با روش تحلیل اجزای اصلی شرح فعالیت‌های فرایند انجام این روش به‌شرح زیر می‌باشد:

۳-٤-٤-۱. تقسیم داده‌ها

در این روش مجموعه داده‌ها به دو بخش داده‌های آموزشی و آزمون تقسیم می‌گردد. در روش ماشین بردار پشتیبان، روش‌های مختلفی برای تعیین نسبت مجموعه داده آموزش به آزمون وجود دارد، که یا از طریق فرمول‌هایی بر اساس عواملی از قبیل پارامترهای دخیل در روش، تعداد کلاس‌ها و تعداد کل مجموعه داده تعیین می‌شوند یا به‌واسطهٔ روش‌های آزمون و خطا نسبت‌های مختلف بهینه‌ای برای مجموعه‌داده‌ها تعیین گردیده است (Guyon 1997). یکی از این بهترین نسبت‌های تقسیم داده‌ها برای مجموعه داده‌های آموزشی و آزمون نسبت ۶۰ درصد به ۴۰ درصد است (Andrew Ng, John Duchi). بر اساس نسبت پیشنهادی، در این پژوهش در حدود ۶۰ درصد از مجموعه داده‌ها (۵۰۰ نمونه) به‌عنوان مجموعه داده آموزش و بقیهٔ آن‌ها (۲۹۱) به‌عنوان مجموعه آزمون در نظر گرفته شده‌اند. البته، با روش آزمون و خطا و مشاهدهٔ میزان خطا در هر سری، می‌توان نسبت بهینهٔ این درصد را خاص هر مجموعه داده هم به دست آورد که در رابطه با این مجموعه داده، و میزان اندک‌بودن میزان خطای نهایی حاصل‌شده در این مقدار نسبت، از مناسب‌بودن آن به‌صورت شهودی هم می‌توان مطمئن بود. نمودارهای رگرسیون هر بخش از مجموعه داده‌ها و ضریب تشخیص نهایی آن‌ها در نمودار ۲ نشان داده شده است.



نمودار ۲. نمودارهای رگرسیون هر بخش از مجموعه داده‌ها و ضریب تشخیص نهایی

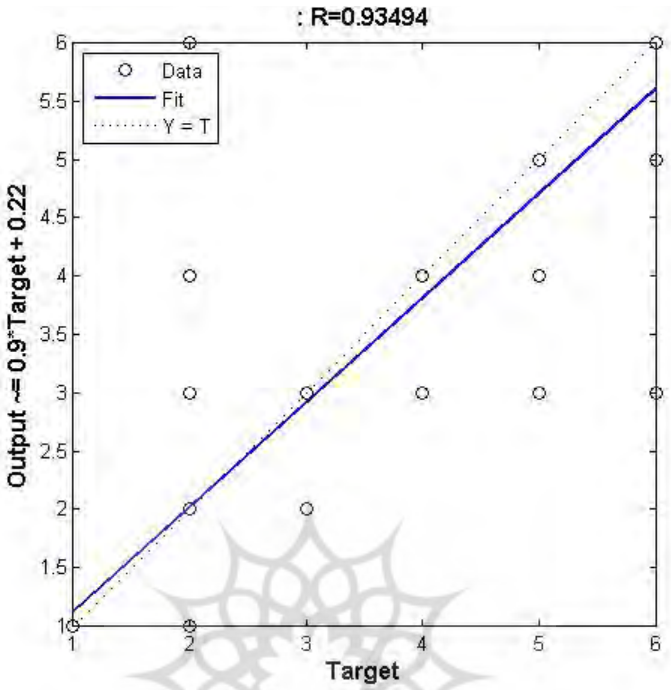
۳-٤-٤-۲. انتخاب تابع هسته و پارامترهای هسته

کل فرایند استفاده از ماشین بردار پشتیبان با دو تابع اصلی SVMClassify و SVMtrain در Matlab صورت گرفته است. در اینجا از هر دو تابع هسته RBF و چندجمله‌ای در ماشین بردار هسته‌ای استفاده شده و مقدار پارامتر آن‌ها همان مقدار پیش‌فرض در خود نرم‌افزار در نظر گرفته شده است.

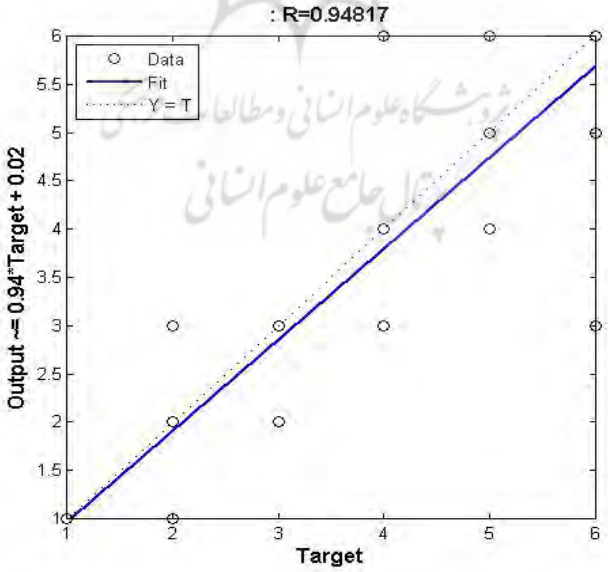
۳-٤-٤-۳. آموزش مجموعه داده‌های آموزش و آزمون داده‌ها

در نهایت، مجموعه داده‌های انتخابی برای آموزش مجموعه داده، آموزش داده شده و پس از تست داده‌ها، مقدار نرخ پیش‌بینی درست برای مجموعه داده‌های آزمون تعیین گردیده است. این مقدار برای ماشین بردار پشتیبان با تابع هسته RBF طراحی‌شده به‌صورت CCR= ۰/۸۰۷۶ می‌باشد. علاوه بر این، میزان ضریب تشخیص که در نمودار۳ نمایش داده شده برابر با ۰,۹۴۸۱۸ می‌باشد.

همچنین، مقدار نرخ پیش‌بینی درست برای ماشین بردار پشتیبان با تابع هسته چندجمله‌ای طراحی‌شده به‌صورت CCR= 0.8488 می‌باشد. علاوه بر این، میزان ضریب تشخیص برای این سیستم ماشین بردار پشتیبان برابر با ۰/۹۳۴۹۴ بوده که در نمودار ۴ نمایش داده شده است.



نمودار ٤. نمودار ماشین بردار پشتیبان با تابع هسته چندجمله‌ای

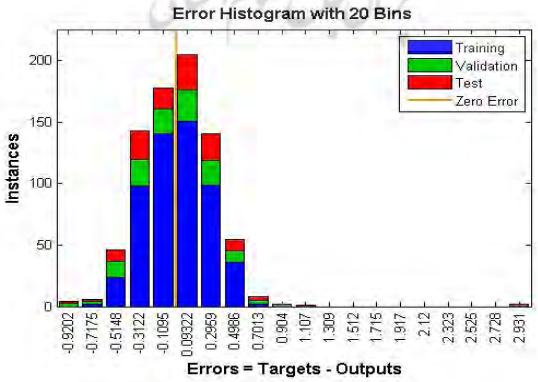


نمودار ۳. نمودار رگرسیون ماشین بردار پشتیبان با تابع هسته RB

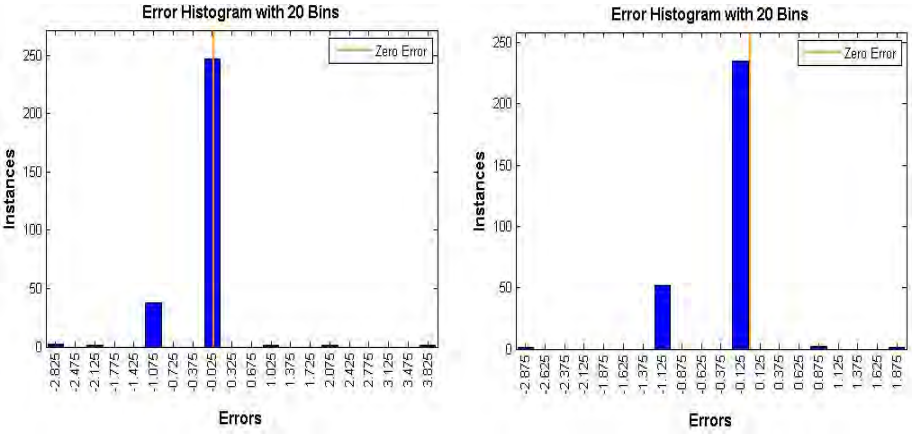
٤. یافته‌های پژوهش

بعد از به‌دست‌آوردن لیستی از ۱۳ شاخص مهم برای ارزیابی کیفیت وب‌سایت‌ها که در پاسخ به سؤال اول پژوهش صورت گرفت، و آموزش و اجرای شبکهٔ عصبی با داده‌های جمع‌آوری‌شده، می‌توان به پاسخ پرسش دوم رسید که امکان طراحی سیستمی به‌صورت خودکار با استفاده از روش‌های یادگیری ماشین امکان‌پذیر است. همان‌طور که در نمودار ۵ دیده می‌شود، میزان خطایی که در پیش‌بینی خروجی داده‌های جدید به‌عنوان داده‌های آزمون وجود دارد، بسیار اندک بوده و همگی به مقدار عددی صفر متمایل هستند. پس، می‌توان به این نتیجه رسید که با دقت بسیار زیاد امکان استفاده از روش‌های یادگیری ماشین، به‌ویژه شبکهٔ عصبی مصنوعی برای پیش‌بینی طبقهٔ وب‌سایت‌ها وجود داشته و می‌توان مطمئن شد که در صورت اضافه‌شدن وب‌سایتی با داده‌های جدید، طبقه این وب‌سایت بسیار به‌سرعت شناسایی شده و در طبقهٔ مورد نظر قرار می‌گیرد.

بعد از آموزش و اجرای ماشین بردار پشتیبان به‌عنوان یکی دیگر از روش‌های یادگیری ماشین می‌توان از نگاهی دیگر دوباره به پاسخ پرسش دوم رسید که امکان طراحی سیستمی به‌صورت خودکار با استفاده از روش‌های یادگیری ماشین امکان‌پذیر است. همان‌طور که در نمودار ۶ دیده می‌شود، میزان خطایی که در پیش‌بینی خروجی داده‌های جدید به‌عنوان داده‌های آزمون وجود دارد، در هر دو روش ماشین بردار پشتیبان چه با تابع هستهٔ چندجمله‌ای و چه با تابع RBF بسیار اندک بوده و همگی به مقدار عددی صفر متمایل هستند. پس، می‌توان به این نتیجه رسید که با دقت بسیار زیاد امکان استفاده از روش‌های یادگیری ماشین، از جمله ماشین بردار پشتیبان برای پیش‌بینی طبقهٔ وب‌سایت‌های موجود وجود دارد و اضافه‌شدن وب‌سایت جدید به مجموعه داده‌های موجود امکان‌پذیر است.



نمودار ۵. نمودار هیستوگرام خطا در شبکهٔ عصبی



نمودار ٦. به‌ترتیب از راست به چپ، هیستوگرام خطای ماشین بردار پشتیبان با تابع هسته RBF و تابع هسته چندجمله‌ای

با انجام مقایسه‌ای بین روش‌های انجام‌شده بر اساس شاخص‌های اعتبارسنجی که در اینجا از شاخص ضریب تشخیص بهره برده‌ایم، می‌توان به این نتیجه رسید که بهترین روش یادگیری ماشین در بین این روش‌ها، همان‌طور که در جدول ۵ مشاهده می‌شود، شبکهٔ عصبی مصنوعی بوده که ضریب تشخیصی بهتری از ماشین‌های بردار پشتیبان دارد و در کل، میان روش‌های ماشین بردار پشتیبان، تابع هستهٔ RBF ضریب تشخیص بهتری نسبت به تابع هستهٔ چندجمله‌ای دارد.

جدول ۵. ضرایب تشخیص روش‌های به‌کاربرده‌شده

روش به‌کاربرده‌شده	ضریب تشخیص
شبکهٔ عصبی	۰,۹۷۳۳۲
ماشین بردار پشتیبان (تابع هسته RBF)	۰,۹۴۸۱۷
ماشین بردار پشتیبان (تابع هسته چندجمله‌ای)	۰,۹۳۴۹۴

۵. نتیجه‌گیری و پیشنهادها

امروزه با توجه به اهمیت وب‌سایت‌های خبری به‌عنوان منابع مهمی از اطلاعات برای شهروندان یک جامعه، این پژوهش به مطالعه و بررسی این وب‌سایت‌ها و ارائهٔ راهکار مناسبی برای ارزیابی خودکار آن‌ها به‌منظور ایجاد پایه‌ای برای مقایسهٔ جایگاه آن‌ها پرداخته است.

نتایج اکثر پژوهش‌های پیشین انجام‌شده در ارتباط با وب‌سایت‌های خبری می‌تواند به‌عنوان

مبنایی در طراحی این دسته از وب‌سایت‌ها در ایران استفاده شود. در واقع، تا کنون خدمات اساسی شناسایی‌شده در ارتباط با نوع کسب‌وکار این نوع وب‌سایت‌ها، مبنا و گام اولیهٔ طراحان این وب‌سایت‌ها در راستای ایجاد آن به‌شمار می‌رود که می‌توانند به افزایش کیفیت و در نتیجهٔ آن، افزایش دفعات بازدید از وب‌سایت منجر شوند.اما پژوهش پیش‌رو، صرفاً از شناسایی این ویژگی‌ها فراتر رفته و به دنبال ارزیابی این وب‌سایت‌ها به‌صورت کاملاً خودکار است، به‌گونه‌ای که در هیچ‌کدام از مراحل جمع‌آوری اطلاعات، وزن‌دهی به معیارها و در نهایت، ارزیابی، هیچ عامل انسانی در این فرایند دخیل نباشد.

در این پژوهش ابتدا با استفاده از ادبیات تحقیق، ۱۳ عامل اصلی که در ارزیابی خودکار وب‌سایت‌های خبری می‌توانند مورد استفاده قرار گیرند، شناسایی شدند. این عوامل، همان‌طور که بیان شد، ۱۳ عامل مهم برای ارزیابی کیفیت وب‌سایت‌ها هستند که تنها با دانستن منبع جمع‌آوری داده‌های آن‌ها با استفاده از روش‌های خودکار، قابل گردآوری می‌باشند. قابل ذکر است که در این مقاله، برای اولین‌بار مجموعه‌ای از شاخص‌های ارزیابی خودکار وب‌سایت‌ها که می‌توانند در پژوهش‌های مشابه استفاده شود، در کنار یکدیگر بیان می‌گردد. در پژوهش‌های مشابه، یا جامعیت این شاخص‌ها به این تعداد نیست، یا در نهایت، در کنار این شاخص‌ها از عوامل دیگری که داده‌های آن‌ها به‌صورت دستی و توسط پرسشنامه از کاربران یا خبرگان حاصل شده بود، استفاده گردیده است. سپس، با استفاده از خزشگرهای خودکار از منابع موجود در وب، داده‌های مرتبط با این ۱۳ عامل، برای ۷۹۱ وب‌سایت خبری که توسط کاربران ایرانی بیشترین مشاهده را توسط آمار «الکسا» دارند، استخراج گردیده‌اند. بدین‌صورت، مجموعه اطلاعاتی که در این پژوهش برای طراحی سیستم خودکار می‌توانند مورد استفاده قرار گیرند، حاصل گردیده است. بعد از آن، نوبت به طراحی سیستمی برای ارزیابی خودکار وب‌سایت‌های فارسی رسید که ابتدا از ترکیب روش «تاپسیس» و آنتروپی اطلاعات، خروجی‌های سیستم برای یادگیری با نظارت استخراج گردیدند. اما، از آنجا که رتبه‌دهی عددی، صرفاً روش خوبی برای رقابت میان وب‌سایت‌ها نبود، از ۶ طبقهٔ بسیار ضعیف تا بسیار عالی برای نگاشت این اعداد در این طبقه‌ها استفاده شد و این ۶ طبقه به‌عنوان خروجی‌های یادگیری با نظارت در دو روش یادگیری ماشین، شبکهٔ عصبی و ماشین بردار پشتیبان استفاده شد.

همان‌طور که از یافته‌های این پژوهش برمی‌آید، بهترین روش برای ارائهٔ سیستم خودکار برای ارزیابی کیفیت وب‌سایت‌های خودکار، روش شبکهٔ عصبی به نسبت روش ماشین بردار پشتیبان می‌باشد. شبکهٔ عصبی با ضریب تشخیص ۰/۹۷۳۳۲ در صدر لیست روش‌های یادگیری ماشین مناسب برای ارزیابی کیفیت وب‌سایت‌ها در این پژوهش می‌باشد.

علاوه بر این، سیستم طراحی‌شده متکی بر ثابت‌بودن مقادیر و تعداد شاخص‌ها نیست و با گذشت زمان و تغییر در این عوامل، امکان ارزیابی و طبقه‌بندی بلادرنگ مجدد وب‌سایت‌ها امکان‌پذیر است. همان‌گونه که با افزایش روزافزون تعداد وب‌سایت‌ها، امکان ارزیابی بیش از ۷۹۱ وب‌سایت به‌کاررفته در سیستم هم به‌صورت لحظه‌ای امکان‌پذیر است و این سیستم وابسته به تعداد ثابتی از وب‌سایت‌ها نیست.

گفتنی است در هیچ‌یک از مقاله‌های مرورشده، برای ارزیابی خودکار وب‌سایت‌ها، جمع‌آوری اطلاعات به‌صورت خودکار نبوده و برای ارزیابی وب‌سایت‌ها در حوزهٔ وب‌سایت‌های ایرانی اولین‌بار است که برای پیش‌بینی رتبهٔ وب‌سایت‌های خبری در آینده از سیستم‌های مبتنی بر یادگیری ماشین استفاده گردیده است.

با ارائهٔ این سیستم، وب‌سایت‌های خبری از قبیل «رویترز»، «سی‌نت»، «قطره»، «روور»، «ویستا»، «اسکای‌اسپرت»، «خبر فارسی»، «خراسان‌نیوز»، «جام‌نیوز»، «مهرنیوز»، «افکارنیوز»، «تابناک»، «عصر ایران»، «ایسنا»، «خبر آنلاین»، «باشگاه خبرنگاران جوان»، «ساینس‌مگ» و «الجزیره» از قبیل سایت‌هایی هستند که در طبقهٔ بسیار عالی قرار گرفته‌اند. از نکات مهمی که می‌توان نتیجه‌گیری کرد، اینکه در پرمشاهده‌ترین وب‌سایت‌های باکیفیت خبری توسط کاربران ایرانی، در کنار وب‌سایت‌های خارجی، تعداد زیادی از وب‌سایت‌های داخلی قرار گرفته‌اند که خود می‌تواند نشان از حضور وب‌سایت‌های داخلی کشور در عرصهٔ رقابت در کنار وب‌سایت‌های نمونهٔ خارجی برای جذب بیشتر بازار و کاربران داخلی ایران باشد. علاوه بر این، نوع وب‌سایت‌های موجود در طبقهٔ بسیار عالی، مشتمل بر نوع وب‌سایت‌های خبری عمومی تا وب‌سایت‌های خبری در حوزهٔ خاصی از قبیل ورزشی، علمی یا فناوری می‌باشد و توزیعی از وب‌سایت‌های برتر خبری در هر حوزه را می‌توان به‌راحتی با این روش شناسایی کرد.

این سیستم به‌عنوان سیستمی برای رتبه‌دهی به وب‌سایت‌ها به‌صورت بسیار کلی، و نه‌تنها در زمینهٔ وب‌سایت‌های خبری، قابل‌تعمیم می‌باشد. برای اغلب وب‌سایت‌های اطلاعاتی مانند وب‌سایت‌های اجتماعی و وب‌سایت‌های سازمان‌ها در صورتی که فقط به‌منظور اطلاع‌رسانی باشند، می‌تواند به‌کاربرده شود. همچنین، برای دیگر وب‌سایت‌ها مانند وب‌سایت‌های حوزهٔ تجارت الکترونیک با اضافه‌کردن شاخص‌هایی از قبیل اعتماد و حریم شخصی و جمع‌آوری داده‌های مناسب می‌توان این وب‌سایت‌ها را هم با این روش رتبه‌بندی نمود.

در صورت اجرایی‌شدن سیستم ارائه‌شده، اولین سیستم خودکار رتبه‌دهی وب‌سایت‌ها را در ایران برای ارزیابی کیفیت آن‌ها به‌صورت تجاری خواهیم داشت. علاوه بر این، بنا بر تعیین ارزش شاخص‌ها و اهمیت نسبی آن‌ها، بهتر است وب‌سایت در جهت افزایش مقدار شاخص‌هایی که

بستگی به مالک وب‌سایت دارد، تلاش نماید. به‌عنوان مثال، سابقهٔ فعالیت وب‌سایت از مواردی نیست که مالک وب‌سایت قادر به افزایش میزان آن باشد، ولی موردی چون کاهش خطاهای HTML جزء مواردی است که در جهت افزایش رتبهٔ بهتر می‌توان به بهبود آن اقدام نمود.

همچنین، صاحبان وب‌سایت‌های خبری و سازمان‌های اطلاع‌رسانی که اولین هدف آنان کسب میزان مشاهده‌کنندهٔ بیشتر است، با استفاده از این سیستم می‌توانند جایگاه خود نسبت به بقیه رقبا را درک نموده و در جهت افزایش آن اقدام نمایند.

پژوهشگران در تحقیقات آتی خود می‌توانند از امکان افزایش شاخص‌های دیگری برای ارزیابی وب‌سایت‌ها که بیشتر متکی‌به‌خود وب‌سایت باشد، استفاده کنند مانند قابلیت سنجش رنگ وب‌سایتی با ایجاد سیستم ماشین بردار پشتیبان برای مشابهت رنگ بقیه وب‌سایت‌ها به وب‌سایت برتر، یا اضافه‌کردن شاخص کپی‌بودن یا نبودن محتوا که به‌صورت میانگین بر روی پست‌های اخیر خبری با استفاده از روش‌هایی مانند هشینگ[1] قابل انجام است، یا افزایش شاخصی که این سیستم را بیشتر بومی نماید، مانند مکان سرور که به‌نحوی می تواند به‌عنوان شاخص زمان پاسخگویی از سرور استفاده شود. علاوه بر این می‌توانند در روش‌های یادگیری ماشین از روش‌های تکاملی برای مشاهده بهبود میزان پیش‌بینی استفاده کنند یا اینکه با تمامی شاخص‌های اعتبارسنجی و مقایسهٔ آن‌ها برای اطمینان بیشتر، اعتبارسنجی سیستم انجام شود.

فهرست منابع

پاشازاده، فریبا، و محسن حاجی‌زین‌العابدینی. ۱۳۸۸. وب کیوام رویکردی کمّی جهت ارزیابی کیفی وب‌سایت کتابخانه‌ها. بابل: همایش منطقه‌ای وب‌سایت کتابخانه‌های دانشگاهی

چنگیز، نفیسه. ۱۳۸۵. ابزارهای کیفیت خودکار وب‌سایت با تأکید بر کاربردپذیری آن‌ها. تحقیقات کتابداری و اطلاع‌رسانی دانشگاهی ۴۵: ۱۴۷-۱۶۲

حسن‌زاده، محمد، و فاطمه نویدی. ۱۳۸۹. مقایسه کاربرد انواع روش‌های ارزیابی دسترس‌پذیری وب‌سایت‌ها (مورد مطالعه وب‌سایت وزارتخانه‌های دولت جمهوری اسلامی ایران) مجله تحقیقات اطلاع‌رسانی و کتابخانه‌های عمومی ۲۸: ۵-۶۱

حسین‌زاده، حسین. ۱۳۷۹. کتاب مدیریت کیفیت جامع نگاهی منسجم. تهران: انتشارات دانشگاه تهران،

زاهدی، شمس‌السادات. ۱۳۸۹. ارزیابی کیفیت وب‌سایت‌ها ابزارها و معیارها. فصلنامه مدیریت توسعه و تحول ۴: ۵-۱۶

1. Hashing

گشمرد، شهروز، علیرضا مهری دهنوی، و حسن ربانی. ۱۳۹۰. استفاده از الگوریتم آنالیز اجزای اصلی جهت تشخیص نرم‌افزاری حروف در متون فارسی. ویژه‌نامه مهندسی پزشکی مجله دانشکده پزشکی اصفهان ۲۹ (۱۷۴): ۱-۲۲

Bauer, Ch., and A. Scharl. 2000. Quantitive evaluation of Web site content and structure. *Internet research* 10 (1): 31-44.

Cho, J., and R. Adams. 2005. *Page Quality: In Search of an Unbiased Web Ranking*. In Proceedings of the 2005 ACM SIGMOD conference, Baltimore, Maryland, PP.551-562.

Derhami, V., E. Khodadadian, M. Ghasemzadeh, and A. M. Zareh Bidoki. 2013. Applying reinforcement learning for web pages ranking algorithms. *Applied Soft Computing* 3 (4): 1686-1692.

Dominic, P., H. Jati, P. Sellappan and G. K. Nee. 2011. A comparison of Asian e-government websites quality: using a non-parametric test. *International Journal of Business Information Systems* 7 (2): 220-246.

Fang, W. 2007. Using Google Analytics for Improving Library Website Content and Design: A Case Study. *Library Philosophy and Practice* 9 (2): 1-17.

Gallagher, M. R. 1999. Multi-layer perceptron error surfaces: Visualization, structure and modeling. PhD dissertation, University of Queensland.

Guyon I., 1997. A scaling law for the validation-set training-set size ratio. *AT&T Bell Laboratories* 1-11.

Hernandez B., J. Jimenex and M. J. Martin. 2009. Key website factors in e-business strategy. *International Journal of information management* 29 (5): 362-371.

Hsiu-Fen, L. 2010. An application of fuzzy AHP for evaluating course website quality. *Computers & Education* 54 (4): 877-888.

Li p., S. Yamada. 2009. Automated Web Site Evaluation – An Approach Based on Ranking SVM, IEEE/WIC/ACM International Joint Conferences on Web Intelligence and Intelligent Agent Technologie. vol. 3, pp: 34-37.

Ng Andrew and Duchi John., 2014. CS229 Machine Learning Online Course. Spring 2014, Online. Available from: http://cs229.stanford.edu/ (accessed 25.09.16)

Reesi Vananie, I, A. Manian, and M. A. Shafia. 2013. Designing a Fuzzy Inference System for Predicting the Implementation Success of ERP Solution. *Journal of Information Technology Management* 5 (1): 98-106.

Scowen G. 2007. Increased Website Popularity through Compliance with Usability Guidelines. A Doctoral Dissertation. University of Otago, New Zealand.

Serhat, A. and K. Cengiz. 2012, Evaluation of e-commerce website quality using fuzzy multi-criteria decision making approach. *IAENG International Journal of Computer Science* 39: 69-70.

Shlens, J. A. 2006. Tutorial on principal component analysis: derivation, discussion and singular value decomposition. [Online] Available from: http://www.snl.salk.edu/~shlens/notes.html. (accessed 11.02.08)

Statestia, The Statistics Portal. 2015. Global number of internet users 2005-2015, [Online] Available from: http://www.statista.com/statistics/273018/number-of-internet-users-worldwide/

Woolrych, A., K. Hornbæk, E. Frøkjær, and G. Cockton. 2011. Ingredients and Meals rather than Recipes: A Proposal for Research that does not treat Usability Evaluation Methods as Indivisible Wholes. *International Journal of Human-Computer Interaction* 27 (10): 940-970.

Younghwa L. and A. K. Kenneth. 2006. Investigating the effect of website quality on e-business success: an analytic hierarchy process (AHP) approach. *Decision support systems* 42 (3): 1383-1401.

Zahedi, Sh. 2010. Website Quality Assessment Tools and Measures. *Journal of Management Development and Evolution* 4: 16-5 (in Persian).

Zahran, D. I., H. A. Al-Nuaim, M. J. Rutter, and D. Benyon. 2014. A Comparative Approach to Web Evaluation and Website Evaluation Methods. *International Journal of Public Information Systems* 10 (1): 21-39.

Zhang Y., H. Cui, J. Burkell and R. E. Mercer. 2014. A Machine Learning Approach for Rating the Quality of Depression Treatment Web Pages, conferances 2014, PP. 192-212.

مولود آرمان

متولد سال ۱۳۶۶، دارای مدرک تحصیلی کارشناسی ارشد در رشتهٔ مدیریت فناوری اطلاعات از دانشگاه تهران است. ایشان هم‌اکنون پژوهشگر فناوری اطلاعات در مرکز تحقیقات مخابرات ایران است.

مدیریت اطلاعات، تحلیل داده و کسب‌وکار هوشمند از جمله علایق پژوهشی وی است.

بابک سهرابی یورتچی

متولد سال ۱۳۴۶، دارای مدرک تحصیلی دکتری در رشتهٔ علم مدیریت از دانشگاه لنکستر انگلستان است. ایشان هم‌اکنون استاد گروه مدیریت فناوری اطلاعات دانشگاه تهران و رییس آموزش‌های الکترونیک دانشگاه تهران است. مدیریت منابع اطلاعاتی، کسب‌وکار هوشمند، یادگیری الکترونیکی و کلان‌داده‌ها از جمله علایق پژوهشی وی است.

امیر مانیان

دارای مدرک تحصیلی دکتری در رشتهٔ مدیریت بازرگانی از دانشگاه تهران است. ایشان هم‌اکنون دانشیار گروه مدیریت فناوری اطلاعات دانشگاه تهران است.

تجارت الکترونیک، مبانی فلسفی سیستم‌های اطلاعاتی و سیستم‌های اطلاعاتی مدیریت از جمله علایق پژوهشی وی است.