سید محسن محمدی تاکامی کارشناسی خود را در سال ۱۳۸۱ در رشته برق - گرایش کنترل و کارشناسی ارشد خود را در سال ۱۳۸۴ در گرایش کنترل هر دو از دانشگاه خواجه نصیرالدین طوسی دریافت کرد. ایشان کارشناسی و کارشناسی ارشد خود را با رتبه نخست در گرایش کنترل به پایان رسانده است و همچنین پایان نامه ارشد ایشان عنوان برترین پایان نامه سال ۱۳۸۴ را اخذ نموده است. ایشان هم اکنون در زمینه‌های تحقیقاتی مرتبط با یادگیری ماشین شامل بینایی ماشین و تشخیص الگو مشغول به کار می باشد.

[8] A. Bosson et al., "Non-retrieval: Blocking Pornographic Images ," *Proc. of Int'l Conf. on the Challenge of Image and Video Retrieval, Lecture Notes in Computer Science,* Vol. 2383. Springer-Verlag, 2002.

[9] Girgis et al., "An Approach to Image Extraction and Accurate Skin Detection from Web Pages," *International Journal of Computer Science and Engineering* Vol.1 No. 2, Spring 2007.

[10] Z. Chen et al., "A Novel Web Page Filtering System by Combining Texts and Images ," *Proc. of the 2006 IEEE/WIC/ACM Int'l Conf. on Web Intelligence*, pp. 732-735, Hong Kong, 2006

[11] P.Y. Lee et al., "Neural Networks for Web Content Filtering ," *IEEE Intelligent Systems,* Vol. 17, No. 5, pp. 48-57, 2002.

[12] T. Bayes, "An Essay Towards Solving a Problem in the Doctrine of Chances," 53:370–418, 1763.

[13] I. Rish, "An Empirical Study of the Naive Bayes Classifier," *IJCAI 2001 Workshop on Empirical Methods in Artificial Intelligence*, 2001.

[14] Foerstner, "A Feature Based Correspondence Algorithm for Image Matching," *Int'l Archives of Photogrammetry and Remote Sensing,*. Vol. 26, No. 3, pp. 150-166, 1986.

[15] S. Theodoridis, K. Koutroumbas, *Pattern Recognition*, Academic Press, 2003.

[16] S. Agarwal, A. Awan, and D. Roth, "Learning to Detect Objects in Images via a Sparse, Part-Based Representation," *IEEE Trans. on Pattern Analysis and Machine Intelligence,* Vol. 26, No. 11, 2004.

[17] W.H. Ho and P.A. Watters, "Statistical and Structural to Filtering Internet Pornography," *Proc. IEEE Int'l Conf. System, Man and Cybernetics*, Vol. 5, pp. 4792-4798, Oct. 2004.

[18] P.Y. Lee, S.C. Hui, and A.C.M. Fong, "An Intelligent Categorization Engine for Bilingual Web Content Filtering," *IEEE Trans. on Multimedia*, Vol. 7, No. 6, pp. 1183-1190, 2005.

[19] R. Du, R. Safavi-Naini, and W. Susilo, "Web Filtering Using Text Classification," *Proc. IEEE Int'l Conf. on Networks*, pp. 325-330, 2003.

[20] http://www.consumersearch.com/parental-control-software.

[21] M. Hammami et al., "Adult Content Web Filtering and Face Detection Using Data-mining Based Skin-color Model ," *Proc. of the Int'l Conf. on Multimedia and Expo (ICME'04),* 2004.

[22] M.J. Jones and J.M. Rehg, "Statistical Color Models with Application to Skin Detection," Proc. of IEEE

[23] Conf. *on Computer Vision and Pattern Recognition*, pp. 274-280, June 1999.

[24] W.A. Arentz and B. Olstad, "Classifying Offensive Sites Based on Image Content," *Computer Vision and Image Understanding*, Vol. 94, Nos. 1-3, pp. 295-310, 2004.

[25] B. Liu, J. Su, Z. Lu, and Z. Li, "Pornographic Images Detection Based on CBIR and Skin Analysis," *Proc. of 4th Int'l Conf. on Semantics, Knowledge and Grid*, pp. 487-488, 2008.

[26] P. Kakumanua, S. Makrogiannisa, and N. Bourbakis, "A Survey of Skin-Color Modeling and Detection Methods," *Pattern Recognition*, Vol. 40, No. 3, pp. 1106-1122, 2007.

[27] V. Vezhnevets, V. Sazonov, and A. Andreeva, "A Survey on Pixel-Based Skin Color Detection Techniques," *Graphicon03*, pp. 85-92, 2003.

[28] J. S. Lee, Y. M. Kuo, P.C. Chung, and E. L. Chen, "Naked Image Detection Based on Adaptive And Extensible Skin Color Model," *Pattern Recognition*, Vol. 40, No. 8, pp. 2261–2270, 2007.

[29] A. Soria-Frisch, R. Verschae, and A. Olano, "Fuzzy Fusion For Skin Detection," *Fuzzy Sets and Systems*, Vol. 158, No. 3, pp. 325-336, 2007.

علی احمدی دارای مدرک مهندسی برق از دانشگاه صنعتی امیرکبیر در سال ۱۳۶۹و کارشناسی ارشد و دکترای هوش مصنوعی از دانشگاه ایالتی اوزاکای ژاپن در سال‌های ۲۰۰۱ و ۲۰۰۴ میلادی است. ایشان پس از پایان دوره دکتری به مدت سه سال در مرکز تحقیقات سیستم‌های نانو در دانشگاه هیروشیمای ژاپن مشغول پژوهش در زمینه "طراحی و پیاده‌سازی سخت‌افزاری مدل‌های یادگیرنده" هوشمند بوده اند که منجر به چاپ مقالات علمی در این زمینه گردیده است. از سال ۱۳۸۶ تاکنون ایشان به عنوان استادیار گروه کامپیوتر در دانشگاه صنعتی خواجه نصیرطوسی مشغول به فعالیت هستند. موضوعات پژوهشی مورد علاقه ایشان عبارتند از محاسبات نرم، مدل‌های هوشمند، طراحی‌های نرم-سخت افزاری، پیاده‌سازی الگوریتم‌های یادگیری.

مهدی زمانیان دارای مدرک مهندسی و کارشناسی ارشد در زمینه کنترل سیستم‌ها از دانشگاه صنعتی خواجه نصیر طوسی در سال‌های ۱۳۷۷و ۱۳۸۰ است. ایشان از سال ۱۳۸۰ تاکنون به عنوان مدرس در دانشکده برق دانشگاه خواجه نصیر طوسی مشغول به فعالیت بوده‌اند. زمینه‌های تحقیقاتی مورد علاقه ایشان سیستم‌های کنترلی خطی و غیر خطی، تعریف سیستم-ها، پیش‌بینی سری‌های زمانی، برنامه‌سازی کامپیوتری و طراحی مدارهای واسط کامپیوتری می‌باشد.

ناحیه پوست نسبت به کل تصویر) بعنوان ویژگی تصویری انتخاب شد. از طبقه‌بندی کننده بیزی برای دسته‌بندی ویژگی‌های متنی استفاده شد. پایگاه داده آزمون همان داده‌های مورد استفاده برای تست سیستم ما بود (۲۰۰ صفحه). نتایج طبقه‌بندی روی ۲۰۰ صفحه آزمون در جدول ۱۸ نشان داده شده است. همانطور که مشاهده می‌شود نتایج طبقه‌بندی سیستم پیشنهادی ما (جدول ۱۴) بطور کلی برتری دارند. در طبقه‌بندی صفحات مجاز (صفحات متعلق به کلاس صفر و بخشی از کلاس یک) دقت دو سیستم تقریبا در یک سطح است اما در طبقه‌بندی صفحات غیراخلاقی(صفحات متعلق به کلاس دو و بخشی از کلاس یک) که عمدتا دارای تصویر هستند، سیستم ما بطرز محسوسی دقت بهتری را ارائه می‌دهد.

جدول۱۸: نتایج دسته‌بندی صفحات آزمون بر اساس سیستم شبیه‌سازی شده از مقاله همامی [۴].

		طبقه‌بندی شده به		
		Cat0	Cat1	Cat2
صفحات ورودی تست	Cat0	(٪۸۲,۵) ۷۱	(٪۱۰,۵) ۹	(٪۷) ۵
	Cat1	(٪۲۱) ۹	(٪۷۰) ۲۱	(٪۹) ۳
	Cat2	(٪۴,۵) ۴	(٪۱۵,۵) ۱۳	(٪۸۰) ۶۵

۸- نتیجه گیری

در این مقاله با استفاده از ترکیب ویژگی های ساختاری، متنی و تصویری صفحات توانستیم یک روش پالایش با دقت قابل قبول را روی صفحات وب ارائه دهیم. روش پیشنهادی مبتنی بر یک دسته بندی ترکیبی با به کار گرفتن یک روش بیزی برای دسته بندی ویژگی های متنی و ساختاری و یک مدل شبکه عصبی برای دسته بندی ویژگی رنگ پوست و نیز ویژگی های عمومی و محلی تصاویر می باشد. این روش در مقایسه با روش های مشابه دارای این برتری است که ویژگی های متنی و ساختاری بیشتر و موثرتری را به کار می گیرد و برای تشخیص رنگ پوست از یک مدل شبکه عصبی با دقت بالا بهره می برد.

در این مسیر مسائل زیادی هنوز هست که می بایست مورد نوجه و مطالعه قرار گیرد از جمله در زمینه شناسایی صفحات وب فارسی، طبقه بندی ویژگی های متنی دارای دشواری هایی است. برای مثال جستجوی کلمات کلیدی می بایست با یک آنالیز صرفی و نحوی و بعضا معنایی همراه باشد. همچنین برای صفحاتی که از دو یا چند زبان استفاده می‌کنند می بایست نوع تشخیص پیچیده تری را بکار برد. برای مثال لیست سیاه کلمات و فرازهای کلیدی موجود و نیز روش جستجوی کلمات می بایست توسعه یابد. در این زمینه ما سعی کردیم از طریق یک مطالعه تحلیلی و آماری، ویژگی های مناسب و موثری را استخراج و با استفاده از چند نوع طبقه بندی کننده (بیزی، شبکه عصبی) نتایج را مورد ارزیابی قرار دهیم. دقت بدست آمده از این طریق اگرچه در حد ایده آل نیست اما با توجه به اینکه این نتایج مبتنی بر داده های واقعی و برگرفته از یک پایگاه داده شاخص است، قابل قبول می باشد. در زمینه ترکیب ویژگی های متنی و تصویری صفحه اگر چه تحقیقاتی تاکنون صورت گرفته است ولی یک کار جدی عملیاتی با دقت و قابلیت اطمینان بالا هنوز عرصه نشده است. در این پروژه سعی کردیم علاوه بر کنکاش روی جنبه های علمی قضیه و ارائه الگوریتم های جدید، یک پیاده سازی نرم افزاری کاربردی هم ارائه دهیم. مشکلات فراوانی در این مسیر وجود داشت که از جمله مهمترین آنها می توان به مساله قابلیت عملکرد برخط سیستم، عدم دسترسی به یک پایگاه داده نمونه کاملا نمایانگر، نحوه آموزش طبقه بندی کننده ها، تنوع بسیار زیاد صفحات و تصاویر از نظر سوژه، سایز، نوع رنگ، کیفیت تصویر، نوع قرار گرفتن سوژه ها، زبان بکار رفته، و ... اشاره کرد. با وجود موارد فوق ما قادر شدیم سیستمی را پیاده سازی کنیم که از طریق استخراج و شناسایی هر سه نوع ویژگی (ویژگی های متنی، ساختاری، تصویری) یک طبقه بندی قابل قبول را روی صفحات وب از نقطه نظر اخلاقی یا غیراخلاقی بودن انجام دهد. علاوه بر این میزان غیر اخلاقی بودن صفحات را از طریق تفکیک صفحات به سه کلاس صفر، یک، و دو تا حد قابل قبولی انجام می دهیم (کلاس صفر شامل صفحات مجاز، کلاس یک شامل صفحات غیراخلاقی با درجه پایین، کلاس دو شامل صفحات غیراخلاقی با درجه زیاد).

سپاسگزاری

این تحقیق بر اساس طرح پژوهشی تعریف شده توسط مرکز تحقیقات مخابرات ایران صورت گرفته است. نویسندگان این مقاله لازم می دانند مراتب تشکر و قدردانی خود را از این مرکز به خاطر پشتیبانی از این پروژه اعلام دارند.

مراجع

[۱] احمدی، علی و دیگران، "پالایش صفحات وب بر اساس تحلیل هوشمند محتوا"، *کنفرانس ملی انجمن کامپیوتر ایران*، اسفندماه ۱۳۸۷.

[۲] احمدی، علی، "گزارش فعالیت فاز یک، طرح پژوهشی: شناسایی صفحات وب غیر اخلاقی با استفاده از پروفایل صفحات"، *مرکز تحقیقات مخابرات ایران*، سال ۱۳۸۷.

[3] Z. Gao et al., "Applying a Novel Combined Classifier for Pornographic Web Filtering in a Grid Computing Environment," *Proc. of 12th Int'l Conf. on Computer Supported Cooperative Work in Design (CSCWD'2008)*, pp. 513-517, 2008.

[4] M. Hammami et al., "WebGuard: A Web Filtering Engine Combining Textual, Structural, and Visual Content-Based Analysis ," *IEEE Trans. on Knowledge and Data Engineering*, Vol. 18, No. 2, Feb. 2006.

[5] W. Hu et al., "Recognition of Pornographic Web Pages by Classifying Texts and Images," *IEEE Trans. on Pattern Analysis And Machine Intelligence*, Vol. 29, No. 6, pp. 1019-1034, June 2007.

[۶] فیلترینگ: مفاهیم و کاربردها.
http://www.knowclub.com/paper/?p=322.

[7] Family Online Safety Institute, FOSI: http://www.fosi.org.

مشاهده می‌شود نتایج به طرز محسوسی بهبود یافته است. این بهبود عمدتا بصورت کاهش در موارد FN هم برای صفحات غیراخلاقی و هم برای صفحات مجاز صورت می گیرد که ناشی از افزایش دقت در تشخیص صفحات دارای تصویر است.

جدول۱۴: ماتریس Confusion برای دسته‌بندی صفحات آزمون بر اساس ترکیب ویژگی‌های متنی و ساختاری و ویژگی‌های تصویری پوست و آبجکت خاص.

		طبقه‌بندی شده به		
		Cat0	Cat1	Cat2
صفحات ورودی تست	Cat0	(٪۸۶,۵) ۷۳	(٪۱۱,۵) ۱۰	(٪۲) ۲
	Cat1	(٪۱۹,۵) ۶	(٪۶۹) ۲۳	(٪۱۱,۵) ۴
	Cat2	(٪۳) ۲	(٪۵) ۴	(٪۹۲) ۷۶

لازم به ذکر است که جداول ۱۲ و ۱۴ دقت طبقه‌بندی را به تفکیک کلاس‌های سه‌گانه نشان می‌دهد اما چنانچه در طبقه‌بندی صفحات فقط غیراخلاقی بودن یا مجاز بودن آنها را در نظر داشته باشیم سیستم یک دقت طبقه‌بندی میانگین ۹۰٪ را برای استفاده از ترکیب ویژگی‌های ساختاری، متنی و تصویری بدست می‌دهد که در جداول ۱۵ و ۱۶ گزارش شده است.

جدول۱۵: نتایج طبقه بندی داده های آزمون بر اساس ویژگی های متنی-ساختاری.

		طبقه‌بندی شده به			
		غیراخلاقی درست (TP)	غیراخلاقی غلط (FP)	مجاز درست (TN)	مجاز غلط (FN)
صفحات ورودی	غیراخلاقی	٪۸۳	۰	۰	٪۱۷
	مجاز	۰	٪۲۵	٪۷۵	۰

جدول۱۶: نتایج طبقه بندی داده های آزمون بر اساس ترکیب ویژگی های متنی-ساختاری و ویژگی‌های تصویری.

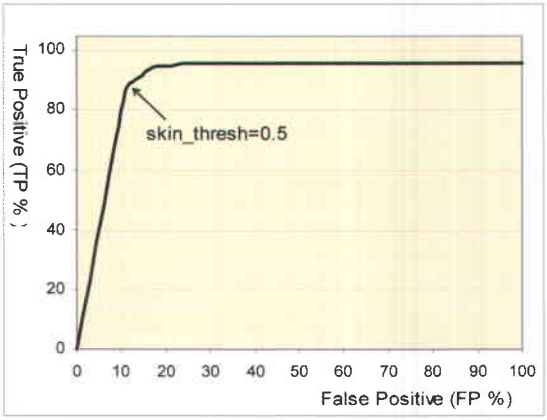
		طبقه‌بندی شده به			
		غیراخلاقی درست (TP)	غیراخلاقی غلط (FP)	مجاز درست (TN)	مجاز غلط (FN)
صفحات ورودی	غیراخلاقی	٪۹۰	۰	۰	٪۱۰
	مجاز	۰	٪۱۶	٪۸۴	۰

جدول۱۷: نتایج طبقه بندی سیستم نهایی روی داده های آزمون بر اساس مقادیر مختلف آستانه skin_thresh (تعداد صفحات غیراخلاقی: ۱۰۰ تعداد صفحات مجاز: ۱۰۰)

Skin_thresh	غیراخلاقی درست (TP)	غیراخلاقی غلط (FP)
٪۴۰	٪۸۶	٪۱۱
٪۵۰	٪۹۰	٪۱۳
٪۶۰	٪۹۲	٪۱۵
٪۷۰	٪۹۳/۵	٪۱۶
٪۸۰	٪۹۵	٪۱۸
٪۹۰	٪۹۵	٪۱۹

همانطور که در فصل ۶ توضیح داده شد برای تشخیص غیراخلاقی بودن تصویر از روی نواحی پوست موجود در آن، از یک حد آستانه skin_thresh استفاده می‌شود که نقش تعیین کننده‌ای در تشخیص درست یا نادرست غیراخلاقی بودن تصویر دارد. در ادامه آزمایش‌ها، دقت طبقه‌بندی سیستم را بر حسب مقادیر مختلف skin_thresh محاسبه کردیم که نتایج در جدول ۱۷ لیست شده است. تعداد کل صفحات غیراخلاقی مورد آزمون ۱۰۰ و تعداد کل صفحات مجاز نیز ۱۰۰ است. شکل ۵ نمودار ROC سیستم را بر اساس مقادیر مختلف skin_thresh نشان می‌دهد.

شکل۵- نمودار ROC سیستم بر اساس مقادیر مختلف skin_thresh



۷-۳- مقایسه با سایر سیستم‌ها

با توجه به پیچیدگی های موجود در دستیابی به راه حلی قابل قبول برای موضوع این پژوهش و وابستگی نتایج طبقه بندی به پارامترهای مختلف از جمله نوع انتخاب و توزیع دینای آموزش و تست، به دشواری می توان مقایسه ای دقیق با نتایج به دست آمده از کارهای دیگران انجام داد. همانطور که در مطالعه کارهای مرتبط (فصل ۲) آورده شده است روش های مختلفی تاکنون برای تشخیص صفحات غیراخلاقی مورد استفاده قرار گرفته و مقادیر متفاوتی برای دقت طبقه بندی گزارش شده است. مشکل اصلی در این میان عدم استفاده از یک پایگاه داده استاندارد برای ارزیابی الگوریتم های مختلف است. به تعبیر دیگر نتایج به دست آمده با به کارگیری پایگاه داده های متفاوت حاصل شده است و این امکان یک مقایسه دقیق را از بین می برد. از سوی دیگر امکان تایید صحت نتایج ادعا شده در مقالات با حتی نرم افزارهای موجود براحتی وجود ندارد. برای مثال در تست هایی که توسط تیم ما روی بعضی نرم افزارهای شرکت YangSky انجام گرفت، نتایج طبقه بندی به هیچ وجه با نتایج ادعا شده در شناسنامه نرم افزار مطابقت نداشت و این در حالی بود که ما صفحات ساده و بدون مشکلی را برای تست به کار بردیم.

با وجود مسائل فوق، برای اینکه مقایسه‌ای عملی بین عملکرد سیستم و سایر سیستم‌های موجود در این زمینه را بدست دهیم، سعی کردیم عملکرد یکی از سیستم‌های مطرح قبلی را شبیه‌سازی کرده و نتایج طبقه‌بندی داده‌های آزمون را بررسی کنیم. برای این کار سیستم ارائه شده در [۴] توسط همامی و دیگران انتخاب شد. توضیح مشخصات این سیستم بطور خلاصه در فصل ۲ این مقاله آمده است. با توجه به اطلاعات ارائه شده در مقاله ایشان، ۱۴ ویژگی مورد استفاده ایشان بعنوان ویژگی‌های متنی-ساختاری و یک ویژگی رنگ پوست (درصد

اسکریپت های متحرک و ثابت تخصیص داده شده است.

در این پایگاه ۶۰۰ صفحه در برگیرنده مطالب و تصاویر غیر اخلاقی گردآوری شده از حدود ۳۰۰ سایت انگلیسی می‌باشد و ۱۰۰ صفحه حاوی مطالب و متن های غیر اخلاقی به زبان فارسی است که در مجموع ۷۰۰ صفحه را شامل می‌شود. این صفحات از طریق جستجوی کلمات کلیدی یا فراوانی link به آنها شناسایی و انتخاب شده اند. ۵۹۵ صفحه دیگر در این پایگاه ماهیت اخلاقی و بهنجار دارند و دارای مضامینی چون اقتصادی-تبلیغاتی، علمی-پزشکی، آموزشی، ورزشی و سرگرمی هستند.

قبل از عملیات طبقه بندی، ابتدا صفحات پایگاه داده فوق بر اساس وابستگی به هر کدام از طبقات سه گانه برشمرده در بالا بصورت دستی برچسب خوردند. جدول ۱۱ توزیع داده‌ها در پایگاه داده نمونه را بر اساس زبان صفحه، نوع متن یا تصویر، مجاز یا غیراخلاقی بودن، و وابستگی به هر یک از کلاس‌های سه‌گانه معرفی شده در فصل ۵ نشان می‌دهد.

جدول ۱۱: توزیع داده‌ها در پایگاه داده نمونه صفحات.

	متنی	متنی-تصویری	مجموع
غیراخلاقی	۲۵۰	۴۵۰	۷۰۰
مجاز	۲۲۵	۳۷۰	۵۹۵
مجموع	۴۷۵	۸۲۰	۱۲۹۵
	انگلیسی	فارسی	مجموع
غیراخلاقی	۶۰۰	۱۰۰	۷۰۰
مجاز	۴۷۲	۱۲۳	۵۹۵
مجموع	۱۰۷۲	۲۲۳	۱۲۹۵
	انگلیسی	فارسی	مجموع
Cat0	۴۵۳	۱۱۲	۵۶۵
Cat1	۱۶۳	۱۹	۱۸۲
Cat2	۴۵۶	۹۲	۵۴۸
مجموع	۱۰۷۲	۲۲۳	۱۲۹۵

۷-۲- ارزیابی تجربی عملکرد سیستم

با توجه به اینکه هر یک از طبقه بندی کننده های مورد استفاده در سیستم نهایی بطور مجزا آموزش داده شده اند، نیازی به آموزش مجدد سیستم نهایی نیست. برای تست عملکرد سیستم، از پایگاه نمونه صفحات برشمرده در فوق، تعداد ۲۰۰ صفحه را بطور تصادفی برگزیدیم. این نمونه ها شامل انواع مختلف صفحات فارسی، انگلیسی و از هر سه کلاس صفر، یک، و دو و دارای تصویر یا بدون تصویر و تصاویر نیز دارای آبجکت مورد نظر (سینه) یا بدون آن بودند. از این میان ۱۰۰ صفحه غیراخلاقی و ۱۰۰ صفحه مجاز هستند (کلاس صفر: ۸۵ صفحه، کلاس یک: ۳۳ صفحه، کلاس دو: ۸۲ صفحه).

ابتدا عملکرد سیستم را تنها بر اساس ویژگی‌های متنی و ساختاری بررسی کردیم. نتایج در جدول ۱۲ نشان داده شده است. سه کلاس Cat0، Cat1، Cat2 بر اساس درجه غیراخلاقی بودن صفحات تعریف می‌شوند (ابتدای فصل ۵ مقاله). همانطور که مشاهده می‌شود بالاترین دقت طبقه‌بندی در کلاس Cat2 (صفحات کاملا غیراخلاقی) بدست آمده است و بیشترین طبقه‌بندی غلط از کلاس Cat2 و Cat0 به کلاس Cat1 و بالعکس صورت می‌گیرد که این به دلیل این است که کلاس Cat1 طبق تعریف در برگیرنده صفحات مشکوک و بینابینی است. دلیل دیگر بروز خطا، وجود صفحات با متن کم یا تصویر زیاد است که از طریق ویژگی‌های متنی قابل تشخیص و تمایز نیستند.

جدول۱۲: ماتریس Confusion برای دسته‌بندی صفحات آزمون بر اساس ویژگی‌های متنی و ساختاری.

		طبقه‌بندی شده به		
		Cat0	Cat1	Cat2
صفحات ورودی تست	Cat0	(٪۷۴,۶) ۶۳	(٪۲۲,۴) ۱۹	(٪۳) ۳
	Cat1	(٪۲۵) ۸	(٪۶۳) ۲۱	(٪۱۲) ۴
	Cat2	(٪۳) ۲	(٪۱۲,۲) ۱۰	(٪۸۴,۷) ۷۰

در مرحله بعد عملکرد سیستم در طبقه‌بندی صفحات آزمون را تنها بر اساس ویژگی تصویری رنگ پوست سنجیدیم. صفحاتی غیراخلاقی (positive) در نظر گرفته شده‌اند که حداقل دارای دو تصویر با سطح نواحی پوست بالاتر از ۵۰٪ باشند. نتایج در جدول ۱۳ گزارش شده است.

جدول۱۳: ماتریس Confusion برای دسته‌بندی صفحات آزمون بر اساس ویژگی رنگ پوست.

		طبقه‌بندی شده به			
		غیراخلاقی درست (TP)	غیراخلاقی غلط (FP)	مجاز درست (TN)	مجاز غلط (FN)
صفحات ورودی تست	Cat0	۰	(٪۱۱,۵) ۱۰	(٪۸۸,۵) ۷۵	۰
	Cat1	(٪۲۷) ۹	(٪۱۶,۵) ۵	(٪۲۸) ۹	(٪۲۸,۵) ۱۰
	Cat2	(٪۷۴) ۶۱	۰	۰	(٪۲۶) ۲۱

همانطور که مشاهده می‌شود در Cat0 (صفحات مجاز) خطای در تشخیص بخاطر وجود تصاویر مجاز دارای رنگ پوست (یا رنگ مشابه) در صفحات است که بعنوان تصاویر غیراخلاقی تشخیص داده شده‌اند. در Cat2 (صفحات کاملا غیر اخلاقی) بخاطر وجود صفحات متنی بدون تصویر، درصد زیادی از صفحات غیر اخلاقی تشخیص داده نشده‌اند. اما بیشترین خطای تشخیص در صفحات کلاس Cat1 ظاهر می‌شود (۴۵٪ یعنی جمع ستونهای غیراخلاقی غلط و مجاز غلط) یعنی صفحات با درجه غیراخلاقی کم و مبهم که از طریق فقط رنگ پوست، قابل تشخیص درست نیستند. میانگین دقت طبقه‌بندی در استفاده از ویژگی رنگ پوست روی کل ۲۰۰ صفحه آزمون با احتساب FP و FN برابر ۷۷٪ است.

در نهایت از طبقه‌بندی کننده ترکیبی شامل ویژگی‌های متنی، ساختاری و ویژگی‌های تصویری پوست و آبجکت خاص مطابق با نمودار شکل ۴ برای دسته‌بندی صفحات آزمون استفاده کردیم که نتایج طبقه-بندی در جدول ۱۴ منعکس شده است. در اینجا حد آستانه Th در رابطه ۴ برابر ۰/۲ و می‌نیمم سایز تصاویر مورد پردازش ۵۰×۵۰ پیکسل و مقدار skin_thresh برابر ۵۰٪ در نظر گرفته شده است. همانطور که

برای محدوده‌ی گرادیان بازه‌ی ۱۹ تا ۲۱ و مقدار بهینه برای طول پنجره ۱۰ به دست آمده است.

پس از استخراج مربع‌های حاوی تصاویر شی، این اشیا بایستی دسته‌بندی شوند. از سه روش مختلف برای خوشه‌بندی و کاهش تعداد استفاده می‌شود. این روش‌ها عبارتند از FCM[14]، PCA[15] و K-Means، که پس از آزمایشات تجربی زیاد، روش K-Means مناسبتر تشخیص داده شد.

معیارهای مختلفی را برای اندازه‌گیری فاصله‌ی بردارهای جدید با مراکز خوشه‌ها می‌توان استفاده کرد. بدیهی است روشی مناسب‌تر است که فاصله‌ی دو تصویر مشابه که تنها دارای جابجایی در راستای عمودی یا افقی هستند را مقدار کوچکی گزارش نماید. در این مرحله از معیار Tanimoto [۱۵] برای ارزیابی شباهت استفاده می‌شود.

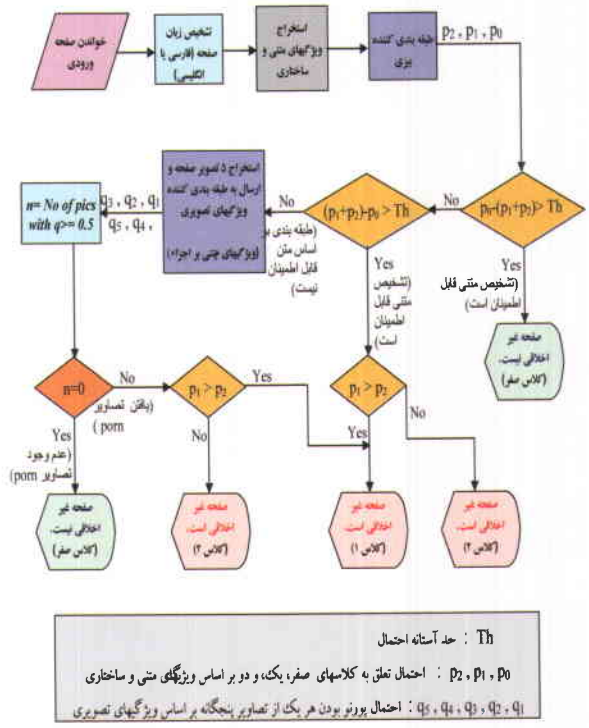
۶- ترکیب روش های طبقه بندی

فرایند نهایی طبقه بندی بصورت زیر است: ابتدا با توجه به کدینگ کلمات موجود در متن، زبان صفحه تشخیص داده می شود. سپس نتیجه طبقه بندی بیزی روی ویژگی های متنی و ساختاری صفحات به دست می‌آید. این نتیجه در قالب سه احتمال برای انتساب صفحه به یکی از طبقات سه گانه برشمرده در فصل ۵ است. احتمال تعلق به هر یک از طبقات را به ترتیب با p_0، p_1، و p_2 نشان می دهیم. چنانچه رابطه زیر برقرار باشد:

$$|(p_1+p_2) - p_0| < Th \qquad (۴)$$

(که در آن Th یک حد آستانه احتمال است) در آن صورت طبقه بندی انجام شده بر اساس ویژگی های متنی و ساختاری قابل اطمینان بوده و بسته به اینکه کدام یک از احتمالات سه گانه p_0، p_1، و p_2 بزرگتر باشد، کلاس مربوط به آن به عنوان طبقه نهایی انتخاب می شود. در صورتی که رابطه فوق برقرار نباشد نیاز به استفاده از ویژگی های تصویری برای طبقه بندی نهایی داریم. برای این کار پنج تصویر از صفحه بطور تصادفی یا بر اساس معیاری دلخواه انتخاب و از طریق طبقه بندی کننده تصویری (بر اساس ویژگی رنگ پوست و ویژگی مبتنی بر اجزاء استخراج شده از ناحیه پوست) میزان غیراخلاقی بودن تصاویر را در قالب پنج سطح احتمال (q_1 تا q_5) معین می کنیم. حال تعداد تصاویر غیراخلاقی (تصاویر با سطح احتمال بالاتر از۰/۵) موجود در میان پنج تصویر را محاسبه و با n نمایش می دهیم. اگر n صفر باشد (یعنی عدم وجود تصویر غیراخلاقی در صفحه) صفحه را صرفنظر از نتیجه طبقه بندی متنی بدست آمده، مجاز طبقه بندی می کنیم. اگر n بزرگتر از صفر باشد (یعنی وجود حداقل یک تصویر غیراخلاقی در صفحه) صفحه را غیراخلاقی طبقه بندی می کنیم حتا اگر نتیجه طبقه بندی متنی غیر از این بوده باشد. حال برای تعیین این که صفحه به کدام یک از کلاس های غیراخلاقی (کلاس ۱ یا کلاس ۲) متعلق است، از مقایسه p_1 و p_2 و این که کدام بزرگتر هستند، استفاده می کنیم. فرایند کامل طبقه بندی در نمودار شکل ۴ نمایش داده شده است.

در فرایند فوق، در قسمت کاربرد ویژگی‌های تصویر برای اینکه هر یک از ۵ تصویر استخراج شده از صفحه غیراخلاقی تشخیص داده شود می‌بایست یا سطح نواحی پوست موجود در آن از یک حد آستانه (skin_thresh) بزرگتر باشد و یا یک آبجکت خاص (سینه) در تصویر یافت شود و یا هر دوی این موارد صدق کند. تشخیص پیکسل‌های رنگ پوست در تصویر توسط شبکه عصبی برشمرده در بخش ۵-۲ صورت می‌گیرد و حد آستانه skin_thresh بصورت تجربی تعیین می‌گردد. تشخیص آبجکت سینه در تصویر توسط فرایند برشمرده در بخش ۵-۳ و شبکه عصبی بکار گرفته شده در آن انجام می‌پذیرد.



Th : حد آستانه احتمال

p_2 ، p_1 ، p_0 : احتمال تعلق به کلاسهای صفر، یک، و دو بر اساس ویژگیهای متنی و ساختاری

q_5 ، q_4 ، q_3 ، q_2 ، q_1 : احتمال پورنو بودن هر یک از تصاویر پنجگانه بر اساس ویژگیهای تصویری

شکل ۴- نمودار سلسله مراتب طبقه بندی نهایی سیستم.

همانطور که اشاره شد در طبقه بندی نهایی در صورتی از ویژگی های تصویری استفاده می کنیم که با استفاده از ویژگی های متنی و ساختاری به یک طبقه بندی قوی و مطمئن نرسیده باشیم و یک فاصله مطمئن بین کلاس برنده و برنده دوم وجود نداشته باشد. ضمنا برای طبقه بندی تصویری می توانستیم از روش دیگری مثل گرفتن میانگین روی پنج احتمال q_1 تا q_5 و مقایسه آن با یک حد آستانه Thresh به جای روش بکار گرفته شده فوق استفاده کنیم.

۷- نتایج تجربی و آنالیز خطا

۷-۱- پایگاه داده نمونه مورد استفاده

برای عملیات آموزش و آزمون سیستم ما از یک مجموعه نمایانگر از صفحات وب که توسط اعضای تیم این پروژه جمع‌آوری شده، استفاده کردیم. ابتدا مجموعه ای شامل ۵۰۰۰ صفحه وب توسط مرورگر Mozilla و به کمک موتور جستجوگر Google انتخاب و سپس از میان صفحات انتخاب شده، تحت یک فرایند اتفاقی ۱۰۷۲ صفحه انگلیسی و ۲۲۳ صفحه فارسی گزینش شدند که در مجموع ۱۲۹۵ صفحه را شامل می‌شود. به طور کلی صفحات وب استفاده شده از نظر ساختاری دو دسته هستند: اول، Plain-text که در آنها درصد زیادی از صفحه به متن و کلمات اختصاص دارد و دوم، Gallery که درصد عمده‌ای از فضای صفحه به نمایش تصاویر و یا Tag های مختلف از جمله کلیپ های فلش و یا

[14] Fuzzy C-means

[15] Principle Component Analysis

شکل۳- نمونه هایی از تشخیص رنگ پوست در تصاویر با به کار گیری الگوریتم پیاده سازی شده.

۵-۳- طبقه بندی کننده شبکه عصبی (تشخیص ویژگی های تصویری مبتنی بر اجزاء)

برای طبقه بندی بردارهای وبژگی بدست آمده از ویژگی های مبتنی بر اجزاء صفحات، از یک شبکه‌ی عصبی از نوع MLP با یکصد ورودی، یک خروجی و ۴۰۰ نرون در لایه‌ی میانی استفاده کردیم. تعداد نرون‌های لایه میانی از طریق آزمودن مقادیر مختلف ۱۰۰، ۲۰۰، ۳۰۰، ۴۰۰، ۵۰۰ نرون و مشاهده عملکرد شبکه با استفاده از MSE (می‌نیمم مربع خطا) تست و روش وارسی اعتبار انتخاب شد. برای تعداد ۵۰۰ نرون حالتی از بیش‌سازگاری[12] مشاهده گردید. توابع فعال سازی این شبکه عبارتند از تابع خطی برای ورودی، تابع tansig برای لایه‌ی میانی و تابع logsig برای خروجی. همچنین از روش آموزش traingdx از نرم‌افزار MATLAB که از شیب گرادیان با اندازه‌ی گام تطبیقی بهره می‌برد، استفاده شده است.

- **آموزش شبکه**

برای آموزش شبکه از یک مجموعه تصاویر شامل ۲۰۰۰ تصویر سینه مربوط به قبل و بعد از عمل جراحی پلاستیک که مجموعا ۴۰۰۰ آبجکت سینه را در بردارد، استفاده کردیم. تصاویر دارای سایز ۱۰۰×۱۰۰ و با زاویه های مختلف نسبت به دوربین هستند. با توجه به آنکه الگوریتم آموزش نیازمند تصاویر غیر اندام مورد نظر نیز می‌باشد، تصاویر دیگری نیز (شبیه یا غیر شبیه به سینه) از نواحی مربوط به شکم، پا، صورت و ... جمع‌آوری گردید. مراحل آموزش به شرح زیر است:

a. خواندن ۴۵۰ تصویر مثبت و ۵۲۰ تصویر منفی که تصاویر مثبت حاوی چهار شی مد نظر است و تمام این تصاویر از تصاویر استفاده شده برای تولید فرهنگ نمونه‌ها متمایز می‌باشند.
b. جارو کردن تصاویر با پنجره‌ی متحرک
c. پیدا کردن نقاط مرکز دایره‌های احتمالی با الگوریتم فورستنر
d. خوشه‌بندی مراکز برای یافتن چهار مرکز احتمالی
e. انتخاب مربع ۳۰×۳۰ در اطراف نقاط مورد نظر
f. کاهش دقت به مربع‌های ۱۵×۱۵
g. مقایسه‌ی کاندیداهای استخراج شده با نمونه های موجود در فرهنگ و اندازه‌گیری تشابه هر یک از این نمونه ها با کاندیدای مورد نظر
h. تولید بردار فاصله که هر عنصر آن نشان‌دهنده‌ی تشابه این کاندیدا با تک‌تک نمونه های موجود در فرهنگ است. (در این مرحله ۳۸۸۰ بردار آماده شده است: ۴×(۴۵۰+۵۲۰))
i. با استفاده از بردارهای تولید شده و مقدار هدف برابر با ۰/۹ برای موارد مثبت و ۰/۱ برای موارد منفی شبکه‌ی عصبی با ۳۰۰۰ نمونه از نمونه‌های ایجاد شده‌ی قبلی آموزش داده می‌شود.

- **تست شبکه**

مراحل طی شده جهت تست تصویر ورودی برای تشخیص آبجکت مورد نظر در آن، مشابه مراحل آموزش است بدین صورت که مراحل a تا h مجددا تکرار می شوند تا بردار ویژگی برای تصویر ورودی ساخته شود. سپس این بردار به شبکه‌ی عصبی اعمال می‌شود و مقدار خروجی شبکه اندازه‌گیری می‌شود. در صورتی که مقدار خروجی شبکه‌ی عصبی برای هر کدام از مربع‌های انتخاب شده بیش از ۰/۵ باشد این تصویر «حاوی شی مورد نظر» یا مثبت شناسایی می‌شود، در غیر این‌صورت نتیجه منفی است و اعلام می‌شود که «شی مورد نظر در تصویر یافت نشده است».

از ۶۰۰ تصویر نمونه شامل ۴۵۰ تصویر مثبت (دارای آبجکت سینه) و ۵۲۰ تصویر منفی (بدون آبجکت سینه) برای تست الگوریتم استفاده شد. تصاویر مثبت برگرفته از پایگاه داده ۲۰۰۰ تایی فوق بودند. نتایج طبقه‌بندی در جدول ۱۰ گزارش شده است.

جدول ۱۰: نتایج تشخیص آبجکت سینه توسط شبکه عصبی بر اساس ویژگی‌های مبتنی بر اجزاء

تعداد نرون لایه میانی	دقت طبقه‌بندی		تعداد خوشه‌ها در فرهنگ نمونه‌ها	تعداد بردارها در فرهنگ نمونه‌ها	داده‌های آزمون	داده‌های آموزش
	False Positive	True Positive				
۴۰۰	٪۲۴	٪۸۵	۵۰	۱۵۰	۶۰۰	۹۷۰
					۲۸۰^{+} ۳۲۰^{-}	۴۵۰^{+} ۵۲۰^{-}

- **عوامل موثر بر دقت عملکرد الگوریتم**

دقت عملکرد این الگوریتم وابسته به چندین پارامتر است که بایستی مقدار مناسبی برای هر کدام از آنها تعیین گردد. این پارامترها شامل موارد زیر هستند:

۱- تعیین اندازه‌ی مناسب برای پنجره
۲- تعیین مقدار مناسب برای آستانه‌ی استفاده شده در الگوریتم فورستنر
۳- تعیین روش خوشه‌بندی مناسب
۴- تعیین معیار مناسب اندازه‌گیری فاصله و شباهت در دو تصویر

اندازه‌ی پنجره و اندازه‌ی آستانه‌ی مناسب برای گرادیان برگردانده شده از الگوریتم فورستنر از هم مستقل نیستند. لذا برای انتخاب بهینه‌ی این دو پارامتر بایستی هر دو پارامتر با هم تغییر داده شوند. معیار بهینه‌گی، انطباق مراکز خوشه‌های تشخیص داده شده با اشیا مورد نظر است. اندازه‌ی پنجره از ۵ تا ۴۰ و اندازه‌ی پارامتر گرادیان از ۱۴ تا ۳۰ در دو مرحله با گام‌های بزرگ و کوچک[13] تغییر داده می‌شود. بهترین مقدار

[12] Overfitting

[13] Coarse and fine tuning

همانطور که از نتایج طبقه بندی مشخص است دقت طبقه بندی در مرحله اول در سطح بالاتری است و حال آنکه در مراحل دوم و سوم با توجه به اینکه داده های پالایش یافته تری را برای آموزش انتخاب کرده ایم، آموزش بهتری صورت گرفته است و انتظار داریم که دقت طبقه بندی هم بالاتر برود. علت این مغایرت این است که در مرحله دوم و سوم داده هایی که برای آزمون (تست) در نظر گرفته شده اند داده هایی هستند که رتبه های ضعیفی را در آموزش داشته اند یعنی داده هایی هستند که شباهت کمی را به داده های شاخص سه کلاس دارند و پس از عملیات آموزش هم این شباهت افزایش چشمگیری نمی یابد. به تعبیر دیگر در مراحل دوم و سوم اگر چه آموزش بهتری داریم ولی در عوض داده های آزمون داده های دشوارتری است و نتیجتا نتایج چندان بهبود نمی یابد.

جدول ۶: نتایج مربوط به طبقه‌بندی کننده بیزی بر اساس ویژگی‌های ساختاری و متنی برای داده آموزش ۲۰٪ و داده آزمون ۸۰٪.

	No of Samples			Accuracy
	Total	Train	Test	
Cat0	453	64	389	79%
Cat1	163	23	140	91%
Cat2	456	65	391	47%

جدول ۷: نتایج مربوط به طبقه‌بندی کننده بیزی بر اساس ویژگی‌های ساختاری و متنی برای داده آموزش ۵۰٪ و داده آزمون ۵۰٪.

	No of Samples			Accuracy
	Total	Train	Test	
Cat0	453	226	227	70%
Cat1	163	81	82	78%
Cat2	456	228	228	24%

جدول ۸: نتایج مربوط به طبقه‌بندی کننده بیزی بر اساس ویژگی‌های ساختاری و متنی برای داده آموزش۷۰٪ و داده آزمون ۳۰٪.

	No of Samples			Accuracy
	Total	Train	Test	
Cat0	453	362	91	31%
Cat1	163	130	33	45%
Cat2	456	364	92	36%

۵-۲- طبقه بندی کننده شبکه عصبی (برای تشخیص رنگ پوست)

شبکه عصبی چند لایه مورد استفاده برای تشخیص پوست دارای یک لایه پنهان و یک نرون در لایه خروجی می‌باشد. تعداد نرون‌های لایه میانی و تعداد ورودی‌ها (تعداد ویژگی‌های مبتنی بر رنگ) از جمله پارامترهای طراحی می‌باشند. در نرون‌ها از توابع غیرخطی سیگموید استفاده شده است. آموزش شبکه با روش لونبرگ مارکوارت و بهینه‌سازی شبکه با استفاده از روش وارسی اعتبار[11] با استفاده از داده‌های آموزش و تست صورت گرفته است. برای انتخاب تعداد نرون بهینه در لایه میانی، تعداد نرون‌های این لایه را از ۵ تا ۵۰ تغییر داده و برای هر تعداد نرون، شبکه ۷ بار و با شرایط اولیه وزن‌ها و بایاس‌های مختلف آموزش داده شد. میانگین خطای شبکه در ۷ آزمایش برای هر تعداد نرون و برای هر دو مورد ویژگی RGB و rgCbCr محاسبه شدند و مشخص شد که در حالت استفاده از ویژگی‌های RGB استفاده از ۳۰ نرون و در حالت استفاده از ویژگی‌های rgCbCr استفاده از ۴۵ نرون در لایه میانی شبکه عصبی مناسب خواهد بود.

برای تولید مجموعه‌های آموزش و تست ۱۶۵ عکس شامل نواحی متنوع از پوست و غیر پوست (در شرایط محیطی، نژاد و دوربین‌های مختلف) از وب جمع‌آوری شد. از این عکس‌ها ۲۰۸۷۷ پیکسل به صورت تصادفی انتخاب شده است که شامل ۱۱۵۷۲ پیکسل پوست و ۹۳۰۵ پیکسل غیر پوست می‌باشد. از این تعداد به صورت تصادفی، ۷۰٪ به عنوان داده آموزشی، ۲۰٪ داده تست و ۱۰ درصد داده ارزیابی در نظر گرفته شده‌اند.

خطای شبکه در دسته‌بندی داده‌های ارزیابی برای هر کدام از دو حالت استفاده از ویژگی‌های RGB و rgCbCr در جدول ۹ گزارش شده است. با توجه به نتایج فوق استفاده از بردار ویژگی rgCbCr منجر به خطای کمتر شبکه می‌شود اما شبکه پیچیده‌تری (با تعداد نرون لایه میانی بیشتر) مورد نیاز است. بنابراین انتخاب نهایی ما شبکه ای با ۴۵ نرون در لایه میانی و بردار ویژگی rgCbCr است. نمونه هایی از تشخیص پوست در تصاویر توسط الگوریتم فوق در شکل ۳ دیده می شود. مقایسه نرخ دسته‌بندی به دست آمده در این روش با نتایج به دست آمده در روش‌های بررسی شده در [۲۵]، [۲۶]، [۲۸] و به خصوص در [۲۷] که تشخیص پوست در جهت تشخیص عکس‌های غیر اخلاقی به کار برده شده است نشان می‌دهد که این روش از دقت قابل قبولی برخوردار است.

جدول ۹: نتایج طبقه‌بندی پوست برای دو نوع بردار ویژگی و تعداد نرون های مختلف در لایه میانی شبکه.

ویژگی مبتنی بر رنگ	تعداد نرون میانی	True Positive	False Positive
RGB	۳۰	%۸۴/۳۴	%۲۱/۲۰
rgCbCr	۴۵	%۸۴/۷۱	%۱۹/۰۷

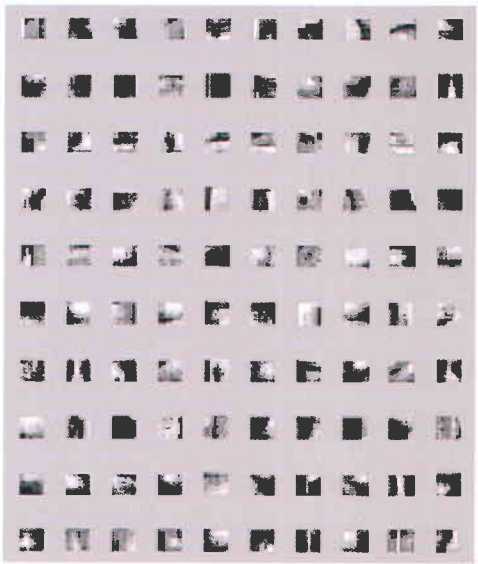
[11] Cross-validation

زمان تشخیص الگوی مورد نظر در تصویر انجام می شود. برای این منظور از روش دسته بندی K-means استفاده می کنیم که در نتیجه ۵۰ خوشه دسته بندی شده خواهیم داشت که میانگین بردار های موجود در هر خوشه، به عنوان نماینده خوشه برای عملیات بعدی در نظر گرفته می شود.

۴-۳-۲- تعیین بردار ویژگی های تصویر ورودی

برای تبدیل هر تصویر به یک بردار، مجددا نقاط ویژگی تمام تصویر از طریق جاروب کردن آن مشخص و با استفاده از الگوریتم فورستنر مرکز دوایر موجود در پنجره ها بازگردانده می شود. مشابه حالت قبل برای کم کردن تعداد نقاط ویژگی از الگوریتم FCM استفاده می کنیم و اطراف هر کدام از نقاط ویژگی یک همسایگی در نظر گرفته می شود و به تعداد نقاط ویژگی، بردارهای آبجکت تولید می شود. سپس برای طبقه بندی هر بردار به یکی از ۵۰ خوشه تعریف شده در فرهنگ نمونه ها از معیار فاصله همبستگی[10] استفاده می کنیم.

حال تصویر را به یک بردار تبدیل می کنیم بدین صورت که مولفه iام بردار حاصل، تعداد دفعاتی است که در آن شباهت هر یک از بردارهای آبجکت به نماینده خوشه iام بیشتر از شباهت به بقیه خوشه ها بوده است. بنابراین به ازای هر تصویر یک بردار ۵۰ تایی خواهیم داشت. نهاینا با داشتن این بردار ویژگی و با استفاده از یک مقدار آستانه برای تعداد شباهت ها، در مورد وجود یا عدم وجود آبجکت مورد نظر در تصویر، تصمیم گیری می کنیم.



شکل۲- تعدادی از بردارهای آبجکت تولید شده.

۵- روش های طبقه بندی

ما سه دسته متمایز را برای طبقه بندی صفحات در نظر می گیریم:

۱- طبقه صفر شامل صفحات مجاز (صفحات کاملا عادی بدون هر نوع آیتم غیراخلاقی)

۲- طبقه یک شامل صفحات غیراخلاقی با شدت کم (بدون کلمات رکیک یا تصاویر جنسی) یا صفحات عادی اما غیراخلاقی نما نظیر برخی صفحات پزشکی، صفحات آموزشی، مد، و غیره.

۳- طبقه دو شامل صفحات غیراخلاقی (با کلمات رکیک و تصاویر جنسی)

طبقه صفر همواره قابل دسترسی خواهد بود. طبقه یک بسته به نوع کاربر و به تشخیص مدیر سیستم قابل دسترسی خواهد بود. طبقه دو غیر قابل دسترسی است مگر برای کاربران خاص و محدود.

۵-۱- طبقه بندی کننده بیزی (Bayesian)

برای دسته بندی صفحات بر اساس ویژگی های ساختاری و متنی آنها از یک دسته بندی کننده بیزی استفاده کرده ایم [۱۲]. روش دسته بندی بیزی بر اساس نظریه بیز [۱۳] بنا نهاده شده است، به این ترتیب که احتمال اینکه رویدادی با بردار ویژگی های $F = (F_1, F_2, ..., F_n)$ در دسته C قرار بگیرد از رابطه (۱) به دست می آید:

$$p(C|F_1,\ldots,F_n) = \frac{p(C)\,p(F_1,\ldots,F_n|C)}{p(F_1,\ldots,F_n)}. \qquad (۱)$$

با فرض اینکه بردار ویژگی ها از یکدیگر مستقل باشند می توان نوشت:

$$p(C,F_1,\ldots,F_n) = p(C)\prod_{i=1}^{n} p(F_i|C). \qquad (۲)$$

بنا براین برای اینکه مشخص کنیم که رویدادی با بردار ویژگی های F، در کدام دسته قرار می گیرد، کافی است احتمال وجود روبداد در هر یک از دسته ها را با توجه به (۲) حساب کرده و بیشترین احتمال بیانگر دسته پیشنهادی خواهد بود:

$$\text{classify}(f_1,\ldots,f_n) = \text{argmax}_c\, p(C=c)\prod_{i=1}^{n} p(F_i=f_i|C=c). \qquad (۳)$$

بردار ویژگی های دسته بندی صفحات ترکیبی از ویژگی های ساختاری و متنی موثر است. احتمال وجود صفحه در هر یک از دسته های سه گانه فوق به صورت $Cat0$ تا $Cat2$ در نظر گرفته شده است. با توجه به اینکه ویژگی های رویداد مشاهده شده ممکن است دقیقاً بر هیچ یک از رویدادهای پیشین منطبق نباشد، فاصله رویداد با رویدادهای پیشین محاسبه شده و به نوعی درجه عضویت برای رویداد جدید با رویدادهای پیشین تخمین زده می شود.

- **آموزش و آزمون**

در مرحله اول از میان کل صفحات پایگاه داده نمونه برشمرده در فصل ۷ این مقاله، ۲۰٪ صفحات شاخص هر کدام از طبقات سه گانه را به عنوان داده آموزش انتخاب شده و بر اساس آنها داده های مرجع الگوریتم بیزی را شکل می دهیم و کلیه صفحات ۸۰٪ باقیمانده را به عنوان داده آزمون، بر اساس میزان شباهتشان به هر یک از دسته های فوق طبقه بندی می-کنیم. نتیجه طبقه بندی ۸۰٪ داده آزمون در جدول ۶ آمده است.

در مرحله دوم ۵۰٪ صفحاتی که در مرحله اول بیشترین امتیاز دسته خود را گرفته بودند به عنوان مرجع انتخاب شده و بر این اساس همه صفحات دوباره رده بندی شدند. نتیجه طبقه بندی ۵۰٪ باقیمانده بعنوان داده آزمون در جدول ۷ ارائه شده است.

در پایان ۷۰٪ صفحاتی که در مرحله پیش بیشترین امتیاز دسته خود را گرفته بودند به عنوان مرجع انتخاب شده و بر این اساس همه صفحات دوباره رده بندی شدند که نتایج دسته بندی ۳۰٪ باقیمانده به عنوان داده آزمون به طور خلاصه در جدول ۸ گزارش شده است.

[10] Correlation

جدول ۴: نمونه ای از کلمات شاخص استخراج شده برای طبقه های مختلف.

طبقه صفر	طبقه یک	طبقه دو
Health	sex	Porn
News	bikini	Sex
information	Breast	sexy
PEOPLE	Girls	Movies
cancer	Drunk	Pictures
sports	nude	Girls
care	videos	TEEN
Time	photos	Hot
Email	Hot	Mature

۴-۱-۲-۲- آنالیز همبستگی روی کلمات

یکی از روش‌های بررسی یک مجموعه ویژگی‌ها، بررسی همبستگی آنها با یکدیگر و بررسی همبستگی آنها با کلاس‌ها می‌باشد. در این دیدگاه، ویژگی‌هایی مناسب‌اند که دارای همبستگی پایین با یکدیگر و همبستگی بالا با کلاس‌ها باشند؛ در این حالت اطلاعات تکراری در ویژگی‌ها کم و تغییر کلاس داده‌ها توسط ویژگی‌ها قابل تشخیص است. در این جا به دلیل این‌که لغات (ویژگی‌ها) به صورت فیزیکی کاملا متفاوت از یکدیگر هستند، همبستگی بین ویژگی‌ها مورد بررسی قرار نمی‌گیرد.

برای محاسبه همبستگی میان لغات و یک کلاس خاص می‌بایست همه داده‌هایی که عضو آن کلاس نیستند (داده‌های دو کلاس دیگر) در یک کلاس قرار گیرند و در حقیقت مساله به حالت دو-کلاسه (به‌جای ۳-کلاسه) در آید. پس از انجام محاسبات همبستگی، لغاتی که بیشترین همبستگی را با کلاس خود داشتند شناسایی شدند که جدول ۵ تعدادی از کلمات دارای همبستگی بالاتر از ۰.۱۵ را نشان می دهد. با توجه به این اطلاعات تنها ۵۵ لغت از کل ۲۵۰ لغت در دسته‌بندی هر سه کلاس به طور موثر شرکت دارند که دلیل این امر می‌تواند کافی نبودن تعداد صفحات مورد آزمایش یا نمایانگر واقعی نبودن صفحات باشد.

جدول۵: تعدادی از لغات دارای همبستگی بیشتر از ۰.۱۵ با سه کلاس و میزان همبستگی آنها.

	Words	Correlation		Words	Correlation
1	'porn'	0.3405	14	0.1915	'housewive'
2	'xxx'	0.3301	15	0.1879	'mature'
3	'Care'	0.3113	16	0.1845	'fuck'
4	'Health'	0.2781	17	0.1832	'Breast'
5	'babes'	0.2749	18	0.1719	'sex'
6	'panties'	0.253	19	0.1645	'bukkake'
7	'Body'	0.2279	20	0.1605	'hardcore'
8	'homemade'	0.2268	21	0.1576	'pornstar'
9	'Fitness'	0.2217	22	0.1567	'hairy'
10	'milf'	0.2198	23	0.1567	'bisexual'
11	'group sex'	0.2141	24	0.1557	'Cancer'
12	'interracial'	0.2114	25	0.1541	'nasty'
13	'teens'	0.2079	26	0.1534	'ebony'

۴-۲- ویژگی رنگ پوست

تشخیص محدوده‌های پوست در عکس از جمله مسایلی است که به دلیل کاربردهای فراوانی که در شناسایی هویت ، ردیابی دست و صورت و فیلترینگ یافته است از اهمیت زیادی برخوردار شده است. روش‌های مختلفی برای تشخیص پوست پیشنهاد شده است که روش‌های مبتنی بر پردازش رنگ به علت سرعت بالا و دقت قابل قبول کاربرد وسیعی یافته‌اند. در این مقاله برای تمایز بخش‌های پوست و غیر پوست از ویژگی‌های مبتنی بر رنگ استفاده کرده‌ایم. وجود فضاهای رنگ متنوع مانند RGB، YCbCr، HSV، YIQ و مانند آنها و همچنین روش‌های مختلف دسته‌بندی داده‌ها اعم از روش‌های کلاسیک، هوشمند و آماری موجب شده است که روش‌های بسیار زیادی جهت تشخیص پوست مبتنی بر رنگ پیشنهاد شود [۲۵] و [۲۶] به بررسی این روش‌ها و مقایسه میزان دقت آنها می‌پردازند. استفاده از ویژگی‌های شامل رنگ در تشخیص پوست به دلیل سهل الوصول بودن موجب می‌شود که تشخیص با سرعت بیشتری انجام شود و همچنین رنگ دارای این مشخصه است که نسبت به تغییر جهت و تغییر مقیاس مقاوم است. اما نقطه ضعف استفاده از رنگ، تاثیر شرایط گوناگونی از جمله نور محیط و دوربین بر آن می‌باشد. در حالت استفاده از RGB حساسیت به شدت روشنایی بیشتر می‌شود. یک راه کاهش حساسیت به نور، تبدیل فضای RGB به فضای YCbCr و حذف مولفه شدت روشنایی (Y) و استفاده از مولفه‌های رنگینی (Cb و Cr) برای تشخیص می‌باشد. در طراحی حاضراستفاده از دو دسته ویژگی RGB و rgCbCr مورد بررسی قرار گرفته است. دسته ویژگی اول شامل مقادیر R، G، B آن پیکسل و چهار پیکسل همسایه (۱۵ مولفه) و دسته دوم شامل r و g نرمال شده و مقادیر Cb و Cr آن پیکسل و چهار پیکسل همسایه آن (۲۰ مولفه) می‌باشند.

۴-۳- ویژگی های تصویری مبتنی بر اجزاء

از نوعی ویژگی های مبتنی بر اجزاء [۱۶] برای تشخیص برخی آبجکت های جنسی (برای مثال سینه) در تصویر استفاده کرده ایم. ابتدا ناحیه پوست تصویر با استفاده از الگوریتم برشمرده در بالا شناسایی و سپس این ویژگی های تصویری از آن استخراج می شود. برای استخراج بردار نهایی ویژگی ها به چند مرحله عملیات پیش پردازش شامل تولید یک فرهنگ از نمونه آبجکت ها و سپس خوشه بندی آنها نیاز است که به احتصار در زیر شرح داده شده اند.

۴-۳-۱- تولید فرهنگ آبجکت های نمونه

ابتدا مجموعه ای از تصاویر نمونه که دارای آبجکت مورد نظر هستند انتخاب می شوند. هر تصویر با پنجره های ۱۰×۱۰ پیکسل جاروب و با استفاده از الگوریتم فورستنر [۱۴] مرکز دوایر موجود در پنجره بازگردانده می شود. پس از جاروب کردن تمام تصویر، از الگوریتم FCM برای کاهش تعداد ویژگی ها به تعداد متناسب با آبجکت موجود در تصویر استفاده می کنیم [۱۵] و با تعریف یک همسایگی ۳۰ × ۳۰ اطراف هر نقطه ویژگی، یک بردار آبجکت برای وارد کردن در فرهنگ نمونه ها می‌سازیم.

در اینجا ما از ۱۵۰ تصویر مثبت (تصاویری که حاوی سینه، با زوایا و اندازه های مختلف بوده اند) برای تولید فرهنگ نمونه ها استفاده کرده ایم. تعدادی از بردارهای آبجکت در شکل ۲ نشان داده شده است. در مرحله بعد بردارهای موجود در فرهنگ بر اساس شباهت دسته بندی می شوند. این امر برای کاستن از تعداد محاسبات و در نتیجه کاهش

به تعبیر دیگر تعداد کلمات مشخصه ای که در متن و دیگر مشخصات متن بکار رفته اند به تفکیک هر طبقه مشخص می شوند و چگالی هر ویژگی از طریق تقسیم تعداد کلمات مشخصه شمارش شده برای آن ویژگی بر تعداد کل کلمات شمارش شده برای آن ویژگی به دست می آید که این چگالی ها برای هر سه طبقه مشخص شده و برای هر یک از پنج مشخصه فوق به صورت جداگانه به دست آمده و نتیجتا ۱۵ ویژگی را به دست می دهند. نتیجه حاصل از طبقه بندی صفحات بر اساس این ویژگیها که با استفاده از روش طبقه بندی بیزی به دست آمد، نرخ طبقه بندی مناسبی را ارائه داد. این روش برای بسیاری از صفحات جواب درست می دهد ولی برخی از صفحات که موضوع آنها درباره کلماتی از یک طبقه مشخص است این روش را به اشتباه می انداخت برای نمونه، صفحه ای با موضوع حقوق جنسی کودکان به دلیل تکرار کلمات جنسی دارای چگالی کلمه زیادی در طبقه غیر اخلاقی است ولی موضوع این صفحات غیر اخلاقی نیست. نکته ای که درباره صفحات این چنین می-توان گفت، این است که در صفحاتی شبیه این تعداد کلمات مشخصه به کار رفته زیاد ولی تنوع آنها کم است و صفحه معمولاً پیرامون یک یا تعداد محدودی کلمه می گردد. برای تشخیص این صفحات ویژگی دیگری تولید شد که در آن تنوع کلمات در نظر گرفته می شود بدین صورت که تعداد کلمات متفاوت متعلق به هر طبقه شمارش و بر تعداد کلمات مجزای متن تقسیم شده و چگالی فراوانی کلمات صفحه در هر طبقه را مشخص می کند. با کاربرد این ویژگی برای طبقه بندی، شاخصهای طبقه بندی بهبود پیدا کردند.

در مرحله بعد با فعال کردن یک ویژگی ساختاری که تعداد پیوندهای غیر اخلاقی صفحه را مستقیما از طریق جستجوی پیوند در یک لیست سیاه از پیوندها به دست می داد، سه ویژگی مربوط به تعداد کلمات موجود در پیوندها، از ویژگی های فوق را حذف کردیم. به عبارت دیگر اکنون پیوندهای غیراخلاقی بصورت مستقیم و نه از طریق کلمات بکار رفته در آنها شناسایی می شوند. بدین ترتیب مجموع ویژگی های متنی نهایی ۱۵ ویژگی خواهد بود که در جدول ۳ نشان داده شده اند.

جدول ۳: ویژگی های متنی مبتنی بر کلمات مشخصه برای یک صفحه نمونه.

#	نام ویژگی
۱	درصد کلمات مشخصه طبقه صفر موجود در متن
۲	درصد کلمات مشخصه طبقه صفر موجود در عنوان
۳	درصد کلمات مشخصه طبقه صفر موجود در نام تصاویر
۴	درصد کلمات مشخصه طبقه صفر موجود در توضیح تصاویر یا
۵	چگالی فراوانی کلمات مشخصه متفاوت طبقه صفر در متن
۶	درصد کلمات مشخصه طبقه یک موجود در متن
۷	درصد کلمات مشخصه طبقه یک موجود در عنوان
۸	درصد کلمات مشخصه طبقه یک موجود در نام تصاویر
۹	درصد کلمات مشخصه طبقه یک موجود در توضیح تصاویر یا پیوندها
۱۰	چگالی فراوانی کلمات مشخصه متفاوت طبقه یک در متن
۱۱	درصد کلمات مشخصه طبقه دو موجود در متن
۱۲	درصد کلمات مشخصه طبقه دو موجود در عنوان
۱۳	درصد کلمات مشخصه طبقه دو موجود در نام تصاویر
۱۴	درصد کلمات مشخصه طبقه دو موجود در توضیح تصاویر یا پیوندها
۱۵	چگالی فراوانی کلمات مشخصه متفاوت طبقه دو در متن

۴-۱-۲- ایجاد پایگاه داده کلمات مشخصه

مبنای طبقه بندی صفحات در کار ما شمارش کلمات مشخصه مربوط به طبقه های مختلف در هر صفحه و بر اساس نتایج به دست آمده، سنجش میزان تعلق صفحه به هر یک از طبقه هاست. به این ترتیب آماده سازی مجموعه کلماتی که با شمارش آنها میزان تعلق صفحه به هر یک از طبقه ها را بتوان سنجید بسیار حائز اهمیت خواهد بود. این کلمات می بایست شاخص طبقه ها بوده و نمایانگر باشند. با توجه به اهمیت فوق العاده ای که ایجاد پایگاه کلمات دارد تلاش ویژه ای کردیم تا این پایگاه را با دقت مناسبی به دست آورده و تلاش کنیم تا بهترین کلمات ممکن به دست آیند. دو مسیر برای این کار طی شد و در نهایت تلفیقی از نتایج دو روش را مورد استفاده قرار دادیم که در ادامه به آنها می پردازیم.

۴-۱-۲-۱- سنجش آماری کلمات در صفحات نمونه

در این روش برای تهیه مجموعه کلمات، صفحات موجود در طبقه های مختلف را مورد بررسی قرار دادیم تا بتوانیم کلمات شاخص را تعیین کنیم. برای هر صفحه در طبقه مشخص، همه کلمات شمارش شده و تعداد هر کلمه به صورت جداگانه محاسبه شده و کلمات بر حسب تعداد مرتب شدند و این کار برای همه صفحات موجود در طبقه تکرار شد که حاصل آن جدولی بسیار بزرگ از کلمات و تعداد تکرار آنها در هر صفحه بود. با نگاهی به جدول تشکیل شده از کلمات و تعداد تکرار آنها می توان تا حدودی کلمات شاخص را تعیین کرد ولی لازم بود که روشی علمی برای این بررسی به کار گیریم. برای این کار در همه صفحات موجود در یک طبقه، چگالی کلماتی که در صفحات گوناگون تکرار شده اند را با هم جمع کردیم تا به چگالی تجمعی در همه صفحات برسیم. برای مثال برای چگالی کلمه jام در صفحه iام خواهیم داشت:

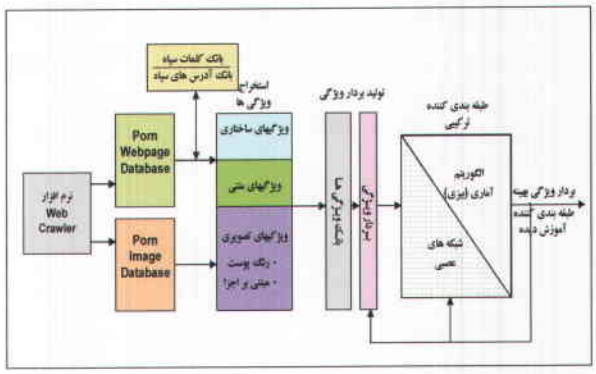
$$\rho_{ij} = \frac{N_{ij}}{N_i}$$

که در آن N_{ij} تعداد کلمه jام در صفحه iام و N_i تعداد کل کلمات صفحه iام است. برای محاسبه چگالی تجمعی کلمه jام در همه صفحات لازم است چگالی کلمه jام در همه صفحات جمع شود:

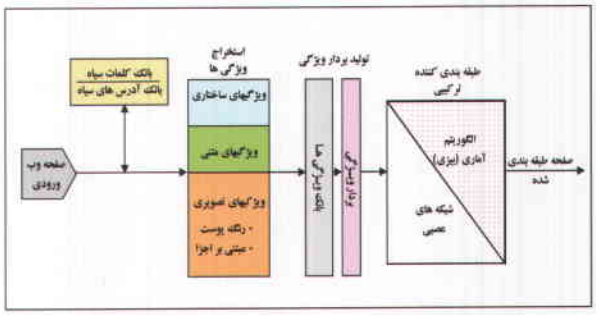
$$\rho_j = \sum_i \frac{N_{ij}}{N_i}$$

چگالی های تجمعی همه کلمات در همه صفحات بر اساس چگالی تجمعی مرتب شده و از آنها کلماتی که چگالی تجمعی بیشتری دارند پس از بررسی توسط فرد خبره انتخاب شده و کلمات شاخص آن طبقه را تشکیل می دهند.

برای انجام این کار برنامه ای نوشته شد که این کار را به صورت خودکار روی مجموعه ای از صفحات فارسی و انگلیسی انجام می داد و حاصل آن، فهرست مرتب شده ای از کلمات در هر طبقه بود که پس از حذف کلمات عمومی، جدولی را تشکیل می داد که بخشی از آن در جدول شماره ۴ نشان داده شده است. جدول مشابهی نیز برای کلمات فارسی تولید شده است. همانگونه که در جدول دیده می شود، کلمات به دست آمده تا حد زیادی شاخص طبقه مربوط به خود هستند.



شکل ۱- فلوچارت سیستم پیشنهادی برای تشخیص صفحات وب غیراخلاقی (مرحله آموزش).



شکل ۲- فلوچارت سیستم پیشنهادی (مرحله تست و کاربرد).

برای هر صفحه ورودی، این بردار ویژگی بهینه، ساخته شده و از طریق تشخیص گر آموزش دیده شده تشخیص می‌دهیم که صفحه ورودی غیراخلاقی هست یا نه. مرحله آموزش و تست سیستم در شکل‌های ۱ و ۲ نشان داده شده است.

نوآوری روش ما در مقایسه با کارهای مشابه در این زمینه را می‌توان بطور خلاصه در محورهای زیر دانست:

- در نظر گرفتن سه طبقه خروجی برای تقسیم‌بندی صفحات بر اساس درجه غیر اخلاقی بودن آنها بجای تنها یک خروجی مثبت یا منفی.
- در نظر گرفتن سه بانک کلمات مشخصه متناظر با سه طبقه خروجی و استفاده از این کلمات مشخصه برای استخراج ویژگی‌های مبتنی بر کلمات مانند متن صفحه، عنوان، نام تصاویر، پیوندها و ...
- استفاده از یک روش سنجش آماری و آنالیز همبستگی برای استخراج ویژگی‌های موثر و همگرا.
- استفاده از ویژگی‌های تصویری مبتنی بر اجزاء برای شناسایی تصاویر دارای آبجکت مشخص. این ویژگی در کنار ویژگی رنگ پوست که خود با روشی جدید استخراج می‌شود ویژگی-های تصویری کارامدی را تشکیل می‌دهد.
- استفاده از یک روش ترکیب سلسله مراتبی برای یکپارچه-سازی طبقه‌بندی کننده‌ها.

۴- استخراج ویژگی های مشخصه

در اینجا به شرح هر یک از سه دسته ویژگی های ساختاری، متنی، و تصویری و روش استخراج آنها می پردازیم.

۴-۱- ویژگی های متنی و ساختاری

ویژگی های متنی شامل کلیه ویژگی های مستخرج از متن می باشد که در جدول ۱ لیست شده اند. ویژگی های ساختاری شامل اطلاعات پروفایل، ساختار و اطلاعات جانبی صفحه است که ویژگی های مهم مورد استفاده، در جدول ۲ لیست شده اند. عملیات استخراج این ویژگی ها و پردازش روی آنها از طریق نرم افزار WebCrawler که به همین منظور در این تیم طراحی شده است، صورت می گیرد.

جدول ۱: ویژگی های متنی صفحات وب.

nwords	تعداد کلمات به کار رفته در صفحه (غیر از نام لینک ها و...)
nxwords	تعداد کلماتی از صفحه که در لیست کلمات سیاه هستند
pcxwords	درصد کلمات سیاه صفحه
nxkeywords	تعداد keywordهایی از صفحه که کلمه ای از لیست سیاه در نام آنها به کار رفته است
pcxkeywords	درصد keywordهای سیاه صفحه
nxdescripts	تعداد descriptionهایی از صفحه که کلمه ای از لیست سیاه در نام آنها به کار رفته است
ntitles	تعداد کلمات title صفحه
nxtitles	تعداد کلماتی از title که در لیست سیاه وجود دارند

جدول ۲: ویژگی های ساختاری صفحات وب.

nimages	تعداد تصاویر صفحه
nximages	تعداد تصاویری از صفحه که کلمه ای از لیست سیاه در نام آنها به کار رفته است
nlinks	تعداد لینک های صفحه
nxlinks	تعداد لینکهایی از صفحه که کلمه ای از لیست سیاه در نام آنها به کار رفته است
nxxlinks	تعداد لینکهایی از صفحه که در لیست لینکهای سیاه هستند
pcxlinks	درصد لینکهای سیاه صفحه
nkeywords	تعداد متا تگ های دارای keyword
nvideos	تعداد ویدیوهای صفحه
nframes	تعداد فریم های بکار رفته در صفحه
ncolors	تعداد رنگهای بکار رفته در صفحه
ndescripts	تعداد متا تگ های دارای description
nwarns	تعداد تگهای warning صفحه
nxwarns	تعداد warningهایی که کلمه ای از لیست سیاه در نام آنها به کار رفته است
ntooltips	تعداد tooltipهای تصاویر
nxtooltips	تعداد tooltipهایی از تصاویر که کلمه ای از لیست سیاه در نام آنها به کار رفته است

۴-۱-۱- تعیین ویژگیهای متنی مبتنی بر کلمات مشخصه

برای تشخیص طبقه هر صفحه نیاز به ویژگیهاییست که تا حد امکان نمایانگر[9] خصوصیات طبقه های مختلف باشند. در ابتدا از ویژگیهای چگالی استفاده کردیم که برای محاسبه آنها، تعداد کلمات موجود در هر یک از مشخصات صفحه یعنی: *متن*، *پیوندها*، *عنوان*، *نام تصاویر* و *توضیحات تصاویر* (و *پیوندها*) شمارش شده و از میان این کلمات تعداد کلماتی که متعلق به کلمات مشخصه هر طبقه خاص می باشند (یعنی متعلق به بانک کلمات مشخصه) نیز به صورت جداگانه حساب می شوند،

[9] Representative

گرفته شده است. از نقائص کار این است که هیچ گونه تحلیلی روی کارآمد بودن ویژگی‌های متنی و ساختاری از طریق آنالیز همبستگی یا روش‌های مشابه صورت نگرفته است. طبقه‌بندی کننده تصویری دارای دقت زیادی نیست زیرا فقط از نسبت تعداد پیکسل‌های پوست به کل پیکسل‌ها به عنوان ویژگی اصلی استفاده می‌کند. بعلاوه روش ترکیب ویژگی ها و ترکیب طبقه‌بندی کننده‌ها و نیز روش تشخیص پوست در تصاویر به روشنی بیان نشده است.

چن [۱۰] از ترکیب ویژگی های متن و تصویر در دسته بندی صفحات وب استفاده می کند. کار ایشان شامل سه مرحله است: دسته بندی صفحات از طریق تحلیل کلمات کلیدی، تحلیل جملات، تحلیل تصاویر. در کنار روشهایی چون تعیین مدل رنگ پوست جهت عملیات دسته بندی، از ویژگی های دیگری بر پایه ROIs[5] استفاده شده است. در ادامه به کمک یک الگوریتم Data Fusion به ترکیب ویژگی‌ها و سپس دسته بندی کننده صفحات به دو دسته Normal و Sensitive پرداخته است. برای تست این سیستم ۱۵۰۰ صفحه وب که بصورت دستی جمع آوری شده است مورد استفاده قرار گرفته و دقتی معادل ۹۱/۸٪ حاصل آمده است. استفاده از آدرسهای صفحات و کلمات کلیدی و عدم توانایی بالا در تشخیص صفحات مجاز از صفحات غیراخلاقی و همچنین پایگاه داده نامناسب و کوچک از ضعفهای این روش است.

اساس کار بوسون در [۸] استفاده از ویژگیهای تصویری است، ابتدا با طراحی یک فیلتر پوست و اعمال آن بر روی تصاویر به تشخیص مکان پوست و پیکسل های متعلق به پوست در تصویر می پردازد. سپس از ویژگی هایی چون مساحت، مرکز جرم، طول محور اصلی و فرعی و ... استفاده می شود و در نهایت یک بردار ویژگی با ۵ درایه تشکیل می شود. برای تست الگوریتم از ۱۰۰۰۵ تصویر متعلق به پنج دسته کلی استفاده شده و پس از اعمال چهار تکنیک دسته بندی، شبکه عصبی MLP[6] با دقت ۸۷/۲٪ کمترین خطای تشخیص را به خود اختصاص داده است. نقطه قوت این روش در طراحی دسته‌بندی کننده بهینه است ولی قسمت تشخیص پوست خصوصا نحوه استخراج ویژگیهای ذکر شده بصورت گویا تشریح نشده و در مورد ارزیابی ویژگیهای تصویری نیز کار مناسبی انجام نگرفته است.

لی [۱۱، ۱۸] از ویژگی‌های متنی و طبقه‌بندی کننده شبکه عصبی شامل SOM[7] و ANN[8] استفاده کرده است. ویژگی‌های متنی شامل عنوان صفحه، بخش قابل دیدن متن، متادیتا شامل توصیف و کلمات کلیدی، و عبارات توضیحی تصاویر می‌شوند. از یک مرحله پیش پردازش برای تبدیل ویژگی‌ها به بردارهای قابل اعمال به شبکه عصبی استفاده شده است. سیستم توسط یک مجموعه مجزا آموزش داده شده است. برای ارزیابی این سیستم از ۵۳۵ صفحه پورن و ۵۲۳ صفحه غیرپورن استفاده شده که دقتی معادل ۹۵٪ را به همراه داشته است. از نقاط قوت این روش امکان تشخیص صفحات دو زبانه (چینی و انگلیسی) و از نقاط ضعف آن می‌توان به عدم توانایی در دسته بندی صحیح صفحات گالری و پر تصویر نام برد چراکه تصمیم گیری فقط بر اساس ویژگی‌های متنی انجام می‌پذیرد.

گیرگیس [۹] یک سیستم برای استخراج تصاویر از صفحات وب و سپس پردازش آنها جهت تشخیص نواحی پوست موجود در تصاویر پیشنهاد کرده است. دو تکنیک تشخیص پوست براساس فضای رنگهای YUVو RGB معرفی شده و روشی پیشنهادی بر اساس روش YUV بهینه شده ارائه شده که ادعا می‌شود این روش بر مشکلاتی چون غیر یکنواخت بودن رنگ پیش زمینه در تصاویر غلبه کرده است. نقطه ضعف این روش هوشمند نبودن آن در شناسایی نواحی پوست است زیرا تنها با اعمال یک تبدیل بر روی تصویر، احتمال تشخیص نواحی شبه پوست بعنوان نواحی پوست بالا خواهد بود.

آرنتز و اولستاد [۲۳] مجموعه‌ای از ویژگی‌های تصویری را روی نواحی پوست پیوسته استخراج می‌کنند که عبارتند از: رنگ، بافت، شکل، مرکز ثقل، و مساحت ناحیه. تمرکز اصلی در این کار روی تشخیص ناحیه پوست است تا اصل تصویر. پایگاه داده این سیستم برای تعیین کارایی آن از ۲۰ صفحه وب که در حدود ۲۰۰۰ تصویر را در خود دارند، تشکیل شده که به دقتی معادل ۸۹٪ دست یافته اند. از مزایای این روش می توان استفاده از الگوریتم ژنتیک جهت انتخاب بهینه ویژگی‌ها نام برد که توانسته است تا حد محسوسی خطای سیستم را کاهش دهد. عدم توانایی در شناسایی و دسته بندی صفحات بدون تصویر و پایگاه داده بسیار کوچک از ضعفهای این روش محسوب می شود.

جونز و رگ[۲۲] از یک تشخیص‌گر ترکیبی بر مبنای متن و تصویر برای تشخیص صفحات غیراخلاقی استفاده می‌کنند. یکی از محدودیت-های کار آنها این است که از عملگر OR برای ترکیب دو طبقه‌بندی کننده استفاده کرده‌اند و نه از ترکیب ریاضی احتمالات. و این باعث افزایش نرخ پذیرش غلط (False positive) در سیستم می‌شود.

لیو و دیگران [۲۴] از یک روش بازیابی محتوای تصویر برای تشخیص تصاویر غیر اخلاقی موجود در صفحات وب استفاده کرده‌اند. ابتدا وجود انسان در تصویر شناسایی می‌شود و سپس از طریق آنالیز رنگ پوست، غیر اخلاقی بودن تصویر تشخیص داده می‌شود.

هر یک از روش های برشمرده در فوق دارای نقاط قوت و نقاط ضعفی از جمله بارگذاری بالای سیستم و سرعت پردازش کم، Over-blocking، misspelled words، عدم پوشش انواع صفحات، عدم امکان استفاده در یک محیط برخط، و غیره هستند که نیاز به استفاده از روشی کارآمدتر را ایجاد می‌کند.

۳- روش پیشنهادی

در جهت رفع نقائص کارهای موجود، در مقاله فعلی از یک روش پیشنهادی که دارای ترکیبی از هر سه نوع ویژگی های ساختاری، متنی، و تصویری صفحات است برای تشخیص صفحات غیراخلاقی استفاده می‌شود. پس از ایجاد یک بانک نمونه از صفحات و تصاویر غیراخلاقی که توسط یک نرم افزار WebCrawler و با اعمال معیارهای مشخص صورت می‌گیرد، در یک فرایند استخراج ویژگی ها، بانکی از ویژگی ها و سپس یک بردار ویژگی نهایی ساخته می‌شود. ویژگی های استخراجی به شرح زیرند: ویژگی های متنی (به شرح مندرج در جدول ۱)، ویژگی های ساختاری (به شرح مندرج در جدول۲)، ویژگی های تصویری شامل ویژگی های رنگ پوست و ویژگی های مبتنی بر اجزاء تصویر.

سپس از طریق یک تشخیص گر ترکیبی که از طبقات مختلف تشخیص شامل روش های آماری همراه با شبکه های عصبی تشکیل شده است ، این ویژگی ها مورد پردازش قرار می‌گیرند و در طی یک مرحله آموزش، بردار ویژگی و نیز تشخیص گر بهینه ساخته می‌شوند. در مرحله تست،

[5] Region of Interest
[6] Multi-layer Perceptron
[7] Self-organizing Map
[8] Artificial Neural Network

غیر اخلاقی صفحات می تواند تاثیرات مخربی بر کارکردهای اجتماعی، فرهنگی و شخصیتی ایشان بگذارد که این موضوع در سالیان اخیر توجه زیادی را در کشورهای مختلف جهان به خود جلب کرده است و اقدامات بسیاری برای دسته بندی صفحات از نظر اخلاقی بودن و محدودسازی دسترسی به صفحات غیر مجاز اخلاقی انجام شده است[۷،۶،۲]. نرم افزار های زیادی نیز برای این منظور تولید شده [۲۰،۲] که عمدتا دو روش اصلی را برای دسته بندی صفحات به کار می بندند: پالایش ایستا و پالایش پویا. مبنای پالایش ایستا بر اساس استفاده از پایگاه داده ای از نشانیهای غیر مجاز اینترنتی[1] و مسدود کردن دسترسی کاربر به آنهاست. با وجود سرعت به نسبت بالا، مشکل این روش نیاز برای به‌روزکردن مداوم فهرست نشانی‌هاست که باتوجه به سرعت و گستردگی تولید محتوا در وب مشکل قابل‌توجهی است. مشکل شناخته‌شده‌ء دیگر این روش‌ها میزان قابل‌توجه اشتباه‌گرفتن صفحه‌های مجاز (Over-blocking) مانند صفحه‌های مربوط به پزشکی، جامعه‌شناسی، ورزشی و یا مسدود کردن سایتی به دلیل وجود تنها یک صفحه غیر اخلاقی در آن است.

در روش پویا، دسته بندی و پالایش بر پایه تحلیل محتوا[2] انجام می‌شود. در تحلیل هوشمند محتوا با استفاده از روش‌های هوشمند نظیر روشهای یادگیرانه، روشهای داده‌کاوانه و مانند آنها به شناسایی صفحه و سپس دسته بندی آن می‌پردازیم که باعث می شود دقت سیستم در عین خودکاربودن شناسایی بالا رود ولی درعین حال بار پردازشی بیشتری نیز به همراه دارد که می‌تواند بویژه در سیستم‌های برخط[3] مشکل‌ساز باشد. در این روش شناسایی صفحه معمولا مبتنی بر استخراج ویژگی های مشخصه صفحه است. پس از بررسی کارهای انجام شده به این نتیجه رسیدیم که تلفیق متناسب و ابتکاری ویژگیهای مشخصه شامل ویژگیهای ساختاری، متنی، و تصویری می تواند دقت دسته بندی و به تبع آن پالایش را بهبود بخشد. ما در این مقاله و در ادامه کارهای قبلی مان [۱]، یک روش پیشنهادی برای بهبود عملکرد رده بندی و پالایش صفحات معرفی کرده ایم و نتایج دسته بندی را بر اساس الگوریتم بیزی و نیز شبکه های عصبی ارائه داده ایم.

ساختار این مقاله به شرح زیر است: ابتدا مروری بر کارهای مرتبط و نقاط ضعف و قوت آنها خواهیم داشت. سپس روش پیشنهادی ما و نقاط تمایز آن با کارهای موجود مورد بحث قرار خواهد گرفت. در فصل ۴ به توصیف ویژگی‌های مشخصه صفحات وب غیراخلاقی و روش استخراج آنها خواهیم پرداخت. در این قسمت ابتدا ویژگی‌های متنی و ساختاری و روش استخراج آنها به همراه سنجش آماری کلمات مشخصه و همچنین آنالیز همبستگی روی کلمات شرح داده می‌شوند و سپس ویژگی‌های تصویری شامل ویژگی رنگ پوست و ویژگی‌های مبتنی بر اجزاء معرفی می‌شوند. در فصل ۵ روش‌های طبقه‌بندی مختلف برای ویژگی‌های استخراج شده ارائه می‌گردد که شامل طبقه‌بندی کننده بیزی برای ویژگی‌های متنی و طبقه‌بندی کننده شبکه عصبی برای ویژگی رنگ پوست و ویژگی مبتنی بر اجزاست. در این فصل همچنین نحوه آموزش و آزمون طبقه‌بندی کننده‌ها به تفصیل توضیح داده می‌شود. فصل ۶ به چگونگی ترکیب طبقه‌بندی کننده‌ها می‌پردازد. در فصل ۷ نتایج تجربی آزمایش‌ها مورد بحث قرار گرفته و در فصل ۸ نتیجه‌گیری مقاله ارائه شده است.

۲- مروری بر کارهای مرتبط

با بررسی های به عمل آمده بر روی تعداد زیادی از مقالات موجود در زمینه پالایش هوشمند صفحات میتوان گفت که این مقالات به طور کلی با استفاده از دو روش به طبقه‌بندی صفحات وب می‌پردازند:

۱- استفاده از ویژگی های متنی صفحات.

۲- استفاده از ویژگی های متنی همراه با ویژگی های تصویری.

مقالاتی که از روش اول استفاده کرده اند [۳، ۱۷، ۱۸، ۱۹] بیشتر با اتکا به کلمات کلیدی موجود در صفحه و مقایسه آنها با یک جدول کلمات مرجع به طبقه بندی صفحه پرداخته اند. مقالاتی که در حوزه دوم فعالیت کرده اند [۴، ۵، ۱۰، ۲۲] علاوه بر ویژگی های متنی عموما با استفاده از تکنیکهای موجود در پردازش تصویر و شناسایی الگو، تصاویر موجود در صفحات وب را شناسایی کرده و براساس یکی از روشهای تشخیص پوست [۴، ۲۱، ۹]، تشخیص ROIs و نقاط مشخص در تصویر [۱۰]، استفاده از تبدیلات تصویر، و یا سایر روشهای پردازش تصویر به دسته بندی تصاویر و در نهایت دسته بندی صفحات وب می پردازند. در زیر به برخی از کارهای مهم انجام شده در این دو حوزه می‌پردازیم.

در کار هو [۵] ابتدا با استفاده از درخت C4.5 صفحات به سه دسته متن پیوسته، متن گسسته، و تصویری تقسیم می‌شوند. از یک شبکه شبه CNN[4] برای کشف ارتباطات معنایی در صفحات متن پیوسته و یک روش بیزی ساده برای تشخیص صفحات متن گسسته استفاده می‌شود. از یک هیستوگرام چندبعدی همراه با الگوریتم EM برای نمایش بردار رنگ پوست و ساختن یک مدل گاوسی ترکیبی جهت تشخیص پوست استفاده می‌شود. در انتها نتایج متنی و تصویری بر اساس تئوری بیزی با هم ترکیب می‌شوند. سیستم با حدود ۱۰۰۰ صفحه از موضوعات مختلف در برگیرنده سه دسته صفحات متن پیوسته، متن گسسته، و تصویری تست شده است و با استفاده از ترکیب طبقه‌بندی کننده‌ها میانگین دقت طبقه‌بندی ۹۱/۶٪ بدست آمده است. استفاده از تحلیل معنایی در تصمیم‌گیری بر روی محتوای متنی و همچنین استفاده از داده های موثر در پایگاه داده از نقاط قوت این روش محسوب می‌شود اما با توجه به اینکه روند الگوریتم به صورت متوالی انجام می‌پذیرد به نظر می‌رسد خطا های هر مرحله به مرحله بعد گسترش می‌یابد. در مرحله استخراج ویژگیهای پوستی نیز به نظر می‌رسد استفاده از هیستوگرام برای محاسبه پارامترهای مدل ترکیبی گزینه مناسبی نباشد، چرا که هیستوگرام علاوه بر زمان‌بری اصولاً یک عملگر global یا کلی محسوب می‌شود و در تصاویری که اشیاء شبه پوست در آنها وجود دارد به نظر نمی‌رسد این روش از دقت بالایی برخوردار باشد.

همامی و دیگران [۴] اساس کار دسته بندی صفحات وب را بر پایه استخراج ویژگی های متنی و نیز ویژگی های تصویری از تصاویر موجود در صفحات وب نهاده اند. در مقاله ایشان حدود ۲۰ ویژگی متنی و ساختاری صفحه استخراج می‌شود و همچنین از طریق تشخیص مکان پوست در تصاویر صفحه، به استخراج ویژگیهای تصویر می‌پردازد. ادعا شده است که با ترکیب سلسله مراتبی ویژگیهای متنی و تصویری دقت سیستم بطور محسوسی بالا رفته است. از نقاط قوت کار ایشان استفاده از ویژگی‌های ساختاری صفحات و تنوع ویژگی‌های متنی و ساختاری به کار

[1] URL Based Filtering

[2] Content Based Filtering

[3] Online

[4] Cellular Neural Network

مجله بین المللی علمی- پژوهشی
ارتباطات و فناوری اطلاعات

ITRC

جلد یکم - شماره ۴ - شهریور ماه ۱۳۸۸ (ش ص. ۴۳-۲۹)

پالایش هوشمند صفحات وب با استفاده از ترکیب ویژگی های متنی، ساختاری و تصویری

علی احمدی
دانشگاه خواجه نصیر طوسی
دانشکده برق و کامپیوتر
ahmadi@eetd.kntu.ac.ir

مهدی زمانیان
دانشگاه خواجه نصیر طوسی
دانشکده برق و کامپیوتر
mzamanian@eetd.kntu.ac.ir

محسن محمدی تاکامی
دانشگاه خواجه نصیر طوسی
دانشکده برق و کامپیوتر
m2takami@gmail.com

تاریخ دریافت: ۱۳۸۸/۱/۱۶- تاریخ پذیرش: ۱۳۸۸/۶/۲۴

چکیده – استفاده از روش های هوشمند برای تحلیل صفحات وب اخیرا مورد توجه قرار گرفته است و یکی از کاربردهای آن در پالایش صفحات غیر اخلاقی است. روش های موجود بیشتر بر مبنای تحلیل ویژگی های متنی و در برخی موارد تصویری صفحه است اما هر یک مشکلاتی را دارند که از آن جمله میزان خطای بالا در تشخیص صفحات سفید (Over-blocking) است. در این مقاله یک روش هوشمند جدید برای پالایش صفحات غیراخلاقی را پیشنهاد کرده ایم که با استفاده از هر سه نوع ویژگی ساختاری، متنی و تصویری و ترکیب سلسله مراتبی آنها از طریق یک طبقه بندی کننده بیزی و نیز شبکه های عصبی، یک طبقه بندی هوشمند با دقت بالا را به دست می-دهد. در بخش ویژگی های متنی و ساختاری، با استفاده از یک بانک کلمات مشخصه و آنالیز همبستگی و تحلیل آماری ویژگی‌های موجود، مجموعه ای کارامد از ویژگی‌ها انتخاب می‌شوند. در مورد ویژگی‌های تصویری، علاوه بر کاربرد ویژگی رنگ پوست بصورت پیکسلی، از مجموعه‌ای ویژگی‌های مبتنی بر اجزاء تصویر نیز استفاده شده است. الگوریتم روی ۱۲۹۵ صفحه وب شامل ۷۰۰ صفحه غیراخلاقی (دارای متن، تصویر، یا هر دو) انگلیسی و فارسی و ۵۹۵ صفحه مجاز شامل صفحات پزشکی، سلامت، ورزشی و غیره مورد آزمایش قرار گرفته و دقت طبقه بندی کلی حدود ۹۰٪ را به همراه داشته است.

کلیدواژه – طبقه بندی هوشمند، پالایش صفحات غیر اخلاقی، تشخیص رنگ پوست، شناسایی صفحات وب، ویژگی های متنی و تصویری.

۱- مقدمه

گسترش چشمگیر وب و افزایش روزافزون تولید صفحات در آن مسائل جدیدی را با خود به همراه داشته است. یکی از این مسائل افزایش صفحات ناهنجار دارای محتوای قبیح نگارانه، خشونت، نژادپرستانه و غیره و نیاز به دسته بندی و پالایش آنها برای کاربردهای خاص است. از موضوعات بسیار مورد توجه در این زمینه، دسته بندی صفحات بر اساس میزان اخلاقی بودن آنهاست. دسترسی افراد به ویژه کودکان به محتوای