ارائه الگوریتمی بهینه و دقیق مبتنی بر خوشه‌بندی مارکوف برای شناسایی روبات‌های وب

مهدیه ذبیحی *، دانشجوی کارشناسی ارشد مهندسی کامپیوتر(نرم‌افزار)، گروه کامپیوتر، دانشگاه بین‌المللی امام رضا(ع)، مشهد،
m.zabihi@imamreza.ac.ir

مجید وفایی‌جهان، عضو هیئت علمی گروه مهندسی کامپیوتر(نرم‌افزار)، دانشگاه آزاد اسلامی، مشهد،
vafaeijahan@mshdiau.ac.ir

چکیده: الگوریتم خوشه‌بندی مارکوف، به‌عنوان یکی از معروف‌ترین روش‌های خوشه‌بندی در تحقیقات بیوانفورماتیک، الگوریتمی مبتنی بر گراف است، که با استفاده از دو عملگر ماتریسی به یافتن خوشه‌های طبیعی موجود در یک گراف می‌پردازد. هدف از این مقاله، ارائه یک فاز پیش‌پردازش مناسب جهت تولید ماتریس ورودی و ایجاد تغییراتی در بدنه اصلی الگوریتم است. روش پیشنهادی برای خوشه‌بندی بازدیدکنندگان یک وب سایت، به دو گروه روبات‌های وب و کاربران انسانی، استفاده و نتایج تحت معیارهای صحت و دقت ارزیابی می‌شوند. در پایان سرعت، تعداد خوشه‌های نهایی و مقدار آزمون همبستگی الگوریتم پیشنهادی با الگوریتم RMCL(نسخه تغییریافته الگوریتم خوشه‌بندی مارکوف) مقایسه می‌شود. ارزیابی‌ها نشان می‌دهد که سرعت و نتیجه آزمون همبستگی برای هر دو الگوریتم تقریبا برابر است اما تعداد خوشه‌های تولیدی توسط روش RMCL بیشتر بوده و همین نشان‌دهنده دقت کم‌تر این الگوریتم نسبت به روش پیشنهادی است.

کلمات کلیدی: خوشه‌بندی مبتنی بر گراف، خوشه‌بندی مارکوف، روبات‌های وب، آزمون همبستگی، شباهت کسینوسی.

۱. مقدمه

الگوریتم خوشه‌بندی مارکوف (MCL) با داشتن ماتریس تلاقی یک گراف و شبیه‌سازی جریان‌های موجود در آن، به خوشه‌بندی داده‌های اولیه می‌پردازد. در این مقاله سعی بر این است که با ارائه یک فاز پیش‌پردازش مناسب جهت تعریف ماتریس تلاقی گراف ورودی و همچنین ایجاد تغییراتی در بدنه الگوریتم اصلی، به روشی بهینه و دقیق برسیم. در پایان نتایج الگوریتم پیشنهادی با نسخه‌ای تغییریافته از الگوریتم خوشه‌بندی مارکوف به نام RMCL[۱۰] مقایسه می‌گردد. توانایی دو الگوریتم در خوشه‌بندی تعدادی از بازدیدکنندگان وب به دو خوشه کاربران انسانی و روبات‌های وب، به چالش کشیده شده و نتایج با هم مقایسه می‌گردد.

۲. الگوریتم خوشه‌بندی مارکوف

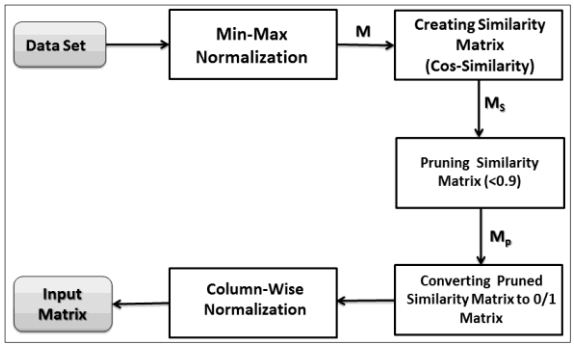
ورودی الگوریتم یک ماتریس نرمال‌سازی شده ستونی است که جمع عناصر هر ستون برابر ۱ است. بدنه اصلی الگوریتم حلقه‌ای تکرارشونده است که از دو عملگر بسط و تورم و یک مرحله هرس‌سازی تشکیل می‌شود. عملگر بسط معادل به توان رساندن ماتریس تلاقی میانی (معمولا توان دو) است که باعث توسعه جریان مابین نودهای گراف ورودی می‌گردد. عملگر تورم نیز از دو مرحله ضرب Hadamard ماتریس تحت پارامتر r و نرمالسازی ستونی آن ساخته می‌شود. مرحله هرس‌سازی جهت صفر کردن درایه‌های کوچک استفاده می‌گردد و در افزایش سرعت الگوریتم تاثیر بسزایی دارد. در پایان و بعد از همگرایی الگوریتم به حالت سکون نهایی، از ماتریس حاصل خوشه‌ها استخراج و تفسیر می‌گردند[۶،۱۰]. در نسخه RMCL این الگوریتم، عملگر بسط به صورت M^*M_G تغییریافته است که M_G ماتریس تلاقی اولیه گراف می‌باشد. هدف از این کار، تولید خوشه‌هایی دقیق‌تر است [۶،۱۰،۱۱،۱۲،۱۳].

۳. الگوریتم پیشنهادی برای فاز پیش‌پردازش

هدف از این فاز، تولید ماتریس ورودی الگوریتم است. در ابتدا کلیه مقادیر مربوط به هر داده با استفاده از نرمال‌سازی min-Max به بازه [۰و۱] انتقال می‌یابند(ماتریس M). سپس با استفاده از فرمول شباهت کسینوسی، شباهت بین هر دو نمونه داده محاسبه و نتایج در ماتریسی مربعی ذخیره می‌گردد(ماتریس شباهت یا M_S). در ادامه، برای تاکید بیشتر بر نمونه داده‌های شبیه به هم، از حدآستانه‌ای استفاده شده و

*ارائه دهنده

درایه‌های کمتر از این حد، در ماتریس شباهت صفر می‌شوند. در پایان، کلیه مقادیر غیرصفر ماتریس به ۱ تغییر کرده و نهایتا ماتریس شباهت بصورت ستونی نرمال‌سازی می‌گردد.



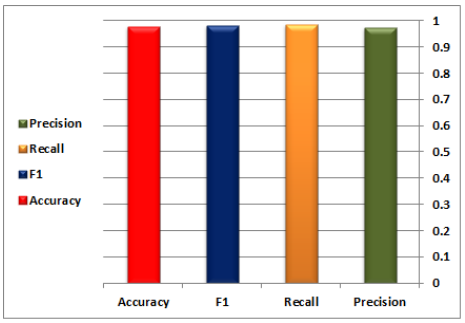
شکل ۲- فلوچارت فاز پیش‌پردازش الگوریتم پیشنهادی

۴. روش پیشنهادی جهت بهینه‌سازی الگوریتم MCL

در الگوریتم پیشنهادی، بعد از اعمال عملگر بسط مقادیر هر ستون از ماتریس هرس می‌گردد. برای تعیین حدآستانه این هرس‌سازی، از تفاضل میانگین و انحراف معیار مقادیر هر ستون استفاده شده‌است. پس جریان بین نودهایی که شباهت کمتری به هم دارند، از ماتریس حذف می‌شوند.

۵. نتایج

در ارزیابی نتایج از فایل ثبت وقایعی با ۲۴۲۲ نشست (۱۰۵۵ انسان و ۱۳۶۷ روبات وب) استفاده شده‌است[۴،۷،۸،۹]. برای هر نشست از ۱۳ خصیصه مرسوم در شناسایی روبات‌های وب استفاده شده‌است[۱،۲،۳].



شکل ۳– نتایج حاصل از ارزیابی الگوریتم پیشنهادی

باتوجه به شکل ۳، دقت خوشه‌بندی روش پیشنهادی ۹۸.۱۱٪ است.

آزمون همبستگی	تعداد تکرار حلقه اصلی	زمان اجرا	تعداد خوشه‌های نهایی	الگوریتم
۰.۶۲۰۶	۳	۱۵.۹۵sec	۲	روش پیشنهادی
۰.۶۲۷۹	۶	۱۶.۹۶sec	۴	RMCL

جدول ۱- مقایسه نتایج الگوریتم پیشنهادی و الگوریتم RMCL

جدول(۱) مقایسه نتایج روش پیشنهادی و الگوریتم RMCL است. آزمون همبستگی هر دو الگوریتم تقریبا برابر است. سرعت RMCL تنها چند ثانیه(۱.۱ sec) کمتر از روش پیشنهادی است، و تعداد تکرار حلقه اصلی در آن ۲ برابر است. اما آن‌چه که باعث برتری روش پیشنهادی نسبت به الگوریتم RMCL می‌گردد، تعداد خوشه‌های نهایی است که در RMCL چهار عدد بوده و نشان‌دهنده دقت پایین‌تر این روش نسبت به الگوریتم پیشنهادی می‌باشد.

۶. مراجع

[۱] Doran, D., Gokhale, Swapna S., "*Web robot detection techniques: overview and limitations*" Data Mining and Knowledge Discovery, Vol.۲۲, pp.۱۸۳-۲۱۰, ۲۰۱۱.

[۲] Stassopoulou, A., Dikaiakos, Marios D., "*Web robot detection: A probabilistic reasoning approach*", Computer Networks: The International Journal of Computer and Telecommunications Networking,Vol.۵۳, pp.۲۶۵-۲۷۸, ۲۰۰۹.

[۳] Stevanovic, D., An, A., Vlajic, N., "*Feature evaluation for web crawler detection with data mining techniques*", Expert Systems with Applications: An International Journal, Vol.۳۹, pp.۸۷۰۷-۸۷۱۷, ۲۰۱۲.

[٤] Rajabnia.J, Zabihi.M, Vafaeijahan.M., "*Web Robot Detection with fuzzy inference system based on decision trees*." IDMC ۲۰۱۳, university of Tehran.

[٥] Kanji, G. (۲۰۰٦). "*۱۰۰ Statistical Tests: STAGE*".

[٦] Dongen.v, S.M."*Graph Clustering by Flow Simulation*". PhD thesis, University of Utrecht (۲۰۰۰).

[۷] Tan, P., Kumar, V., "*Discovery of Web Robot Sessions Based on their Navigational Patterns*", Data Mining and Knowledge Discovery, Vol.٦, pp.۹-۳۵, ۲۰۰۲.

[۸] Bomhardt, C., Gaul, W., Schmidt-Thieme, L., "*Web Robot detection pre-processing web log files for Robot Detection*", New Developments in Classification and Data Analysis Studies in Classification, Data Analysis, and Knowledge Organization, pp.۱۱۳-۱۲٤, ۲۰۰۵.

[۹] Tan P.-N., Kumar V. "*Modelling of Web robot navigational patterns*". In Workshop on Web Mining for E-Commerce. Challenges and Opportunities Working Notes (KDD۲۰۰۰), Boston, MA; August ۲۰۰۰. p. ۱۱۱–۱۷.

[۱۰] Sataluri.V, Parthasarathy.S, "Scalable Graph Clustering Using Stochastic Flows: Application to Community Discovery",KDD۲۰۰۹,june ۲۸-July ۱,Paris,France.

[۱۱] Enright. A, Dongen.S, Ouzounis.A,"*An efficient algorithm for large-scale detection of protein families*", Nucleic Acids Research,۲۰۰۲,Vol.۳۰,No.۷ ۱۵۷۵-۱۵۸٤.

[۱۲] Szilagyi.L, Medves.L, Szilagyi.M,"*A modified markov clustering approach to unsupervised classification of protein sequence*".Neurocomputing Journal ۷۳ (۲۰۱۰) ۲۳۳۲-۲۳٤٥.

[۱۳] Stephen.F, Thomas.L, "*Gapped BLAST and PSI-BLAST: a new generation of protein database search programs*", Nucleic Acids Research,۱۹۹۷,Vol. ۲٥,No.۱۷ ۳۳۸۹-۳٤۰۲.