University of Qom

Engineering Management and Soft Computing

JEMSC.QOM
Online ISSN: 2538-2675

Presenting a method to reduce the sensitivity of incremental clustering algorithms of XML documents based on collective intelligence algorithms

Mohammad Fakhri Nazari[1]✉, Ebrahim Fakhri Nazari[2] and Ali Nouroz Bakhsh[3]

1. Corresponding author, Ph.D. Student in Information Technology Management, Faculty of Management University of Science and Research, Tehran, Iran. Email: M_kasitman@yahoo.com
2 .Associate Prof., Faculty of Management University of Science and Research, Tehran, Iran. Email: E60_itmggtn@yahoo.com
3. MSc. in Information Technology, Faculty of Management, University of Science and Research, Tehran, Iran. Email: Ali.nb12605@gmail.com

Article Info

Article type:
Research Article

Article history:
Received 2023 October 24
Received in revised form 2023 November 21
Accepted 2024 March 10
Published online 2024 March 15

Keywords:
collective intelligence,
incremental clustering,
particle optimization algorithm,
semistructured documents.

ABSTRACT

Many internet data are semi-structured, including information about the document and its data. Using documents that include data and their structure simultaneously increases the documents' complexity and, as a result, makes processing these documents more difficult. The increase in the complexity of semi-structured documents, the extraction of useful information, and the development of this technology face many difficulties. Until now, various methods have been presented for storing and retrieving data from semi-structured documents, most placed in two groups with batch and incremental approaches. In the batch or cluster approach, it is assumed that all the documents can be accessed and clustered, and the documents can be processed several times, which increases the execution time of such algorithms. In the incremental approach, all the records do not exist in one place. However, over time, they are provided to the classification method, and from this point of view, the execution time of such algorithms is less compared to the batch method. As a result, their execution speed is faster. This research compared our proposed method with XCLS and XCLS+ methods in three evaluation criteria: Entropy, Purity, and Fscore. The results showed that the proposed method is preferable to the XCLS and XCLS+ methods in terms of Entropy, Purity, and Fscore, and it is slightly less efficient than the XCLS+ method only in the Fscore criterion.

Cite this article: Fakhri Nazari, M. Fakhri Nazari, E. & Bakhsh, N. (2023). Presenting a method to reduce the sensitivity of incremental clustering algorithms of XML documents based on collective intelligence algorithms. *Enginiiring Management and Soft Computing, 9 (2)*. 177-187. **DOI**: https://doi.org/

DOI: https://doi.org/

Publisher: University of Qom

JEMSC.QOM
شاپا الکترونیکی: ۲۶۷۵-۲۵۳۸

مدیریت مهندسی و رایانش نرم

دانشگاه قم

ارائه روشی برای کاهش حساسیت الگوریتم‌های خوشه‌بندی افزایشی اسناد XML مبتنی بر الگوریتم‌های هوش دسته‌جمعی

محمد نظری فرخی[1]✉، ابراهیم نظری فرخی[2] و علی نوروزبخش[3]

۱. نویسنده مسئول، گروه مدیریت فناوری اطلاعات، دانشکده مدیریت، دانشگاه علوم و تحقیقات، تهران، ایران. رایانامه: M_kasitman@yahoo.com

۲. گروه مدیریت فناوری اطلاعات، دانشکده مدیریت، دانشگاه علوم و تحقیقات، تهران، ایران. رایانامه: E60_itmggtn@yahoo.com

۳. گروه مدیریت فناوری اطلاعات، دانشکده مدیریت، دانشگاه علوم و تحقیقات، تهران، ایران. رایانامه: Ali.nb12605@gmail.com

اطلاعات مقاله

نوع مقاله: مقاله پژوهشی

تاریخ دریافت: ۱۴۰۲/۰۸/۰۲

تاریخ بازنگری: ۱۴۰۲/۰۸/۳۰

تاریخ پذیرش: ۱۴۰۲/۱۲/۲۰

تاریخ انتشار: ۱۴۰۲/۱۲/۲۵

کلیدواژه‌ها:
اسناد نیمه‌ساخت‌یافته، الگوریتم بهینه‌سازی ذرات، خوشه‌بندی افزایشی، هوش دسته‌جمعی.

چکیده

بسیاری از داده‌های اینترنتی، نیمه‌ساخت‌یافته می‌باشند که شامل اطلاعاتی درمورد سند و داده‌های بکار رفته در آن می‌باشند. بکارگیری اسنادی که همزمان شامل داده‌ها و ساختار آنها باشد، باعث افزایش پیچیدگی اسناد و درنتیجه پردازش مشکل‌تر این اسناد می‌شود. افزایش پیچیدگی اسناد نیمه‌ساخت‌یافته، استخراج اطلاعات مفید و توسعه این فناوری را با دشواری‌های فراوانی مواجه می‌نماید. تاکنون روش‌های مختلفی برای ذخیره‌سازی و بازیابی اطلاعات اسناد نیمه‌ساخت‌یافته ارائه شده‌است که بیشتر آنها در دو گروه با رهیافت دسته‌ای و افزایشی قرار می‌گیرند. در رهیافت دسته‌ای یا خوشه‌ای فرض بر این است که کل اسناد قابل دسترسی و خوشه‌بندی است و اسناد می‌توانند چندین بار مورد پردازش قرار گیرند که باعث افزایش زمان اجرای اینگونه الگوریتم‌ها می‌شود. در رهیافت افزایشی کل اسناد تماما یکجا وجود ندارند بلکه به مرور زمان دراختیار روش دسته‌بندی قرار می‌گیرد که از این نظر زمان اجرای اینگونه الگوریتم‌ها نسبت به روش دسته‌ای کمتر و درنتیجه سرعت اجرای آنها بیشتر است. در این پژوهش روش پیشنهادی ما با روش‌های XCLS و XCLS+ در سه معیار ارزیابی Entropy، Purity و Fscore مورد مقایسه قرار گرفت. نتایج نشان داد روش پیشنهادی در معیارهای Entropy، Purity و Fscore نسبت به دو روش XCLS و XCLS+ ارجحیت دارد و فقط در معیار Fscore نسبت به روش XCLS+ اندکی کارآیی کمتری از خود نشان می‌دهد.

استناد: نظری فرخی، محمد؛ نظری فرخی، ابراهیم و نوروزبخش، علی. (۱۴۰۲). «ارائه روشی برای کاهش حساسیت الگوریتم‌های خوشه‌بندی افزایشی اسناد XML مبتنی بر الگوریتم‌های هوش دسته‌جمعی». *مدیریت مهندسی و رایانش نرم، دوره ۹ (۲)*. صص: ۱۸۷-۱۷۷. https://doi.org/

ناشر: دانشگاه قم

۱) مقدمه

امروزه دنیای وب به بستر عظیمی از داده‌ها و اطلاعات تبدیل شده‌است که اطلاعات آن به‌صورت افزایشی تولید، ذخیره و مبادله می‌شود (نایاک، ۲۰۰۸). از بین روش‌های موجود برای نمایش و انتقال اطلاعات، اسناد XML[۱۰۶] به دلیل انعطاف‌پذیری بالا مورد استقبال قرار می‌گیرند (نایاک و تران، ۲۰۰۷). از آنجاییکه در اسناد XML، برچسب‌ها مفهوم ساختاری و معنایی اطلاعات در اسناد متنی را توصیف می‌کنند، این اسناد به‌عنوان نیمه‌ساخت‌یافته[۱۰۷] شناخته می‌شوند (نایاک، ۲۰۰۸)، (هوانگ و همکاران، ۲۰۱۰)، (پیرنیک و همکاران، ۲۰۱۶). در سال ۱۹۷۰، در پاسخ نیاز به یک زبان قدرتمند برای مدلکردن اطلاعات وب، XML توسط کنسرسیوم جهانی گسترده[۱۰۸] پیشنهاد گردید. XML یک استاندارد غیررسمی[۱۰۹] پذیرفته‌شده در سطح گسترده و جهانی به‌منظور تبادل اطلاعات و انتقال بین سیستم‌ها می‌باشد (گوئل، ۲۰۰۸). خوشه‌بندی یکی از بهترین روش‌هایی است که برای کار با داده‌ها ارائه شده‌است. خوشه‌بندی قابلیت ورود به فضای داده و تشخیص ساختارش را امکان‌پذیر می‌نماید (کیم و همکاران، ۲۰۰۴). به‌عبارت‌دیگرخوشه‌بندی قرا دادن داده‌ها در گروه‌هایی است که اعضای هر گروه از زاویه خاصی شبیه یکدیگرند (کاستا و همکاران، ۲۰۱۳). فرآیند خوشه‌بندی اسناد XML، نقش حیاتی در بسیاری از حوزه‌ها مانند بازیابی اطلاعات[۱۱۰]، بهبود پردازش پرس‌وجو[۱۱۱]، یکپارچه‌سازی داده‌ها[۱۱۲]، وب‌سرویس‌ها[۱۱۳] مانند خوشه‌بندی سرویس‌های وب به‌منظور بهبود فرآیند کشف سرویس، وب‌کاوی[۱۱۴] مانند خوشه بندی نتایج جستجو در موتورهای جستجو[۱۱۵] و بیوانفورماتیک[۱۱۶] دارد (الجرگوی و همکاران، ۲۰۱۱).

تاکنون روش‌های مختلفی برای خوشه‌بندی اسناد XML ارائه شده‌است که بیشتر آنها بر مبنای دو روش دو به دو[۱۱۷](کل اسناد) و روش افزایشی[۱۱۸](ترتیب اسناد) عمل می‌نمایند (آلجرگیوی و همکاران، ۲۰۱۱). در روش دو به دو، برنامه کاربردی در مرحله اول به کل اسناد تماما یکجا دسترسی دارد و جهت کشف دانش موردنظر به تمام اسناد و محتوی آنها دسترسی داشته و این احتمال وجود دارد که محتوا و ساختار اسناد بارها توسط تکنیک خوشه‌بندی مورد پردازش قرار گیرد. مزیت عمده روش دو به دو دقت مناسب و در عین حال زمان بالای خوشه‌بندی و دسته‌بندی اطلاعات است. در مقابل روش دو به دو، روش افزایشی وجود دارد که در مرحله اول به تمام اسناد دسترسی ندارد بلکه ورودی الگوریتم ترتیبی از اسناد است که از این نظر سرعت اجرای این روش بالا ولی در عین حال دقت آن تحت تاثیر ترتیب ورودی‌ها قرار دارد (علیشاهی و همکاران، ۲۰۱۰).

خوشه‌بندی اسناد XML می‌تواند از نظر دقت و زمان اجرای خوشه‌بندی امری چالش‌برانگیز باشد که روش‌های

---

106. Extensible Markup Language
107.Semi-Structured
108. world Wide Web Conserciom
109. De facto
110. Information Retrieval
111. Query Processing
112. Data Integrity
113. Web Service
114.Web Mining
115. Search Engine
116 Bioanformatic
117. Pair wise
118. Incremental

افزایشی سعی نموده‌اند تاحدودی بر این مشکلات به‌ویژه زمان اجرا غلبه نمایند. خوشه‌بندی اسناد را می‌توان یک مسئله سخت و دشوار[۱۱۹] از نوع مسائل بهینه‌سازی[۱۲۰] درنظر گرفت که هدف آن ارائه یک خوشه‌بندی دقیق با شباهت بیشینه[۱۲۱] درون خوشه‌ها و کمینه[۱۲۲] نمودن شباهت اسناد متعلق به خوشه‌های متفاوت می‌باشد. الگوریتم هوش دسته‌جمعی ذرات[۱۲۳] را می‌توان یکی از نمونه‌های تکاملی برای یافتن راه‌حل‌های بهینه[۱۲۴] یک مسئله نظیر خوشه‌بندی درنظر گفت. با توجه به اینکه خوشه‌بندی اسناد XML یک مسئله بهینه‌سازی[۱۲۵] محسوب می‌شود، تلاش شده‌است در این پژوهش به کمک الگوریتم هوش دسته‌جمعی ذرات[۱۲۶] خوشه‌بندی افزایشی دقیق‌تری از این اسناد ارائه شود.

۲) اهداف پژوهش

- بالا بردن دقت خوشه‌بندی اسناد XML.
- کاستن از تاثیر ورود اسناد در دقت خوشه‌بندی به کمک الگوریتم هوش دسته‌جمعی ذرات.

۳) فرضیات پژوهش

- ترکیب الگوریتم هوش دسته‌جمعی ذرات و خوشه‌بندی افزایشی، دقت خوشه‌بندی را افزایش می‌دهد.
- افزایش تعداد خوشه‌ها، دقت را افزایش و در مقابل سرعت روش موردنظر را کاهش می‌دهد.
- استفاده از خوشه‌بندی اسناد به کمک الگوریتم ذرات، تاثیر ورود اسناد در خوشه‌بندی را کاهش می‌دهد.

۴) سئوالات پژوهش

- چگونه ترکیب الگوریتم هوش دسته‌جمعی ذرات و خوشه‌بندی افزایشی، دقت خوشه‌بندی اسناد XML را افزایش می‌دهد؟
- افزایش تعداد خوشه‌ها چه تاثیری بر دقت و سرعت روش پیشنهادی دارد؟
- آیا میزان تاثیر ورود اسناد در دقت روش پیشنهادی به کمک الگوریتم ذرات کاهش می‌یابد؟

۵) روش پیشنهادی

در این قسمت، روش پیشنهادی به همراه ملزمات آن برای توسعه یک الگوریتم افزایشی شرح داده می‌شود و با توجه به اینکه الگوریتم ذرات نقش مهمی در روش پیشنهادی دارد در ابتدا الگوریتم موردنظر تشریح شده و در ادامه نحوه استفاده از این الگوریتم برای خوشه‌بندی افزایشی شرح داده می‌شود.

---

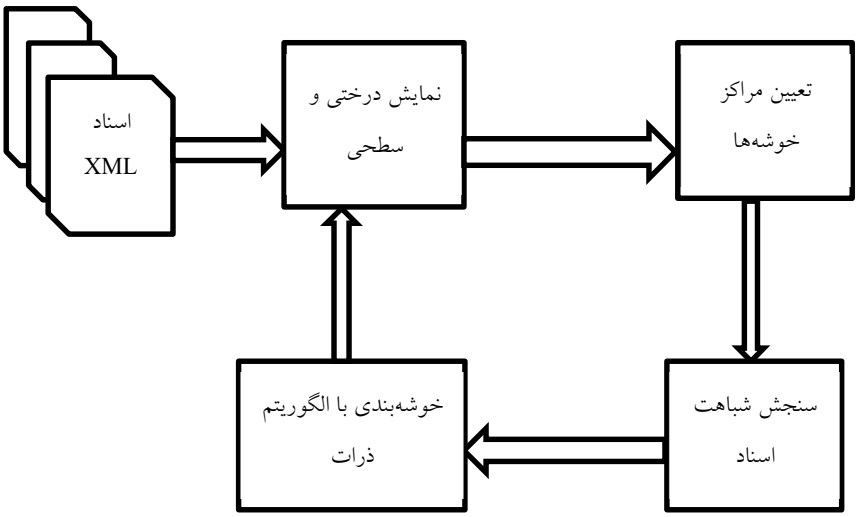
119. NP Hard
120. Optimization problems
121. Maximum
122. Minimum
123. Particle Swarm Optimization
124. Optimal Solutions
125. Optimization problem
126. Particle Swarm Optimization Algorithm

۵-۱) الگوریتم ذرات

الگوریتم بهینه‌سازی هوش دسته‌جمعی ذرات [۱۲۷] در سال ۱۹۹۵ توسط دو روانشناس کندی[۱۲۸] و ابرهارت [۱۲۹] ارائه شد. الگوریتم ذرات براساس رفتار گروهی دسته‌های ماهی و پرندگان در یافتن غذا و فرار از شکارچیان مدل‌سازی شده‌است (گاد، ۲۰۲۲)، (ابرهارت و شی، ۲۰۰۱). رفتارهای دسته‌جمعی فقط مختص پرندگان یا ماهیان نیست بلکه در جانداران دیگر نظیر مورچه‌ها [۱۳۰] (دی کاپریو و همکاران، ۲۰۲۲)، (سانتوس و همکاران، ۲۰۱۰)، کرم‌های شب‌تاب [۱۳۱] (یسودا و آمودا، ۲۰۲۲)، (فیستر و همکاران، ۲۰۱۳)، خفاش [۱۳۲] (زان و همکاران، ۲۰۲۲)، (میشرا و همکاران، ۲۰۱۲) و عنکبوت‌ها [۱۳۳] (جیمز و همکاران، ۲۰۱۵) نیز دیده می‌شود. الگوریتم ذرات برخلاف الگوریتم‌های دیگر از روابط ساده‌ای جهت یافتن بهینه‌های سراسری استفاده می‌نماید و در عین حال دارای نرخ همگرایی بسیار بالاتری نسبت به الگوریتم‌های نظیر الگوریتم ژنتیک [۱۳۴] (زانگ و همکاران، ۲۰۱۰) می‌باشد.

۵-۲) چارچوب روش پیشنهادی

روش پیشنهادی برای خوشه‌بندی اسناد XML یک الگوریتم افزایشی است که برای بهبود خوشه‌بندی از الگوریتم ذرات استفاده نموده است. در شکل ۱ چارچوب این روش پیشنهادی نشان داده شده‌است:



شکل ۱. چارچوب کلی روش پیشنهادی

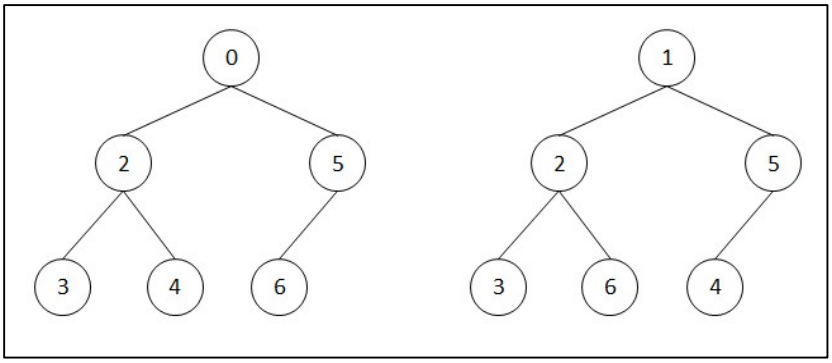
در این چارچوب هر سند در ابتدا به شکل ساختار درختی نمایش و برچسب‌های آن شناسایی می‌شود تا در مرحله سنجش شباهت اسناد و خوشه‌ها بکار گرفته شود. در مرحله ابتدایی هر سند ورودی می‌تواند به‌عنوان یک مرکز خوشه یا نمانده خوشه‌ها درنظر گرفته‌شود. در این مرحله یک آستانه اولیه برای تعیین مراکز خوشه‌ها درنظر گرفته می‌شود و تعدادی

127. Particle swarm optimization (PSO)
128. Kennedy
129. Eberhart
130. Ant clone algorithm
131. Firefly algorithm
132. Bat algorithm
133. Spider algorithm
134. Genetic algorithm

از اسناد ورودی به‌عنوان مراکز خوشه‌ها تعیین می‌شوند. در مراحل اولیه می‌توان هر سند ورودی را به‌عنوان یک مرکز خوشه درنظر گرفت که این عمل علی‌رغم سادگی کار، ریسک بالایی در دقت خوشه‌بندی فراهم می‌نماید لذا در مرحله اول الگوریتم به کمک آستانه تعیین مراکز خوشه‌ها سعی می‌شود این مشکل تا حد بالایی تعدیل شود. بعد از تعیین مراکز اولیه خوشه‌ها تعداد دیگری سند به‌عنوان ورودی انتخاب شده و برحسب معیار شباهت سطحی میزان شباهت آنها با مراکز خوشه‌ها تعیین می‌شود و در ادامه به تعداد جمعیت اولیه تعیین‌شده در الگوریتم ذرات، خوشه‌بندی ایجاد نموده و معیار شایستگی را مجموع شباهت درون‌خوشه‌ای درنظر می‌گیریم تا بهترین ذره که شامل بهینه‌ترین خوشه‌بندی است، انتخاب شود.

۵-۳) معیار شباهت

در روش پیشنهادی یک معیار شباهت پیشنهادی برای مقایسه شباهت دو سند، ارائه نموده‌ایم که براساس شباهت سطوح مختلف درخت ساختار این اسناد عمل می‌نماید. شکل ۲ نمایش دو سند مختلف XML را به شکل درختی نمایش داده‌است. در اینجا میزان شباهت این دو سند به کمک الگوریتم XCLS و روش پیشنهادی شرح داده‌می‌شود.



شکل ۲. مختلف XML را به شکل درختی

جهت محاسبه شباهت در الگوریتم XCLS از رابطه ۱ استفاده می‌شود:

$$Sim = \frac{0/5 \times \sum_{i=0}^{L-1} CN_1^i \times r^{L-i-1} + 0/5 \times \sum_{j=0}^{L-1} CN_1^j \times r^{L-j-1}}{(\sum_{k=0}^{L-1} N^k \times r^{L-k-1}) \times Z} \qquad (1)$$

که محاسبه این معیار شباهت در الگوریتم XCLS بصورت زیر است:

$$Sim = \frac{0/5 \times (0 \times 2^2 + 0 \times 2^1 + 0 \times 2^0) + 0/5 \times (0 \times 2^2 + 0 \times 2^1 + 0 \times 2^0)}{1 \times 2^2 + 2 \times 2^1 + 3 \times 2^0} = 0 \qquad (2)$$

حال در روش پیشنهادی با اعمال تغییراتی در رابطه ۱ می‌توانیم رابطه ۳ را معرفی نمایم که در آن میزان شباهت عناصر سطوح متناظر نیز به رابطه موردنظر اضافه شده‌است:

$$Sim = \frac{\frac{1}{3} \times \sum_{i=0}^{L-1} CN_1^i \times r^{L-i-1} + \frac{1}{3} \times \sum_{j=0}^{L-1} CN_1^j \times r^{L-j-1} + \frac{1}{3} \times \sum_{h=0}^{L-1} CN_{1,2}^h \times r^{L-h-1}}{(\sum_{k=0}^{L-1} N^k \times r^{L-k-1}) \times Z} \qquad (3)$$

حال برای محاسبه میزان شباهت دو سند شکل ۱، مطابق ذیل از شباهت در روش پیشنهادی استفاده می‌شود:

$$Sim = \frac{\frac{1}{3}\times(0\times 2^2+0\times 2^1+0\times 2^0)+\frac{1}{3}\times(0\times 2^2+0\times 2^1+0\times 2^0)+\frac{1}{3}(0\times 2^2+2\times 2^1+3\times 2^2)}{1\times 2^2+2\times 2^1+3\times 2^0} = 0/51 \quad (4)$$

مشاهده می‌شود که معیار شباهت بکار رفته در روش پیشنهادی دو سند را به اندازه 0/51 درنظر گرفته است (این عدد بین ۰ و ۱ متغیر است) و این در حالی است که معیار شباهت الگوریتم XCLS، این دو سند را بدون شباهت یا با شباهتی به اندازه صفر درنظر گرفته‌است. نتایج بررسی‌های دیگر نشان می‌دهد که این معیار پیشنهادی برای سنجش شباهت سندها از معیاز شباهت الگوریتم XCLS بهتر عمل می‌نماید.

۶) سخت افزار و نرم افزار شبیه‌سازی، مجموعه داده و معیارهای ارزیابی

جهت پیاده‌سازی روش پیشنهادی، از نرم‌افزار متلب به‌عنوان محیط برنامه‌نویسی استفاده شده و خروجی موردنظر روی سیستمی با مشخصات کارت گرافیک ۴ گیگ، حافظه 8 گیگابایت و پردازنده 7 هسته‌ای تهیه شده‌است. مجموعه داده بکار رفته در این پژوهش تعدادی از اسناد XML می‌باشند که قصد داریم آنها را توسط الگوریتم پیشنهادی خوشه‌بندی نمایم و توسط آنها روش پیشنهادی را با تعدادی از الگوریتم‌های خوشه‌بندی اسناد مقایسه نماییم. در ادامه معیارهای ارزیابی که برای تجزیه‌وتحلیل روش پیشنهادی بکار گرفته شده، آورده شده‌است.

۶–۱) معیار Entropy

اگر q را تعداد خوشه‌ها و N تعداد کل اسناد مجموعه داده درنظر گرفته شود، می‌توان آنتروپی کلی خوشه‌بندی را طبق رابطه ۵ تعریف نمود:

$$E(C_i) = \frac{1}{\log(k)} + \sum_{r=1}^{k} \frac{n_i^r}{n_i} log \frac{n_i^r}{n_i} \quad (5)$$

۶–۲) معیار Purity

می‌توان برای کل خوشه‌ها از معیار Purity کل مطابق رابطه ۶ که یک مجموع وزنی است استفاده نمود:

$$Purity = \sum_{i=1}^{q} \frac{n_i}{N} P(C_i) \quad (6)$$

۶–۳) معیار Fscore

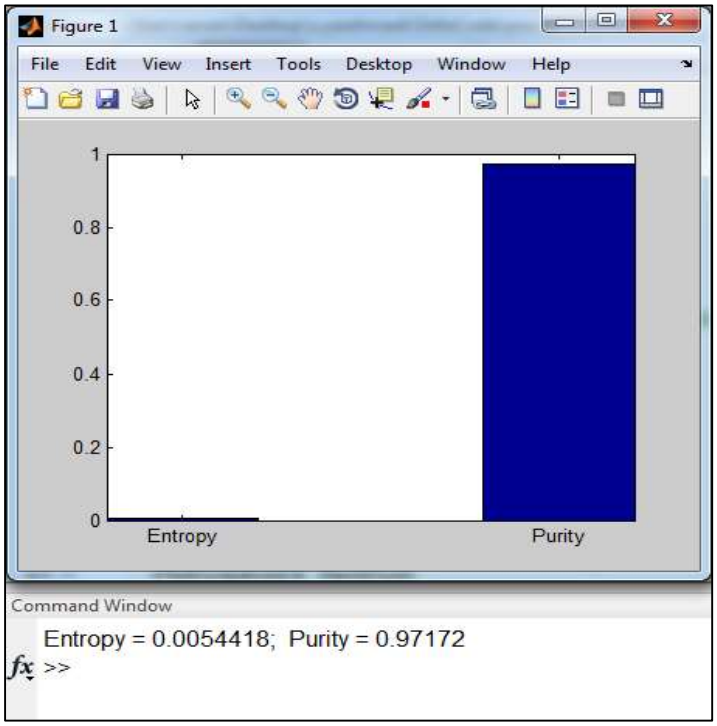
معیار ارزیابی FScore یک معیار مناسب‌تر برای ارزیابی خوشه‌بندی نسبت به معیارهای ارزیابی Entropy و Purity است. FScore ترکیبی از دو معیار دقت ۱۳۵ و حساسیت ۱۳۶ است.

135. Precision
136. Recall

$$F(Z_{r1}, C_i) = \frac{2n_i^r}{n_i + n_r} \qquad (7)$$

۶-۴) نمونه خروجی در الگوریتم پیشنهادی

در شکل ۳ یک نمونه از خروجی الگوریتم پیشنهادی و محاسبه دو معیار Entropy و Purity را نشان داده‌است. همانگونه که خروجی برنامه نشان می‌دهد، میزان Entropy و Purity در خروجی موردنظر به ترتیب برابر 0/0054 و 0/9717 می‌باشد که نشان‌دهنده کیفیت مناسب خروجی خوشه‌بندی این اسناد در روش پیشنهادی است. کمینه‌بودن مقدار Entropy در این خروجی نشان می‌دهد که آشفتگی و عدم شباهت درون خوشه‌ها در مجموع اندک است و این موضوع نشان‌دهنده این است که اسناد به‌خوبی درون خوشه‌های خود قرار گرفته‌اند. از طرفی مقدار نزدیک به یک معیار Purity نیز یک‌دست‌بودن اسناد هر یک از خوشه‌ها را نشان می‌دهد. به عبارت بهتر خروجی شکل ۳ نشان‌دهنده خوشه‌بندی مناسب اسناد توسط الگوریتم پیشنهادی است. در بخش‌های بعدی یک مقایسه بین روش پیشنهادی و دو روش خوشه‌بندی افزایشی XCLS و XCLS+ خواهیم داشت.

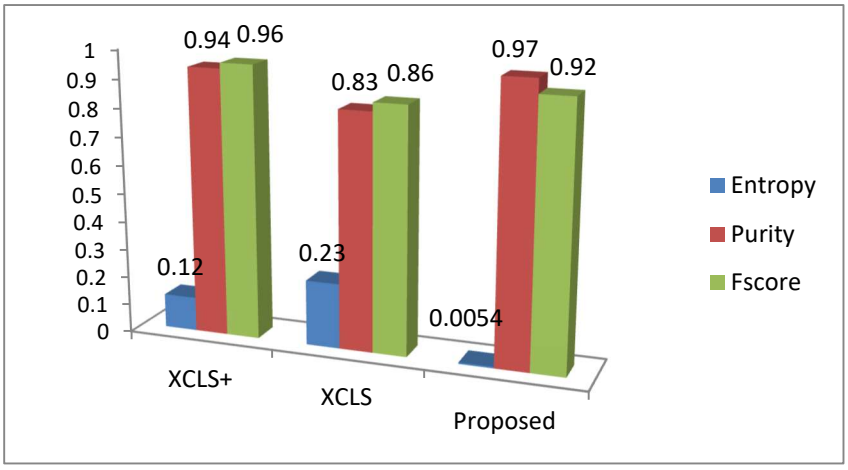


شکل ۳. محاسبه معیار Entropy و Purity در روش پیشنهادی

۶-۵) مقایسه با الگوریتم‌های XCLS و XCLS+

در این قسمت یک مقایسه بین روش پیشنهادی و الگوریتم‌های XCLS و XCLS+ برای خوشه‌بندی اسناد XML برحسب سه معیار اصلی Entropy، Purity و Fscore خواهیم داشت و در ادامه یک مقایسه زمانی نیز بین این سه الگوریتم صورت خواهد گرفت. جهت ارزیابی و مقایسه فرض‌شده در سه روش پیشنهادی، XCLS و XCLS+ مقدار حد آستانه شباهت

برابر ۰/۹ و مقدار فاکتور وزن‌دهی r=۲ می‌باشد. در روش پیشنهادی اندازه جمعیت اولیه ذرات و تعداد تکرار به‌ترتیب برابر ۵۰ و ۲۰۰ و ضرایب یادگیری هر دو برابر ۲ انتخاب شده‌است. نتایج مقایسه روش پیشنهادی، XCLS و XCLS+ در شکل ۴ نشان داده شده‌است:

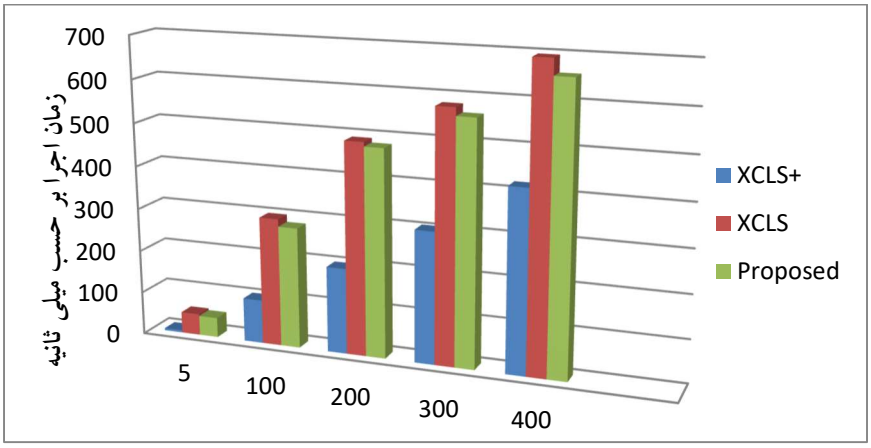


شکل ٤. محاسبه معیار Entropy و Purity در روش پیشنهادی

بررسی نمودار و آزمایشات مشابه نشان می‌دهد که Entropy روش پیشنهادی به مراتب نسبت به دو روش XCLS و XCLS+ کمینه‌تر شده و از این نظر، روش پیشنهادی در معیار آنتروپی از دو روش XCLS و XCLS+ بهتر عمل می‌نماید. در معیار Purity، روش پیشنهادی بهتر از روش‌های XCLS و XCLS+ عمل می‌نماید بطوریکه میزان Purity در آنها به ترتیب برابر 0/97، 0/83 و 0/94 می‌باشد. مقایسه سه روش خوشه‌بندی در معیار Fscore نشان می‌دهد که روش XCLS+ نسبت به روش پیشنهادی و روش XCLS برتری دارد به‌گونه‌ای که مقدار معیار Fscore در روش XCLS+ پیشنهادی و روش XCLS به ترتیب برابر 0/96، 0/92 و 0/86 می‌باشد. نتایج معیارهای Entropy، Purity و Fscore بر روی سه روش بکار رفته در این پژوهش نشان می‌دهد که این نتایج در آستانه‌های مختلف نظیر 0/8، 0/7، 0/6 و 0/1 نیز قابل استناد است.

۶-۶) مقایسه زمانی

پیچیدگی زمانی الگوریتم XCLS و XCLS + برابر با $O(m \times c \times p \times n)$ است که m تعداد عناصر در اسناد، c تعداد خوشه‌ها، p تعداد تکرار و n تعداد عناصر مشخص در خوشه‌هاست. سندهایی که در یک خوشه جای می‌گیرند بایستی از لحاظ ساختاری و عناصر شبیه‌به‌هم باشند. بنابراین تعداد عناصر مشخص در خوشه‌ها همیشه از عناصر مشخص در اسناد کمتر هستند. تعداد تکرار نیز معمولا عدد کوچکی است و در نهایت از عدد شش بیشتر نمی‌شود بنابراین اگر تعداد خوشه‌ها از تعداد اسناد کمتر باشد هزینه زمانی این دو روش خطی و متناسب با تعداد اسناد است. در شکل ۵ یک مقایسه زمانی بین روش پیشنهادی و روش‌های XCLS و XCLS+ برحسب اندازه و تعداد سندهای ورودی را نشان داده‌است.



شکل ۵. مقایسه زمانی بین روش پیشنهادی و روش‌های XCLS و +XCLS

تحلیل نمودار موردنظر نشان می‌دهد که زمان اجرای روش پیشنهادی اندکی از روش XCLS بهتر است و نسبت به روش +XCLS زمان اجرای بیشتری دارد. این نتایج نشان می‌دهد که روش پیشنهادی در زمان اجرا فقط، نسبت به روش XCLS بهتر عمل کرده‌است و دلیل آن زمان‌بر بودن خوشه‌بندی به روش تکاملی است. روش +XCLS برای پیداکردن میزان شباهت بین دو شی تنها یکبار ساختار سطحی مربوط به هر شی را پیمایش می‌کند درحالیکه روش پیشنهادی و XCLS برای محاسبه میزان شباهت، دوبار به بررسی ساختارهای سطحی دو شی می‌پردازد و در پایان بزرگترین را انتخاب می‌کرد. این امر باعث می‌شود که روش پیشنهادی و XCLS با تأخیر بیشتری انجام شود اما در عین حال باعث بهبود Entropy و Purity روش پیشنهادی نست به روش+XCLS می‌شود.

۷) نتیجه‌گیری

ترتیب ورود اسناد در اکثر روش‌های افزایشی بر روی کیفیت خوشه‌بندی این اسناد تاثیر منفی می‌گذارد. در این پژوهش یک روش پیشنهادی جهت بهبود خوشه‌بندی افزایشی به کمک الگوریتم هوش دسته‌جمعی ذرات[۱۳۷] ارائه شد تا در زمان منطقی یک خوشه‌بندی دقیق‌تر از اسناد ورودی به عمل آید. نتایج آزمایشات و شبیه‌سازی‌ها نشان می‌دهد که میزان Entropy و Purity در روش پیشنهادی به ترتیب برابر 0/0054 و 0/9717 می‌باشد که نشان‌دهنده کیفیت مناسب خروجی خوشه‌بندی اسناد XML است. نتایج تحقیق نشان می‌دهد که مقدار Entropy روش پیشنهادی به مراتب نسبت به دو روش XCLS و +XCLS کمینه‌تر شده و از این نظر روش پیشنهادی در معیار آنتروپی از دو روش XCLS و +XCLS بهتر عمل می‌نماید. در معیار Purity، روش پیشنهادی بهتر از روش‌های XCLS و +XCLS عمل می‌نماید بطوریکه میزان Purity در آنها به ترتیب برابر 0/97، 0/83 و 0/94 می‌باشد. مقایسه سه روش خوشه‌بندی در معیار Fscore نشان می‌دهد که روش +XCLS نسبت به روش پیشنهادی و روش XCLS برتری دارد به گونه‌ای که مقدار معیار Fscore در روش+XCLS پیشنهادی و روش XCLS به ترتیب برابر 0/96، 0/92 و 0/86 می‌باشد. نتایج معیارهای Entropy، Purity و Fscore بر روی سه روش بکار رفته در این پژوهش نشان می‌دهد که این نتایج در آستانه‌های مختلف نظیر 0/8، 0/7، 0/6 و 0/1 نیز قابل

[137]. Particle Swarm Optimization

استناد است. تحلیل زمانی روش پیشنهادی، XCLS و+XCLS نشان می‌دهد که زمان اجرای روش پیشنهادی اندکی از روش XCLS بهتر است و نسبت به روش +XCLS زمان اجرای بیشتری دارد.

منابع

Algergawy, A., Mesiti, M., Nayak, R., & Saake, G. (2011). XML data clustering: An overview. ACM Computing Surveys (CSUR), 43(4), 1-41. https://doi.org/10.1609/icwsm.v7i1.14380

Alishahi, M., Naghibzadeh, M., & Aski, B. S. (2010). Tag name structure-based clustering of XML documents. International Journal of Computer and Electrical Engineering, 2(1), 119. https://doi.org/10.1609/icwsm.v7i1.21369

Costa, Gianni, Giuseppe Manco, Riccardo Ortale, and Ettore Ritacco. "Hierarchical clustering of XML documents focused on structural components." Data & Knowledge Engineering 84 (2013): 26-46. https://doi.org/10.1609/icwsm.v7i1.95647

Di Caprio, D., Ebrahimnejad, A., Alrezaamiri, H., & Santos-Arteaga, F. J. (2022). A novel ant colony algorithm for solving shortest path problems with -fuzzy arc weights. Alexandria Engineering Journal, 61(5), 3403-3415. https://doi.org/10.1609/icwsm.v7i1.6257

Eberhart, R. C., & Shi, Y. (2001). Particle swarm optimization: developments, applications and resources. In evolutionary computation, 2001. Proceedings of the 2001 Congress on (Vol. 1, pp. 81-86). IEEE. https://doi.org/10.1609/icwsm.v7i1.62957

Fister, I., Yang, X. S., & Brest, J. (2013). A comprehensive review of firefly algorithms. Swarm and Evolutionary Computation, 13, 34-46. https://doi.org/10.1609/icwsm.v7i1.62148

Gad, A. G. (2022). Particle Swarm Optimization Algorithm and Its Applications: A Systematic Review. Archives of Computational Methods in Engineering, 1-31. https://doi.org/10.1609/icwsm.v7i1.75924

Gürel, G. (2008). Mining XML documents with association rule algorithms (Doctoral dissertation, Izmir Institute of Technology (Turkey)). https://doi.org/10.1609/icwsm.v7i1.62597

Hwang, J. H., & Ryu, K. H. (2010). A weighted common structure based clustering technique for XML documents. Journal of Systems and Software, 83(7), 1267-1274. https://doi.org/10.1609/icwsm.v7i1.75391

James, J. Q., & Li, V. O. (2015). A social spider algorithm for global optimization. Applied Soft Computing, 30, 614-627. https://doi.org/10.1609/icwsm.v7i1.95173

Kim, J., & Kim, H. J. (2004). A partition index for XML and semi-structured data. Data & Knowledge Engineering, 51(3), 349-368. https://doi.org/10.1609/icwsm.v7i1.96358

Mishra, S., Shaw, K., & Mishra, D. (2012). A new meta-heuristic bat inspired classification approach for microarray data. Procedia Technology, 4, 802-806.

Nayak, R. (2008). Fast and effective clustering of XML data using structural information. Knowledge and Information Systems, 14(2), 197-215. https://doi.org/10.1609/icwsm.v7i1.712596

Nayak, R. (2008). XML data mining: Process and applications. Idea Group Inc./IGI Global, 22. https://doi.org/10.1609/icwsm.v7i1.7193685

Nayak, R., & Tran, T. (2007). A progressive clustering algorithm to group the XML data by structural and semantic similarity. International Journal of Pattern Recognition and Artificial Intelligence, 21(04), 723-743. https://doi.org/10.1609/icwsm.v7i1.63215

Piernik, M., Brzezinski, D., & Morzy, T. (2016). Clustering XML documents by patterns. Knowledge and Information Systems, 46(1), 185-212. https://doi.org/10.1609/icwsm.v7i1.75931

Santos, L., Coutinho-Rodrigues, J., & Current, J. R. (2010). An improved ant colony optimization based algorithm for the capacitated arc routing problem. Transportation Research Part B: Methodological, 44(2), 246-266. https://doi.org/10.1609/icwsm.v7i1.71937

Yesodha, R., & Amudha, T. (2022). A bio-inspired approach: Firefly algorithm for Multi-Depot Vehicle Routing Problem with Time Windows. Computer Communications, 190, 48-56. https://doi.org/10.1609/icwsm.v7i1.93817

Zan, Z., Cong, Y., & Zhang, X. (2022, May). An Improved Bat Algorithm for Solving Nonlinear Algebraic Systems of Equations. In Proceedings of the 7th International Conference on Big Data and Computing (pp. 75-81). https://doi.org/10.1609/icwsm.v7i1.71987

Zang, H., Zhang, S., & Hapeshi, K. (2010). A review of nature-inspired algorithms. Journal of Bionic Engineering, 7, S232-S237. https://doi.org/10.1609/icwsm.v7i1.63158