سومین کنفرانس فناوری اطلاعات و دانش
۶ تا ۸ آذرماه ۱۳۸۶
دانشگاه فردوسی مشهد
IKT 2007 کنفرانس بین المللی فناوری اطلاعات و دانش

بهبود ساختارهای درختی برای یافتن خوشه های چگال از مستندات وب

محمد رحیمی* ، حسن ابوالحسنی† و مرتضی حقیر چهرقانی†

*دانشکده مهندسی کامپیوتر دانشگاه صنعتی امیرکبیر، †دانشکده مهندسی کامپیوتر دانشگاه صنعتی شریف

m_rahimi@cic.aut.ac.ir , abolhassani@sharif.edu , haghir@ce.sharif.edu

چکیده - یکی از مسائل بسیار با اهمیت مطرح در خوشه بندی داده ها ، محاسبه ی میزان فاصله ی میان اشیاء (عدم شباهت) است که می تواند دارای هزینه های پردازشی و ورودی/خروجی بسیار زیادی باشد. در این مقاله روشی برای کاهش این هزینه ها در خوشه بندی مبتنی بر چگالی پیشنهاد شده که بر پایه‌ی ذخیره داده ها در ساختار درختی خاصی استوار است و تا کنون در مورد مستندات وب اعمال نشده است. همچنین با انجام عملیات پیش پردازشی بر روی درخت مستندات ، سرعت الگوریتم در حذف نویزها و عملیات خوشه بندی، بهبود داده شده است. در نهایت مقایسه ای میان این روش با حالت معمول خوشه بندی مبتنی برچگالی(بدون استفاده از ساختار درختی)، انجام گرفته است که کارایی روش ارائه شده را نشان می دهد.

کلید واژه- خوشه بندی مبتنی برچگالی ، شاخص گذاری درختی، مستندات وب

۱- مقدمه

از جمله کاربردهای مهم خوشه بندی ، بعنوان پیش‌پردازش، استخراج دانش است که بویژه در محیط وب اهمیت زیادی دارد. یکی از مسائل اصلی در خوشه بندی این نوع داده ها ، ابعاد بسیار زیاد آنهاست. انجام عملیات خوشه بندی برای چنین داده‌هایی، مستلزم صرف هزینه های زمانی زیاد برای محاسبه‌ی فاصله میان اشیاء (یا به عبارت دیگر میزان عدم شباهت اشیاء نسبت به یکدیگر) می‌باشد. برخی از ایده هایی که برای حل این مسئله پیشنهاد شده است، استفاده از ساختار های درختی مانند VP-tree [1] ، Rtree [2] ، Mtree [3] و موارد مشابه دیگر برای شاخص گذاری (Indexing) اشیاء است. در این مقاله برای اولین بار از ساختار Mtree (به دلایلی که به آنها اشاره خواهد شد) ، در شاخص گذاری مستندات وب استفاده خواهد شد. همچنین با پیشنهاد یک عملیات پیش پردازشی برروی درخت، در سرعت حذف نویز ها و عملیات خوشه بندی بهبود داده شده است.

ایده ی اصلی مطرح در خوشه بندی مبتنی بر چگالی ، استفاده از مفهوم فیزیکی چگالی می باشد. درواقع در این روش، میزان تراکم اشیاء در یک محدوده ی فضایی خاص به عنوان معیاری برای تشخیص خوشه ها استفاده می شود. در روش خوشه بندی مبتنی برچگالی، یک خوشه با توجه به اشیاء همسایه ی آن (در یک شعاع خاص) شروع به رشد می کند و این کار را آنقدر ادامه می دهد که تعداد اشیاء موجود در آن همسایگی، کمتر از یک حد آستانه ای مشخص شوند. در این صورت ، رشد خوشه ی فعلی متوقف شده و خوشه های دیگری شکل خواهند گرفت.

استفاده از این روش خوشه بندی چندین مزیت را به دنبال خواهد داشت. در روش های خوشه بندی مبتنی بر تقسیم-بندی ، خوشه ها براساس فاصله ی میان اشیاء شکل می گیرند که منجر به تولید خوشه هایی با شکل صرفاً کروی می‌شود و در کشف خوشه های با اشکال دیگر دچار مشکلات جدی می شوند. در حالی که در روش های مبتنی بر چگالی، خوشه ها با هر نوع شکل دلخواه ، قابل کشف هستند. همچنین در روش های مبتنی برچگالی ، امکان