مجلـه‌ی علمی پژوهش در علوم رایانه ، شماره ۶ ، تابستان ۱۳۹۶، ص ۵۹- ۴۶

ISSN: 2745-3673
http://www.Rcsj.ir

مجلـه‌ی علمی پژوهش در علوم رایانه (سال دوم)
شماره ۶ / تابستان ۱۳۹۶ / ص ۵۹- ۴۶

خلاصه سازی خودکار متن با استفاده از کلونی زنبورعسل

فرشاد کیومرثی[۱]، شقایق بختیاری[۲]، مریم هادی پور[۳]

[۱] گروه کامپیوتر، دانشکده فنی و مهندسی ، دانشگاه آزاد اسلامی واحد شهرکرد، شهرکرد، ایران
[۲] گروه کامپیوتر، دانشکده فنی و مهندسی ، دانشگاه آزاد اسلامی واحد شهرکرد، شهرکرد، ایران
[۳] گروه کامپیوتر، دانشکده فنی و مهندسی ، دانشگاه آزاد اسلامی واحد شهرکرد، شهرکرد، ایران

نام نویسنده مسئول:
مریم هادی پور

چکیده

خلاصه‌سازی عبارتست از نمایش فشرده‌ی متن ورودی، به گونه‌ای که متن خروجی دارای مفاهیم متن ورودی باشد. به دلیل انفجار اطلاعات از طریق اینترنت و گسترش هرروزه‌ی آن از طریق وب سایت‌ها، نیاز به خلاصه‌سازی خودکار متن احساس می‌شود. خلاصه‌سازی استخراجی شامل سه مرحله‌ی پیش‌پردازش، پردازش و تولید خلاصه‌ی نهایی می‌شود. تمرکز این تحقیق بر قسمت تولید خلاصه نهایی، یعنی انتخاب جملات نهایی است و با توجه به این‌که الگوریتم کلونی زنبور عسل می‌تواند در یافتن راه‌حل‌های بهینه کارا باشد، در این تحقیق از آن بهره گرفته می‌شود. در انتها رویکرد پیشنهادی با دو روش خلاصه‌سازی با کلونی زنبور عسل و خلاصه‌سازی با الگوریتم مورچگان مقایسه می‌شود و بهتر شدن عملکرد روش پیشنهادی از جهت دقت اثبات و نتایج در قالب نمودار نمایش داده می‌شود.

واژگان کلیدی: خلاصه‌سازی خودکار متن، الگوریتم کلونی زنبور عسل، راه‌حل‌های بهینه، پیش‌پردازش، خلاصه‌ی نهایی

ISSN: 2745-3673
http://www.Rcsj.ir

مقدمه

به‌دلیل افزایش سریع اطلاعات از طریق وب‌سایت‌ها و سیستم‌های چندرسانه‌ای، خلاصه‌سازی متن به‌دست انسان کاری دشوار و نشدنی است [۱]. به‌جهت تسریع در کاهش افزونگی اطلاعات موجود در اینترنت، برنامه‌های کامپیوتری به کمک انسان آمده و با بکارگیری فرمول و الگوریتم‌هایی به خلاصه‌سازی متون می‌پردازد [۲]. خلاصه‌سازی خودکار متن را می‌توان به دو دسته طبقه‌بندی کرد:i)خلاصه-سازی چکیده ii)خلاصه‌سازی استخراجی [۳]. در خلاصه‌سازی چکیده ای لذا از کپی جملات متن اصلی در خلاصه استفاده نمی‌شود و به همین دلیل به ماشین قدرتمندتر و محاسبات پیچیده‌تری نسبت به خلاصه‌سازی استخراجی نیاز دارد و محتوای غنی‌تری از متن ورودی در خلاصه گنجانده شده و خلاصه‌ی فشرده‌تری به‌دست می‌آید [۴]. در خلاصه‌سازی استخراجی که رویکردی تهی از دانش است از بین جملات متن ورودی با توجه به فرآیند پیش‌پردازش و سپس وزن‌دهی به لغات و جملات، جملاتی که دارای وزن بیشتری هستند دارای اهمیت بیشتری بوده و به‌عنوان خلاصه‌ی نهایی عرضه می‌شوند. در خلاصه‌سازی استخراجی پیچیدگی محاسباتی کمتر بوده لذا بیشتر خلاصه‌سازها بر اساس خلاصه‌سازی استخراجی کار می‌کنند [۱].

۱- بیان مسأله

با رشد روزافزون منابع متنی در اینترنت، هر روزه بر اطلاعات قابل دسترس برای کاربران افزوده می‌شود و سامانه‌ی خلاصه‌سازی خودکار متن راهکاری برای جلوگیری از افزونگی اطلاعات و سردرگمی کاربران در گزینش اطلاعات است. از آنجایی که در کارهای قبلی از رویکردهای حریصانه برای انتخاب جملات استفاده می‌شده است و رویکرد حریصانه در دسترسی به بهینه سراسری مناسب نیست به این جهت از آخرین نسخه‌ی الگوریتم کلونی زنبور مصنوعی(ABC) استفاده می‌شود، زیرا برای دستیابی به راه‌حل‌های بهینه در سطح سراسری الگوریتم هوشمندانه ایست که از رفتار خوراک جویی زنبور عسل الهام گرفته است [۵].

الگوریتم کلونی زنبور عسل دارای سه نوع زنبور کارگر، ناظر و نگهبان است. در روشی که در این تحقیق استفاده می‌شود به‌جهت اینکه توسط کنترل رفتار زنبور نگهبان باعث بهبود عملکرد اکتشاف می‌شود باعث دستیابی به خلاصه‌ی بهتری در سطح سراسری می‌شود. با مطالعاتی که انجام گرفت، شکافی در مرحله‌ی پایانی خلاصه‌سازی که همان انتخاب جملات خلاصه است کشف شد و آن هم این بود که با رویکردهایی که تا کنون برای خلاصه‌سازی استفاده می‌شد، رسیدن به خلاصه بهینه سراسری مشکل بوده است [۵].

۲- روش‌های پیشین

در روش منطق فازی مورد استفاده برای خلاصه‌سازی متن، در مرحله‌ی اول متن ورودی پیش‌پردازش می‌شود و ارزش ویژگی‌هایی همچون: ویژگی عنوان، طول جمله، وزن عبارت، شباهت جمله به جمله و. .... بدست آورده می‌شوند، سپس این مقادیر ویژگی‌ها در سیستم منطق فازی جهت تولید امتیاز برای هر جمله استفاده می‌شوند. در مرحله‌ی بعد با استفاده کردن قوانین فازی در موتور فازی بر اساس وزن اختصاص یافته به هر جمله با تراز کردن همه‌ی ویژگی‌ها برای انتخاب کردن یا انتخاب نکردن جملات، خلاصه نهایی ایجاد می‌شود [۶].

اختصاص پارامترهای باینری به ویژگی‌های جملات شکافی بود که در منطق فازی با تخصیص مقدار بین ۰ و ۱ (پیوسته) تابع تخصیص ویژگی دقیق‌تر عمل کرده و با تعریف ویژگی‌های داده شده به عنوان کیفیت‌های فازی این مشکل حل می‌شود لذا در خصوص هر ویژگی هر چه مقدارش به ۱ نزدیک‌تر باشد امتیاز بالاتری به جمله داده می‌شود [۶].

در [۷] تکنیک خلاصه‌سازی موضوع محور و از نوع خلاصه‌سازی استخراجی بوده و از آنجایی که براساس پرسش و پاسخ است، پاسخی دقیق‌تر در قالب خلاصه تولید می‌کند. این روش به صورت مختصر به‌شرح زیر است:

A-در ابتدا یک برچسب به جمله تعلق می‌گیرد.

B-سپس با تجزیه و تحلیل کلمات اولیه، نوع سؤال تعیین می‌شود: مثلاً اگر سؤال با "کجا" شروع شود، سپس باید "مکان" برگردانده شود.

C-در صورت امکان تعیین سؤال، فیلتر انجام و روند تطبیق الگو صورت می‌پذیرد، در غیر این‌صورت پاسخ به "خالی"[1]تنظیم می‌شود.

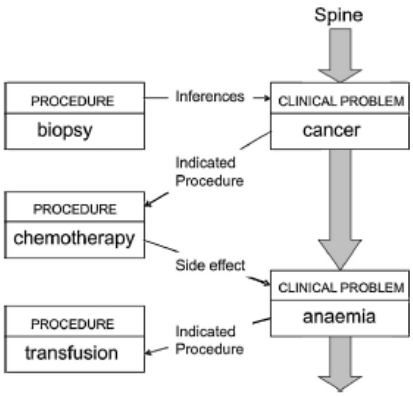
در روشی که خلاصه‌سازی سند تعاملی(IDS)[2]است، کنترل پویای ویژگی‌های جملات سند را فراهم می‌کند، به‌طوری که تغییرات ایجاد شده توسط کاربر فوراً در خلاصه روی صفحه نمایش بازتاب داده شود. در روش مورد بحث تکنیک استخراج جمله با بهره‌گیری از

[1]- Nil
[2]- Interactive Document Summarisation

ISSN: 2745-3673
http://www.Rcsj.ir

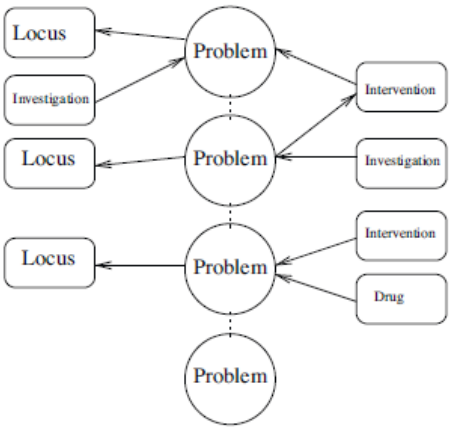
عبارات کلیدی[3]انجام می‌شود. قابل ذکر است که این الگوریتم از تکنیک‌های یادگیری ماشین[4]استفاده می‌کند و برای یادگیری به یک مجموعه از اسناد آموزشی تأکید می‌کند، برای اینکه یک مجموعه از عبارات کلیدی الگو برای آموزش داشته باشد [۸].

روش [۹] که بر روی خلاصه‌سازی داده-به-متن یادداشت‌های پزشکی توسط کامپیوتر با هدف بی‌خطا بودن و دقیق‌تر بودن خلاصه‌ها برای استفاده‌ی عموم پزشکان متمرکز است با به‌کار بردن تکنیک‌های NLP[5]و به کمک گراف‌ها ارتقاء یافته است و بیشتر خلاصه‌ی تولیدی نمایشگر این است که "چه بیماری"، "چه زمانی"، "چه معالجه‌ای" و "توسط چه پزشکی" و اطلاعات دیگر که به صورت طبقه‌بندی شده برگردانده می‌شود که رویدادنامه نامیده می‌شود. برای ارتباط بین رویدادها، پیامدها در طراحی انتخاب مفهوم اهمیت دارد و مرحله‌ی ساختاربندی متن، این کار و تولید متن را تسهیل می‌کند و اجازه می‌دهد درجه‌ی بالاتری از انعطاف‌پذیری متن ایجاد شود. طراحی سیستم از مولد گزارش یک معماری کانال ارتباطی کلاسیک را با یک انتخاب کننده‌ی مفهوم، یک ریزترتیب دهنده و فهمیدن نحوی پیگیری می‌کند



شکل ۱- نتیجه‌ی تعیین محتوا[۹]

در شکل(۱) که در رویکرد موردنظر کاربرد دارد اتفاقات درونی ساختار نشان داده نمی‌شود، و معیار انتخاب مفهوم به‌طور نمادین یک گراف معنایی شامل یک خط‌الرأس از اتفاقات کانونی بازیابی می‌کند.



شکل ۲- مثالی از خلاصه‌سازی داده به متن با یک کانون و عمق۱[۹]

در شکل(۲) رویدادها به رویدادهایی از نوع تشخیص به یک سطح عمیق متصل می‌شوند، که عمق ۰ تنها مصداق تشخیص بیماری و عمق ۱ پیامد تشخیص بیماری مثل "جراحی" را استخراج می‌کند اما هیچ اطلاعات بیشتری روایت نمی‌شود.

[3]- Key Terms
[4]- Machine Learning
[5]- Natural Language Process

ISSN: 2745-3673
http://www.Rcsj.ir

تصمیم‌گیری چگونگی برای گفتن آن با شروع از یک گراف معنایی برجستگی-پایه، یک ترتیب از پاراگراف‌ها معمولاً یکی برای همه برنامه‌ریزی می‌شود بنابراین قوانین تولیدشده‌ی اختصاصی می‌تواند به‌طور مساوی برای تولید خلاصه‌های پزشکی انگلیسی استفاده شود.

در نهایت مفهوم در میان یک مجموعه از جملات ساخته شده‌ی بالای پاراگراف توزیع می‌شود.

در روش بالا سیستم تولید متن دارای دو نوع از قوانین گرامری روابط (۱) و (۲) است شامل:

۱)گرامر تولیدی استاندارد برای گزاره‌ها و جملات انگلیسی، که متشکل از قوانین نحوی مانند زیر است:

(۱) (برای گزاره‌ها)---> {اسم}+{تعیین کننده}+{بند معین}=گزاره‌های اسم معین

+اتصال دهنده‌ی واژه‌ی سببی + واژه‌ی سببی اصلی = ارتباط واژه‌ی سببی(مانند:therefor و for)

(۲) (برای جملات) -------> واژه‌ی سببی ثانویه

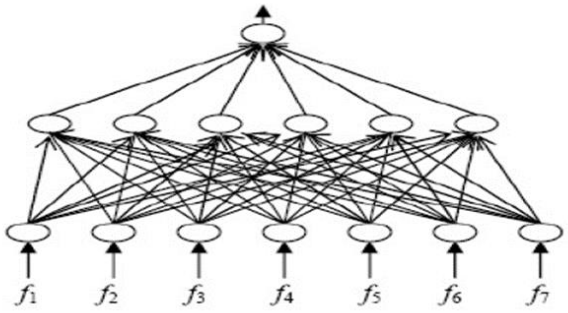
مجموعه‌ی دوم که از قوانین تولید به‌دست می‌آیند به کنترل اندازه‌ی خلاصه مربوط می‌شوند و روش بیرون کشیدن اتصالات بین کلمات در واژگان را اداره می‌کنند.

روش دیگری که برای خلاصه‌سازی متن استفاده می‌شده است، نمایندگی لغت توزیع شده با تجزیه و تحلیل تفکیک[۶](mRMR) برای امتیاز دادن عبارات و سپس جملات بررسی شد و در نهایت یک الگوریتم استخراجی جدید برای مقابله با مشکل افزونگی پیشنهاد شد، در این روش ابتدا پیش‌پردازش چند زبانه به منظور اینکه وابستگی زبان حداقل شود صورت می‌پذیرد. در مرحله‌ی بعد که خوشه بندی جملات است، جملات مشابه در یک خوشه قرار گرفته و طی پروسه‌ی بعدی جملات خلاصه انتخاب می‌شوند. برای محاسبه‌ی ویژگی شباهت جمله به جمله با رابطه (۴) به‌جای استفاده از شباهت کسینوس که با متریک (۳) محاسبه می‌شده است، استفاده می‌شود [۱]:

$$match(w_i) = \arg \max_{w_J \in S_2} Sim\,(Rep(w_i).Rep(w_j))$$

$$Sim(S_1.S_2) = \frac{\Sigma_i \,\mathrm{Match}(w_i) + \Sigma_j \,\mathrm{Match}(w_j)}{|S_1| + |S_2|} \qquad (۴)$$

(۳) روش خلاصه‌سازی متن با شبکه‌های عصبی [۱۰] که در آن با استفاده‌ی شبکه‌ی عصبی یاد می‌گیرد که چه جملاتی را باید در خلاصه قرار دهد، بر روی چند پاراگراف ایجاد شده و تشخیص می‌دهد که جمله در خلاصه بیاید یا نه؟ شبکه‌های عصبی از الگوهای ذاتی برای انتخاب جملات و قرار دادن یا قرار ندادن در خلاصه نهایی استفاده می‌کنند. سطح اول که جهت انتخاب جملات به شبکه‌ی عصبی داده می‌شود به شکل (۴) است.

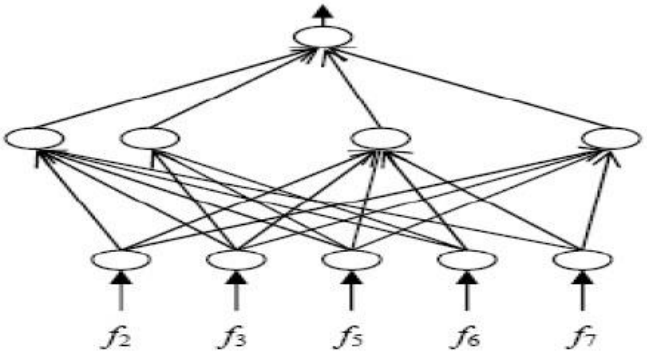


شکل ۴ –شبکه عصبی پس از آموزش[۱۰]

هنگامی‌که شبکه آموزش دید و ویژگی‌ها را یاد گرفت، نیاز به کشف روندها و روابط بین جملات است که متشکل از دو مرحله می-شود: ۱) از بین بردن ویژگی‌های غیر معمول ۲) حذف تأثیرات ویژگی‌های مشترک اتصالاتی که وزن سبک دارند، چون تأثیری بر عملکرد شبکه ندارند، قابل حذف‌اند و در انتها تمام ارتباطات بی‌تأثیر در لایه‌ی شبکه مطابق شکل (۵) از بین می‌روند.

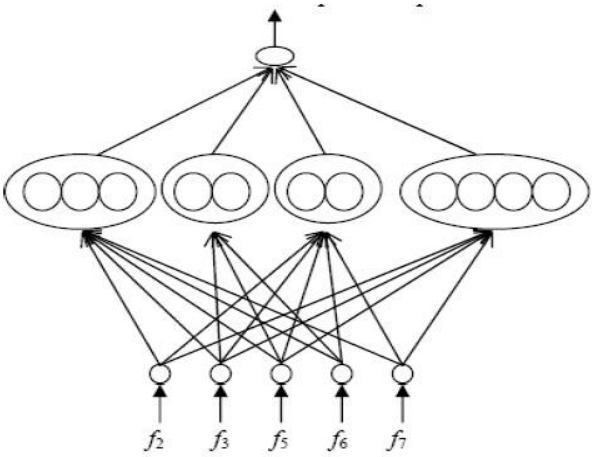
[6]- Minimum Rdundancy Maximum Relevance

ISSN: 2745-3673
http://www.Rcsj.ir



شکل ۵-شبکه عصبی پس از هرس[۱۰]

با خوشه‌بندی تطبیقی، لایه‌ها خوشه‌بندی شده و هر خوشه بر اساس جرم آن شناخته می‌شود، همان‌طور که در شکل(۶) نشان داده می‌شود در نهایت پس از محاسبه و ارائه پارامترهای کنترل برای رتبه‌بندی خوشه‌ها خلاصه نهایی ایجاد می‌شود.



شکل ۶-شبکه عصبی پس از ترکیب ویژگی[۱۰]

تکنیکی که یک الگوریتم تکاملی را برای خلاصه‌سازی خودکار متن توسعه بخشید به طور مختصر به‌شرح زیر است [۱۱]:
در این روش برای تطبیق دادن فرآیند خوشه‌بندی جمله از اندازه‌گیری شباهت ضریب جاکارد استفاده می‌کند، مطابق رابطه (۵) :

$$sim_{jaccard}(S_k, S_l) = \frac{S_k \cap S_l}{S_k \cup S_l} \qquad (۵)$$

در رابطه جهت بهبود کارایی خوشه استفاده می‌شود و برای انتخاب جملات خلاصه از روش ویژگی-پایه به‌جای رویکرد جمله پایه استفاده می‌کند، دلیلش این است که روش‌های جمله-پایه در دریافت ارتباط بین یک جمله و جملات دیگر ناتوان هستند. تفاوت دیگرش با روش‌های قبلی این است که در تطبیق فرآیند خوشه‌بندی جمله یک مدولاتور واقعی-به-صحیح استفاده شده است و روال کار با تخصیص یک فضای برداری امتیاز-ویژگی به متن پیش‌پردازش شده آغاز می‌شود، سپس راه‌اندازی تمامی خوشه‌ها برای جملات متن هر دو اندازه-گیری شباهت ضریب جاکارد و مدولاتور واقعی -به- صحیح اعمال می‌شود و بدین طریق خوشه‌ها بهینه می‌شوند، در مرحله‌ی بعد یک کروموزوم نماینده‌ی جملات تحت عنوان راه‌حل‌ها به منظور قرارگیری در خلاصه، فرموله می‌شود که ژن‌ها نیز نشان دهنده‌ی جملات هستند.

فرمولی که دامنه‌ی مدولاتور واقعی- به-صحیح را محاسبه می‌کند، در (۶) آمده است:

$$R_{ch} = \frac{m}{k} \qquad (۶)$$

ISSN: 2745-3673
http://www.Rcsj.ir

که در رابطه‌ی (۶)، m بالاترین مقدار مرزی، از مقادیر تولید شده‌ی تکامل دیفرانسیل(DE) است و k تعداد خوشه‌ی موردنیاز است.

در انتها برای تطبیق دادن کیفیت خوشه‌بندی پارتیشن شده، تابع هدف وارد کار می‌شود که ترکیب دو تابع معیار به‌نام‌های شباهت درون-خوشه و اختلاف درون-خوشه است، هدف این تابع تطبیق درجه‌ی شباهت بین جملات، بدست آوردن امتیاز شباهت ماکسیمم، مینیمم اختلاف خوشه و سپس بالانس(تراز) کردن دو تابع مطابق فرمول‌های (۲-۱۳)، (۱۴-۲) و (۱۵-۲) است :

$$F1 = \sum_{l=1}^{k} |C_l| \sum_{s_t, S_j \in C_l} sim_{NGD}(S_i, S_j) \rightarrow max \quad (۷)$$

$$F2 = \sum_{l=1}^{k-1} \frac{1}{|C_l|} \sum_{m=l+1}^{k} \frac{1}{|C_m|} \sum_{S_i \in C_t} \sum_{S_j \in C_m} sim_{NGD}(S_i, S_j) \rightarrow min \quad (۸)$$

$$F = (1 + sigm(F_1))^{F_2} \rightarrow \max \quad (۹)$$

که (۷) و (۸) فرمول‌های توابع معیار هستند و (۹) فرمولی برای بالانس بین این دو تابع است.

در روش [۱۲] که مبنای کار برای خلاصه‌سازی متن، الگوریتم بهینه‌سازی کلونی مورچگان (ACO)[7]بوده است، و دلیل کاربرد ACO در این حوزه این بود که الگوریتم مذکور از رفتار هوش توده‌ای مورچه‌ها برای یافتن منابع غذایی استفاده می‌کند. با به‌کار بردن این ایده‌ی الهام گرفته از طبیعت، سعی در بالا بردن دقت خلاصه‌سازی داشته است و نیز با انتخاب جملات بر اساس امتیازدهی آن‌ها با استفاده‌ی بهترین ویژگی‌ها و تطبیق وزن‌دهی به هر یک از آنها نتایج کل را بهبود بخشیده است. در این روش خلاصه‌سازی بر اساس رویکرد استخراجی بوده که برای استخراج ویژگی‌ها از متن اصلی، الگوریتم ACO وارد عمل می‌شود. الگوریتم کلونی مورچگان بر اساس ایده زیر کار می‌کند:

یک مورچه در حال حرکت، مقداری فرومون(در اندازه‌های مختلف) از خود بر زمین باقی می‌گذارد و بدین ترتیب مسیر را به وسیله-ی بوی این ماده مشخص می‌سازد. هنگامی که یک مورچه به‌طور تصادفی و تنها حرکت می‌کند با مواجه شدن با مسیری که دارای اثر فرومون بیشتری است، به احتمال زیاد مسیر فوق را انتخاب کرده و با فرومونی که از خود برجای می‌گذارد، اثر آن را در مسیر مذکور تقویت می‌نماید. در الگوریتم ACO رفتار مورچه‌ها در یافتن غذا نوعی هوشمندی توده‌ای[8]است نه هوشمندی اجتماعی، به این معنی که کلونی مورچه‌ها نه بر اساس هوشمندی یک مغز مرکزی بلکه بر اساس توده‌ای از عامل‌های هوشمند(مورچه‌ها) و رابطه‌ی بین آن‌ها عمل می‌کند، در هوشمندی توده‌ای عناصر، رفتاری تصادفی دارند، بین آنها هیچ نوع ارتباط مستقیمی وجود ندارد و آن‌ها به طور غیر مستقیم و با استفاده از نشانه‌ها با هم در تماس هستند.

در آخرین تحقیقی که در خصوص خلاصه‌سازی خودکار متن بررسی شد، با استفاده از بهینه‌سازی زنبور عسل مصنوعی(ABC)[9]، نسخه‌ی(۲۰۱۴) آن که ABC-سریع نام داشت، برای حل مشکل بهینه‌سازی خلاصه‌سازی متن، یک رویکرد جدید پیشنهاد شد [۱۳] .

اصلاح معادلات هسته‌ای از الگوریتم کلونی زنبور عسل مصنوعی-سریع در طول مراحل مختلف، یعنی مقداردهی اولیه، فاز زنبور عسل کارگر، فاز زنبور عسل ناظر(تماشاچی)، محاسبات تناسب[10]انجام شد. در این روش به منظور افزایش اثربخشی در کل روند، در مرحله-ی دوم یعنی پس از پیش‌پردازش، طبقه‌بندی سطح بالا برای کمک به طبقه‌بندی اولیه انجام می‌شود به طوری که فاز بعدی شناسایی موضوع قادر به تولید خوشه‌ی همگن‌تر شود و کیفیت این خلاصه‌سازی توسط وسعت پوشش محتوا وکاهش افزونگی صورت می‌پذیرد [۱۳].

۳- روش پیشنهادی

رویکردی که از الگوریتم کلونی زنبور عسل مصنوعی(ABC) جهت بهینه‌سازی خلاصه‌سازی خودکار متن در خلاصه‌سازی خودکار متن استفاده کرده بود، نسبت به استفاده از روش حریصانه بهتر عمل کرده و در به حداکثر رساندن پوشش محتوا و به حداقل رساندن افزونگی موفق بوده است زیرا الگوریتم‌های هوش جمعی مانند الگوریتم ABC در یافتن راه‌حل بهینه بسیار کارا عمل می‌کنند به دلیل این-که ذاتاً دارای مکانیسم بهبود راه‌حل و جستجوی محلی است [۱۴].

الگوریتم ABC که شامل سه نوع زنبور کارگر، ناظر و نگهبان است که هر کدام در عملکرد خود متفاوت هستند، در الگوریتم مذکور موقعیت منبع غذایی نشان‌دهنده‌ی یک راه‌حل ممکن برای مشکلی است که قرار است بهینه شود و شهد منبع غذایی به کیفیت راه حل

[7] Ant Colony Optimization
[8] - Swarm
[9]-Artificial Bee Colony
[10] -Fitness

ISSN: 2745-3673
http://www.Rcsj.ir

مرتبط است [۱۵]. در طول هر چرخه، اتفاقی که می‌افتد این است که زنبورهای کارگر و ناظر به سمت منابع غذایی به صورت تصادفی حرکت می‌کنند، سپس محاسبه‌ی شهد، تعیین زنبور نگهبان و در آخر حرکت به سوی منابع غذایی جدید اتفاق می‌افتد اگر راه‌حل بهبود نیافت منبع غذایی رها می‌شود. در الگوریتم ABC مقدار حد[1]یک پارامتر کنترل مهم است که پس از حد مجاز زنبور کارگر به زنبور نگهبان تبدیل شده و شروع به جستجو برای منابع غذایی جدید می‌کند.

همان‌گونه که در الگوریتم ABC پیداست زنبور نگهبان زمانی عمل می‌کند که شاخص محاکمه از پارامتر حد بیشتر شود و شاخص محاکمه زمانی افزایش می‌یابد که راه‌حل بهبود نیابد، اگر در طول فرآیند، زنبورهای کارگر و ناظر بهبود یابد شاخص محاکمه به صفر تنظیم می‌شود. برای مشکلات خاصی مانند این‌که اگر راه‌حل های محلی به‌طور مداوم بهبود یابد، شاخص محاکمه همیشه به صفر تنظیم می‌شود و بنابراین شاخص محاکمه از حد(Limit) تجاوز نمی‌کند در نتیجه باعث می‌شود اکتشاف سراسری توسط زنبور عسل نگهبان به مشکل برخورد کند پس باید به دنبال برطرف کردن این محدودیت راه‌حلی جایگزین کرد به همین دلیل در بهینه‌سازی خلاصه‌سازی متن الگوریتم بهبود یافته‌ی ABC جایگزین می‌شود. فرآیند زنبور نگهبان توسط نرخ تغییر کلونی زنبور عسل(ABC-ROC) بهبود یافته و بجای راه-حل‌های محلی، توسط شیب در نمودار، راه‌حل‌های سراسری بدست خواهد آمد [۹].

در ابتدا فایل متنی به‌صورت یک گراف نمایش داده می‌شود زیرا می‌توان ویژگی‌های مؤثرتری را از تئوری گراف‌ها استخراج کرد. پس از انجام پیش پردازش متن یعنی؛ مرزبندی جملات، حذف کلمات مقطع(کلماتی که دارای مفهوم خاصی نیستند) و ریشه‌یابی، هر جمله در یک بردار (گره گراف) لحاظ می‌شود و با این روش متن را به آرایه‌ای از اعداد تبدیل کرده که پردازش آن توسط نرم‌افزار راحت‌تر شود. گراف بکاربرده شده شامل دو مجموعه‌ی (V,E) است که در آن V شامل رئوس و E شامل یال‌ها می‌باشد، برای ساخت گراف یک متن مراحل زیر طی می‌شود.

-به ازای هر S_i عضو فایل، S_i را به V اضافه می‌کنیم.

-به ازای هر S_i و S_j عضو V، اگر S_i از نظر تقدم زمانی قبل از S_j در فایل آمده باشد، آنگاه (S_i, S_j) را به E اضافه می‌کنیم.

بعد از ساخت گراف باید مرحله‌ی امتیازدهی به جملات انجام شود. در روش پیشنهاد شده مطابق فرمول‌های (۱۰) و (۱۱) از سیستم وزن‌دهی $tf - isf$ استفاده شده و هر کدام از جملات بر اساس تعداد تکرار کلمات، وزنی به خود تخصیص می‌دهد در انتها طبق رابطه‌ی (۱۲) از ضرب دو رابطه‌ی اخیر، وزن نهایی در قالب Wi,j بدست می‌آید [۱۶] :

$$tfi,j = \frac{\text{freq } i,j}{\max freq\ i,j} \quad (۱۰)$$

$$isf\ i = \text{Log}\frac{N}{n\,i} \quad (۱۱)$$

$$Wi,j = tfi,j \times isf\ i \quad (۱۲)$$

به‌گونه‌ای که tfi,j، بیانگر تعداد کلمات iام در جمله‌ی jام و isf نشان‌دهنده‌ی عکس تعداد تکرار جمله از کلمه‌ی i است. در رابطه‌ی (۱۱)، N نشان دهنده‌ی تعداد کل جملات و ni تعداد جملاتی است که کلمه‌ی iام در آن وجود دارد.

فاکتورهای دیگری که در مرحله‌ی پردازش باید روی متن پیش‌پردازش شده اعمال شود، مانند شباهت جملات با عنوان متن به صورت روابط (۱۳) و (۱۴) محاسبه میشوند:

$$sim(Si,q) = \frac{\sum_{i=1}^{t} Wi,j \times Wi,q}{\sqrt{\sum_{i=1}^{t} Wi,j \sqrt{\sum_{i=1}^{t} Wi,q}}} \quad (۱۳)$$

$$Wi,q = (0.5 + \frac{0.5 \times freq\ i,q}{\max freq\ i,q}) \times isf \quad (۱۴)$$

[1] -Limit [1]

ISSN: 2745-3673
http://www.Rcsj.ir

که $sim(Si,q)$ نشان دهنده‌ی شباهت جمله‌ی iام و عنوان است، W هم که وزن جملات بوده و توسط رابطه‌ی (۶-۳) و (۳-۸) محاسبه شده و در رابطه‌ی شباهت استفاده می‌شود، $freq\ i,q$ تعداد تکرار عنوان متن در جمله‌ی iام است و $\max freq\ i,q$ مؤلفه‌ایست که حداکثر تکرار عنوان متن در جمله‌ی iام است. رابطه‌ی دیگری که در مرحله بعد وارد عمل شده و مرحله‌ی پایانی شباهت جملات با عنوان را قابل محاسبه می‌سازد در رابطه (۱۵) آمده است:

$$TR = \frac{\sum_{i=1}^{t} sim(Sj,q)}{\max sim(Sj,q)} \qquad (۱۵)$$

پارامتر دیگری که در رویکردهای دیگر نیز فاکتور مهمی بوده است، شباهت دو جمله نسبت به هم است که در رابطه (۱۶) تعریف می‌شود:

$$sim(Sm,Sn) = \frac{\sum_{i=1}^{t} Wi,m \times Wi,n}{\sqrt{\sum_{i=1}^{t} Wi,m^2}\sqrt{\sum_{i=1}^{t} Wi,n^2}} \qquad (۱۶)$$

از آنجایی که طول جمله هم در امتیاز گرفتن جملات مؤثر بوده (جملات خیلی کوتاه و جملات خیلی بلند در متن خلاصه نباید بیایند که اصلی مسلم در تولید خلاصه‌ی خوب است)، لذا از رابطه‌ی (۱۷) می‌توان به این پارامتر مهم دست یافت:

$$Length = \frac{leni}{2\max leni} \qquad (۱۷)$$

که $Length$ طول جملات، $leni$ ، طول جمله‌ی iام و $2\max leni$، دو برابر ماکسیمم طول جملات است. مزیت استفاده از الگوریتم ABC-ROC در روش ایده به‌گونه ایست که در وضعیت عدم بهبود راه‌حل (در صورت رکود یا گیر کردن در مینیمم محلی) به‌طور مستقیم از نمودار می‌توان وضعیت را تعیین کرد الگوریتم مذکور از سه پارامتر کنترل به‌نام‌های max- ,max-Roc ,max-Trace تشکیل شده و کارشان به ترتیب زیر است: max-Tace نقطه‌ی شروع برای محاسبه‌ی شیب را تصمیم می‌گیرد، max-Roc یک مقدار حداکثر از شیب، قبل از تماس زنبور نگهبان در نظر گرفته می‌شود و max-flag وضعیتی که نمودار بهبود نمی‌یابد و شیب=۰ است را پیگیری می‌کند [۱۴] .

۴- یافته‌های پژوهش

در این فصل، به ارزیابی رویکرد معرفی شده و تشریح نتایج حاصل از آن خواهیم پرداخت. نمایش فایل‌های متنی به صورت یک گراف، یک ایده قدیمی است که به صورت گسترده استفاده می‌شود. دلیل این نوع نمایش این است که ما می‌توانیم ویژگی‌های مؤثری را از تئوری گراف‌ها استخراج کنیم و مسأله را به راحتی با رویکرد گراف-مبنا تطبیق دهیم. به منظور آزمایش رویکرد پیشنهادی، ما ابتدا یک فایل کوتاه را برای خلاصه‌سازی در نظر می‌گیریم. در ادامه، سعی خواهیم کرد که با استفاده از رویکرد معرفی شده، خلاصه‌ای از این متن را در قالب n جمله استخراج کنیم که n تعداد جملاتی است که باید در خلاصه بیاید و باتوجه به ضریب فشردگی میتواند متغیر باشد .

۴-۱- نمایش

در این بخش، ما متن مورد نظر را به صورت گراف بدون جهت[2]نمایش می‌دهیم. به این منظور، متن را به جملاتی تقسیم می‌کنیم و هر جمله را در یک بردار (که می‌تواند شاخه یا گره گراف باشد) لحاظ می‌کنیم. شکل ۴-۲ تقسیم متن به جملات را نشان می‌دهد. به عنوان یک عمل پیش‌پردازش، می‌توان در این مرحله متن را ویرایش کرد . این ویرایش شامل موارد زیر می‌باشد:

[1] -Directed Acyclic Graph [2]

ISSN: 2745-3673
http://www.Rcsj.ir

حذف حروف
حذف کلمات اضافه
حذف فعل‌ها
حذف علائم نگارشی
جایگزینی کلمات با ریشه‌ی آن‌ها

گراف نمایشی، از دو مجموعه تشکیل می‌شود، (V, E) که در آن V شامل رئوس و E شامل یال‌ها می‌باشد. برای ساخت گراف یک متن، باید مراحل زیر را طی کرد:

- به ازای هر s_i عضو فایل، s_i را به V اضافه می‌کنیم.
- به ازای هر s_i و s_j عضو V، اگر s_i قبل از s_j در فایل آمده باشد(از نظر زمانی)، آنگاه (s_i, s_j) را به E اضافه می‌کنیم.

پس از ساخت گراف، طبق مراحل بخش ۲-۵، حال می‌توان امتیازها را به آن اختصاص داد.

۴-۲- امتیازدهی

برای وزن‌دهی به جملات، ما از سیستم وزن‌دهی $tf - isf$ استفاده می‌کنیم. این سیستم وزن‌دهی مشابه مدل برداری کلاسیک IR است. وزن‌های $tf - isf$ برای هر جمله حساب شده سپس برای هر جمله و عنوان، یک بردار ایجاد می‌شود. عنوان نقش استفسار[1][3] را ایفا می‌کند

.

۴-۳- تابع هدف

ما تابع هدف را با داشتن خاصیت انعطاف‌پذیری طراحی می‌کنیم، به این معنی که، پارامترهایی را دارد که توسط کاربر می‌توانند تنظیم شوند. شخصی ممکن است خواهان یک متن خلاصه با بیشترین قابلیت خوانایی باشد، در حالی که، شخص دیگری ممکن است خواهان متن خلاصه‌ای باشد که بیشترین ارتباط را با عنوان داشته باشد.

برای داشتن چنین عملکردی، ما یک تابع هدف را که به صورت میانگین وزنی فاکتورها تعریف می‌شود، همانند فرمول (۱۸) طراحی خواهیم کرد :

$$F = \frac{a*W_1+\beta*W_2+\gamma*W_3+\mu*W_4}{a+\beta+\gamma+\mu} \qquad (۱۸)$$

در رابطه فوق، a، β، γ و μ اعداد حقیقی هستند که توسط کاربر مقداردهی می‌شوند. W_iها، $i = 1, \dots, 4$، فاکتورهای تعریف شده در بخش قبل هستند که مقادیر آن‌ها بین ۰ و ۱ قرار دارد.

۴-۴- نتایج

جدول (۱) تنظیمات الگوریتم ABC-ROC را نشان می‌دهد. در این جدول، NB تعداد زنبورها، NF شمار غذاها و $maxcycle$ تعداد دور الگوریتم را نشان می‌دهد.

جدول ۱- تنظیمات الگوریتم ABC-ROC

پارامتر	NB	NF	Maxcycle
مقدار	۱۰	۵	۵۰

[1] - Query [3]

ISSN: 2745-3673
http://www.Rcsj.ir

۴-۵- ارزیابی الگوریتم ABC-ROC

ما از رویکرد معرفی شده برای خلاصه‌سازی دو مجموعه داده متنی DUC2001 و DUC2002 به کار خواهیم برد. این دو مجموعه داده، منابع آزادی هستند که می‌شود آن‌ها را از سایت http://duc.nist.gov تهیه کرد. در مرحله پیش‌پردازش، حروف اضافه و توقف[۴]را در هر جمله، با استفاده از لیست توقف تهیه شده در ftp://ftp.cs.cornell.edu/pub/smart/english.stop، حذف می‌کنیم و ریشه کلمات باقیمانده را با استفاده از طرح Porter پیدا می‌کنیم.

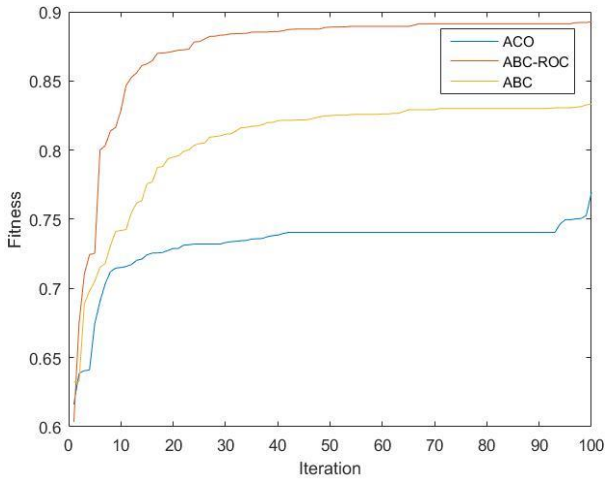
برای مقایسه، ما علاوه بر رویکرد معرفی شده خود، از الگوریتم‌های ABC و الگوریتم کلونی مورچگان نیز به منظور خلاصه‌سازی متن استفاده می‌کنیم. جدول (۲) تنظیمات این الگوریتم‌ها را برای مجموعه داده‌های DUC2001 و DUC2002 نشان می‌دهد.

جدول۲- تنظیمات الگوریتم‌های ABC و ABC-ROC

تنظیمات / روش‌ها	مجموعه داده‌ها	Limit	NF	α	P	nPop	MaxIt
ACO	DUC2001	-	-	۲	۰,۰۲	۳۰	۱۰۰
	DUC2002	-	-	۲	۰,۰۲	۲۰	۸۰
ABC	DUC2001	۱۰۰	۱۵	-	-	۳۰	۱۰۰
	DUC2002	۱۰۰	۱۰	-	-	۲۰	۸۰
ABC-ROC	DUC2001	۱۰۰	۱۵	-	-	۳۰	۱۰۰
	DUC2002	۱۰۰	۱۰	-	-	۲۰	۸۰

در جدول (۲)، MaxIt تعداد تکرار، nPop جمعیت اولیه (مورچه‌ها و زنبورها)، α و β پارامترهای الگوریتم مورچگان و NF و Limit پارامترهای مربوط به الگوریتم زنبور عسل هستند. به دلیل حجم کمتر متن DUC2002 نسبت به DUC2001، ما تعداد تکرارها را برای متن اول بیشتر در نظر گرفته‌ایم.

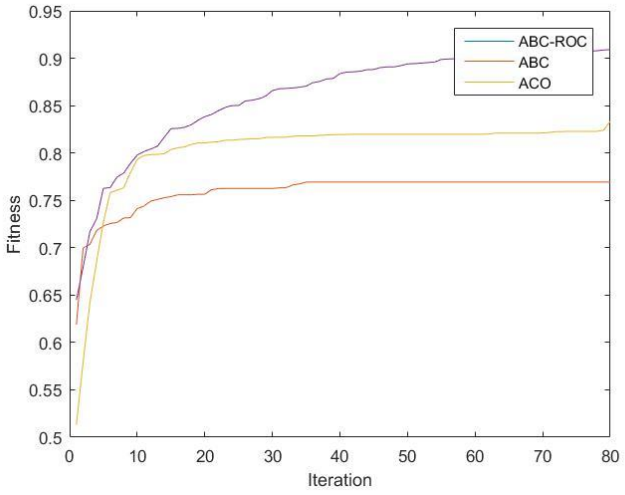
شکل ۷-الف و ب، به ترتیب، نتایج دقت خلاصه‌سازی بدست آمده از اعمال الگوریتم‌ها، به متن‌های DUC2001 و DUC2002 را نشان می‌دهد. مشاهده می‌شود که، روند افزایش تابع ارزیابی در الگوریتم ABC-ROC، در خلاصه‌سازی هر دو متن، از عملکرد بهتری نسبت به بقیه الگوریتم‌ها برخوردار است. جدول ۴-۵ معیارهای Precision، Recall و F1 بدست آمده از این سه الگوریتم در حل مسأله خلاصه‌سازی متن را نشان می‌دهد.



(الف)

[1] Stopword - [4]

ISSN: 2745-3673
http://www.Rcsj.ir



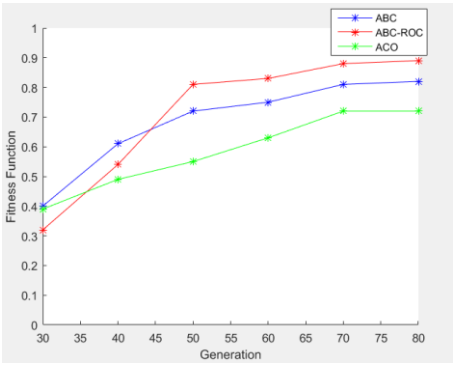
(ب)

شکل۷- نتایج بدست آمده از خلاصه سازی متون (الف) DUC2001 و (ب) DUC2002؛ با استفاده از الگوریتم‌های ACO، ABC و ABC-ROC

جدول ۳-معیارهای ارزیابی الگوریتم‌های ACO، ABC و ABC-ROC در مساله خلاصه‌سازی متن

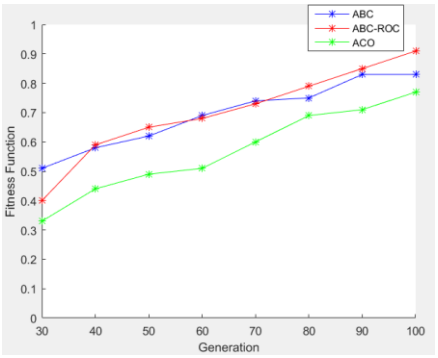
روش	داده	Precision	Recall	F1
ABC-ROC	DUC2001	۰,۸۹	۰,۶۰	۰,۶۷
	DUC2002	۰,۹۱	۰,۵۹	۰,۶۵
ABC	DUC2001	۰,۸۲	۰,۶۳	۰,۷۰
	DUC2002	۰,۸۳	۰,۶۵	۰,۷۲
ACO	DUC2001	۰,۷۲	۰,۷۶	۰,۷۵
	DUC2002	۰,۷۷	۰,۷۱	۰,۷۳

شکل ۸-الف و ب، تابع ارزیابی بدست آمده از روش‌های معرفی شده، برای مجموعه داده‌های DUC2001 و DUC2002 را بر حسب نسل‌های مختلف نشان می‌دهد. در رابطه با هر دو مجموعه داده، می‌توان گفت الگوریتم ABC-ROC عملکرد بهتری در بهینه-سازی تابع هزینه داشته و بدترین عملکرد را الگوریتم ACO داشته است.



(الف)

ISSN: 2745-3673
http://www.Rcsj.ir



(ب)

شکل۸-تابع ارزیابی بدست آمده از روش‌های معرفی شده، برای مجموعه داده‌های (الف) DUC2001 و (ب) DUC2002 را بر حسب نسل‌های مختلف

جدول (۴) ، مقایسه بین روش پیشنهادی با خلاصه‌سازهای Microsoft word و Copernic، برای مجموعه داده‌های Duc2001 و Duc2002 را نشان می‌دهد. این خلاصه‌سازها، به عنوان معیاری برای مقایسه الگوریتم پیشنهادی با خلاصه‌سازهای مرسوم به کار برده می‌شوند.

فرآیند خلاصه‌سازی، با سه پارامتر دقت (P)، خوانایی (R) و تناسب کلی[۵](F) ارزیابی می‌شود. با توجه به این جدول، مشاهده می‌شود که دقت رویکرد پیشنهادی در خلاصه‌سازی مجموعه داده‌های Duc2001 و Duc2001، به ترتیب، برابر با ۸۹ و ۹۱ درصد می‌باشد، که بیشترین دقت در بین روش‌های معرفی شده در این جدول است. این مقایسه نشان می‌دهد که روش معرفی شده، که بر مبنای الگوریتم بهینه‌سازی ABC-ROC، طراحی شده است، بهترین نتایج را در مقایسه با روش‌های دیگر می‌دهد. این نتیجه به این خاطر است که روش معرفی شده می‌تواند ویژگی عمومی و محلی متن را به خوبی بهینه کند، در عین حال، همپوشانی را کاهش و حضور جملات مهم در متن را افزایش دهد.

جدول۵- مقایسه پارامترهای دقت (P)، خوانایی (R) و F1 با استفاده از روش‌های مختلف

Our approach			Copernic Summarizer			MS Word Summarizer			ابزار / متن
F1	R	P	F	R	P	F1	R	p	
۰,۷۵	۰,۶۲	۰,۸۱	۰,۷۳	۰,۶۸	۰,۴۸	۰,۲۸	۰,۴۳	۰,۳۵	Duc2001
۰,۷۲	۰,۶۵	۰,۸۲	۰,۷۵	۰,۷۴	۰,۵۲	۰,۲۲	۰,۴۷	۰,۳۷	Duc2001

[1] -Overall Fitness [5]

ISSN: 2745-3673
http://www.Rcsj.ir

نتیجه‌گیری

رویکرد معرفی شده برای خلاصه‌سازی متن، به صورت آزمایشگاهی، به کار برده شد و رویکرد ایده را برای خلاصه‌سازی دو مجموعه داده متنی بزرگ و کوچک به کار بردیم. در مرحله ابتدایی، متن ورودی بصورت کلمه به کلمه تفکیک شده، کلمات بی‌تأثیر از متن اصلی حذف و سپس مرز بین جملات مشخص می‌شود. امتیاز هر جمله با توجه به ویژگی‌های آن محاسبه می‌شود. سپس، نتیجه بدست آمده و تعداد کل جملات انتخاب شده به عنوان ورودی به الگوریتم بهینه‌سازی ABC-ROC داده می‌شود. مزیت اصلی رویکرد معرفی شده، استفاده از الگوریتم بهینه‌سازی زنبور عسل بهبود یافته و تعریف یک تابع هدف جدید بود، که نشان دادیم، استفاده از این‌ها می‌تواند دقت خلاصه‌سازی را به طور مؤثری بهبود بخشد.

ISSN: 2745-3673
http://www.Rcsj.ir

منابع و مراجع

[1] H. O. B, P. Blache, and O. Nouali, “and mRMR Discriminant Analysis for Multilingual Text Summarization.”
[2] M. M. Al-tahrawi and S. N. Al-khatib, “Arabic text classification using Polynomial Networks,” J. King Saud Univ. - Comput. Inf. Sci., vol. 27, no. 4, pp. 437–449, 2015.
[3] M. Joshi, H. Wang, and S. Mcclean, “Generating Object-Oriented Semantic Graph for Text Summarisation,” pp. 298–311, 2014.
[4] M. S. Binwahlan, N. Salim, and L. Suanmali, “Swarm Diversity Based Text Summarization,” pp. 216–225, 2009.
[5] P. Perera and L. Kosseim, “Compression for Text Summarisation,” pp. 126–139, 2013.
[6] S. Jones, G. W. Paynter, P. Bag, and N. Zealand, “Interactive Document Summarisation Using Automatically Keyphrases Extracted,” vol. 0, no. c, pp. 1–10, 2002.
[7] A. Sarker and D. Moll, “An Approach for Query-Focused Text Summarisation for Evidence Based Medicine,” pp. 295–304, 2013.
[8] O. Article, “Automatic text summarization based on latent semantic indexing,” pp. 25–29, 2010.
[9] M. E. Hannah, T. V Geetha, and S. Mukherjee, “Based on Fuzzy Logic : A Sentence Oriented Approach,” pp. 530–538.
[10] H. Moen, L. Peltonen, J. Heimonen, and A. Airola, “Artificial Intelligence in Medicine Comparison of automatic summarisation methods for clinical free text notes,” Artif. Intell. Med., vol. 67, pp. 25–37, 2016.
[11] S. Chakraborti, “Student Research Abstract : Multi-Document Text Summarization for Competitor Intelligence : A Methodology based on Topic Identification and Artificial Bee Colony Optimization,” pp. 1110–1111, 2015.
[12] T. Summarization, U. Ant, O. Algorithm, and O. F. Hassan, “TEXT SUMMARIZATION USING ANT COLONY قنسحملا لمنلا ة رمعتسم قيمزراوخ مادختساب صوصنلا صيخلت,” no. February, 2015.
[13] B. Nozohour-leilabady and B. Fazelabdolabadi, “On the application of Artificial Bee Colony (ABC) algorithm for optimization of well placements in fractured reservoirs; efficiency comparison with the Particle Swarm Optimization (PSO) methodology,” Petroleum, vol. 2, no. 1, pp. 79–89, 2015.
[14] S. Anuar, A. Selamat, and R. Sallehuddin, “A modified scout bee for artificial bee colony algorithm and its performance on optimization problems,” J. King Saud Univ. - Comput. Inf. Sci., 2016.
[15] Y. Tong, M. Liu, Y. Zhang, X. Liu, R. Huang, F. Song, R. D. Cannon, R. Calderone, H. Tomoda, S. Omura, and L. Zhang, “Beauvericin counteracted multi-drug resistant Candida albicans by blocking ABC transporters,” vol. 1, pp. 158–168, 2016.
[16] Y. Ledeneva, R. G. Hernández, and R. M. Soto, “EM Clustering Algorithm for Automatic Text,” no. February 2003, pp. 305–315.