بسم الله الرحمن الرحيم

دانشگاه اصفهان

دانشکده فنی و مهندسی

گروه مهندسی کامپیوتر

پایان‌نامه کارشناسی ارشد رشته‌ی مهندسی کامپیوتر گرایش نرم‌افزار

خلاصه‌سازی خودکار متون فارسی به روش معنایی

استادان راهنما:

دکتر محمد علی نعمت بخش

دکتر احمد رضا نقش نیلچی

پژوهشگر:

فراز محمدیان جدول قدم

شهریورماه 1391

دانشگاه اصفهان

دانشکده فنی و مهندسی

گروه کامپیوتر

پایان‌نامه‌ی کارشناسی ارشد مهندسی کامپیوتر گرایش نرم‌افزار

آقای فراز محمدیان جدول قدم تحت عنوان

ارائه یک روش خلاصه‌ساز خودکار متون فارسی به روش معنایی

در تاریخ 91/12/13 توسط هیأت داوران زیر بررسی و با درجه بسیار خوب به تصویب نهایی رسید.

1- استاد/ استادان راهنمای پایان نامه	دکتر محمدعلی نعمت بخش با مرتبه‌ی علمی دانشیار	امضا
2- استاد/ استادان راهنمای پایان نامه	دکتر احمد نقش نیلچی با مرتبه‌ی علمی دانشیار	امضا
3- استاد/ استادان داور داخل گروه	دکتر بهمن زمانی با مرتبه‌ی علمی استادیار	امضا
4- استاد/ استادان داور خارج از گروه	دکتر بتول علی نژاد با مرتبه‌ی علمی دانشیار	امضا

امضای مدیرگروه

با تشکر و قدردانی از

اساتید گرانقدر آقایان دکتر محمدعلی نعمت بخش و دکتر احمدرضا نقش نیلچی که اساتید راهنمای این پروژه بودند و مرا در تمام مراحل انجام این کار یاری رساندند.

تقدیم به

پدر و مادر و همسرم

چکیده:

خلاصه‌سازی خودکار متن یکی از جنبه‌های مهم در بازیابی اطلاعات می‌باشد. در این پایان‌نامه یک روش جدید برای خلاصه‌سازی خودکار متون فارسی پیشنهاد شده است که رویکردی مبتنی بر معنا دارد. روش پیشنهادی، دارای سه فاز پیش پردازش، پردازش و تولید خلاصه نهایی می‌باشد. در مرحله پیش پردازش، ریشه کلمات با استفاده از یک روش مبتنی بر فرهنگ واژگان بدست می‌آید. مجموعه ترادف جملات موجود در متن از طریق همین فرهنگ واژگان بدست می‌آید. این کار منجر به یک زنجیره لغوی قوی می‌شود. در مرحله پردازش، با استفاده از زنجیره لغوی و روش تحلیل معنای پنهان، ارتباط بین جملات مهم بدست می‌آید تا جملات مرتبط با هم گزینش و در خروجی قرار گیرند. در مرحله تولید خلاصه، برای رفع افزونگی و تکرار، جملات مشابه در خروجی با جملات انتخاب شده مقایسه شده و در صورت داشتن شباهت نسبی (با انتخاب ضریب میزان تشابه 80٪) از قرار دادن آن در خلاصه نهایی جلوگیری می‌شود.

دو شیوه برای ارزیابی خلاصه‌ساز پیشنهادی ارائه شده است. یکی روش ارزیابی مستقیم و دیگری روش ارزیابی مبتنی بر LSA (تحلیل معنایی پنهان) است. نتایج ارزیابی نشان می‌دهد به دلیل استفاده از فرهنگ واژگان غنی، ریشه‌یابی با دقت بالائی انجام می‌گیرد که در مقایسه با روش‌های مشابه کارایی بالائی دارد. به دلیل استفاده از همین فرهنگ واژگان غنی، مجموعه ترادف و زنجیره لغوی ارتباط معنای کلمات و جملات در روش تحلیل معنایی پنهان بسط داده شده به درستی بدست آورده می‌شود که نتایج حاکی از این امر دارد. در نهایت نتایج ارزیابی نشان می‌دهد که روش ارائه شده خلاصه قابل قبول و منسجمی ارائه می‌دهد.

واژگان کلیدی: خلاصه‌سازی، ریشه‌یابی، زنجیر لغوی، تحلیل معنایی پنهان، ارتباط معنایی.

فهرست مطالب

عنوان **صفحه**

عنوان **صفحه**

عنوان **صفحه**

فصل ششم: جمع‌بندی و کارهای آینده

فهرست شکل ها

عنوان **صفحه**

فهرست جداول

عنوان **صفحه**

فصل اول

طرح تحقیق

1-1- مقدمه

امروزه بیش از 80 درصد از دانش ما به صورت متن، مستندات[1] و دیگر صورت‌های رسانه‌ای نظیر ویدیو و صدا نگهداری می‌شود. اگر از منظر علوم کامپیوتری به این مستندات نگاه کنیم همه‌ی آنها به طبیعتی غیرساخت‌یافته وابسته‌اند. با رشد روز افزون منابع اطلاعاتی، پید کردن منابع مربوط به موضوع مورد نظر و گزینش[2] مطالب مفید از میان منابع مربوطه به یک معضل تبدیل شده است. این مشکل توسط سیستم‌های خلاصه‌سازی قابل حل است. از این‌رو پیدا کردن اطلاعات متنی مناسب و ارائه یک چکیده[3] (خلاصه) به یک چالش بزرگ تبدیل شده است [1].

1-2- خلاصه

خلاصه، در لغت به معنی خالص، برگزیده، منتخب و کوتاه‌شدن مطلب است. خلاصه، یک ارائه دقیق از محتوای یک متن است. برای رسیدن به این منظور در زیر تعریفی از خلاصه ارائه شده است: «کاهش و تغییرات متن اصلی به خلاصه متن، از طریق انتخاب یا عمومیت دادن به آنچه که در متن است.» استفاده از خلاصه، زمان خواندن متن را متناسب با اطلاعات مورد نیاز تا حد زیادی کاهش می‌دهد. یک فرد برای دریافت دانش از اطلاعات یک متن، بایستی ابتدا آنرا درک کند، سپس آن را پردازش کرده تا بفهمد چه معانی و مفاهیمی در متن موجود است و از میان این مفاهیم کدام جدید و کدام قدیمی است. بعد از درک مفاهیم، خلاصه‌ای متناسب با

1. Document’s
2. Election
3. Abstract

نیاز خود تولید نماید. این کار سخت و طاقت فرسا می‌باشد و از طرفی چون قضاوت انسانی بی‌ثبات است و ماهیت افکار آن‌ها در محیط‌های مختلف قابل تغییر است، خلاصه‌های مختلفی را می‌توان از افراد داشته باشیم. بنابراین اگر بتوان خلاصه را توسط ماشین و بصورت خودکار تولید نمود این مشکل قابل حل است. این رهیافت مستلزم استفاده از خلاصه‌سازی خودکار متن است [3، 5].

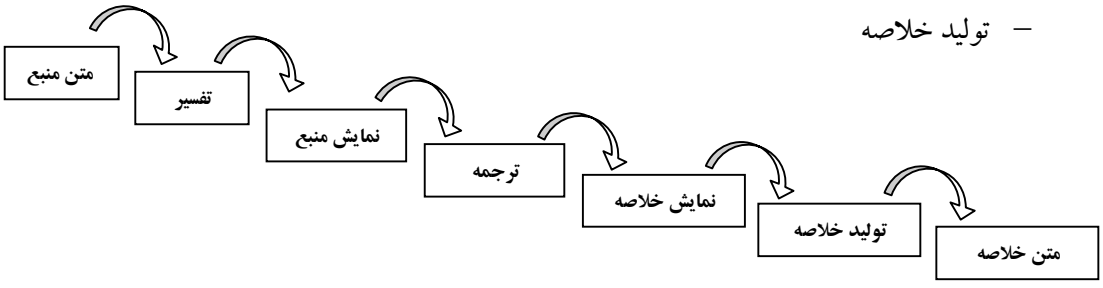
خلاصه‌سازی خودکار متن[1] عبارت است از: کوتاه‌کردن متن از طریق انتخاب جملات مهم، به طوری که متن نهایی مهم‌ترین نکات متن اصلی را نیز در برداشته باشد. خلاصه‌سازی خودکار متن، تکنیکی است که از طریق یک برنامه کامپیوتری متنی را خلاصه می‌کند و در اختیار درخواست کننده قرار می‌دهد. این برنامه یک متن از ورودی دریافت کرده و یک خلاصه از متن اصلی به خروجی می‌فرستد. بنابراین در خلاصه‌سازی خودکار، اطلاعات از منبع اصلی استخراج، آنگاه محتوای آن بررسی شده و محتوای مهم برای کاربر به شکل متراکم و به هم پیوسته ارائه می‌شود [۲،۶،17].

از جمله مزیت‌های خلاصه سازی خودکار متن نسبت به خلاصه‌سازی دستی می‌توان به موارد زیر اشاره کرد:

- ارتباط بین موجودیت‌های متن، در خلاصه و موقعیت آن‌ها در متن اصلی به آسانی قابل ایجاد است.
- اندازه خلاصه و محتوای آن قابل کنترل و معین است.

فرآیند تبدیل منبع به خلاصه در سه مرحله ذیل انجام می‌پذیرد که در شکل 1-1 نشان داده شده اند:

- تجزیه و تحلیل[2] منبع
- انتخاب محتوا
- تولید خلاصه



شکل 1-1- فرآیند خلاصه‌سازی

1. Automatic summarization
2. Analysis
3. Preprocessing

فرایند تجزیه و تحلیل منبع، که به آن پیش‌پردازندش[1] هم گفته می‌شود خصیصه‌هائی که به یک خلاصه منتج می‌شود را تعیین کرده و یک نمایش ساخت‌یافته از متن اصلی به دست می‌آورد. فرایند انتخاب محتوا یا پردازش الگوریتمی، اطلاعات متن را براساس خصیصه‌هائی که در مرحله قبل بدست آمده است به ساختار خلاصه تبدیل می‌کند، نهایتاً در فرایند تولید خلاصه، خلاصه نهایی با استفاده از ساختار خلاصه بدست می‌آید [۳،۸،19].

روش‌های خلاصه‌سازی می‌تواند براساس سطح زبان به دو گروه عمده تقسیم شود:

1. دیدگاه سطحی که محدود به نمایش قواعد نحوی‌است و سعی در استخراج قسمت‌های برجسته متن به روش مناسبی دارد.
2. دیدگاه عمیق‌تر که شامل سطح معنایی از متن اصلی و فرایندهای زبانی است.

در دیدگاه اول هدف از پیش‌پردازش متن، کاهش بعد فضایی آن است و به صورت طبیعی شامل موارد زیر می‌باشد:

الف) حذف حروف اضافه[2] (لغات ارتباط دهنده رایج بین کلمات بدون هیچ معنایی خاص مثل a و The)

ب) نادیده گرفتن حروف بزرگ و کوچک و تبدیل تمامی کاراکترهای متن به فرم خاص از حروف (حروف بزرگ یا کوچک)

ج) تبدیل کلمات به ریشه[3] آنها، کلماتی با قواعد نحوی مشابه مثل کلمات جمع، حالت های مختلف فعل و غیره به صورت یکسان در نظر گرفته می‌شوند. هدف از این مرحله به دست آوردن ریشه کلمه است که در پردازش معنایی استفاده می‌شود [۵،11].

3-1-انواع خلاصه‌سازی

خلاصه را می‌توان به دو دسته استخراجی[4] و غیراستخراجی (چکیده[5]) تقسیم‌کرد. درخلاصه‌ی استخراجی جمله‌ها از متن اصلی گزینش می‌شوند و در خلاصه کپی می‌گردند، به عنوان مثال عبارات، جملات یا پارگراف‌های کلیدی در متن اصلی عیناً در خلاصه کپی می‌شوند. در این روش خلاصه نهایی انسجام و یکپارچگی را بین پاراگراف‌ها وجملات تضمین نمی‌کند ولی نمایش تقریبی از محتوای متن را به نمایش می‌گذارد. خلاصه غیر استخراجی مطالب مفید متن اصلی را، انتخاب می‌کند و آنها را تفسیر می‌نماید و در

2. Stop word
3. Stem
4. Extractive
5. Abstracting

خلاصه بازنویسی می‌کند. به طورکلی خلاصه غیراستخراجی قوی‌تر از خلاصه استخراجی می‌تواند متن را خلاصه کند، برنامه‌هایی که این نوع خلاصه را تولید می‌کنند، نیاز به بکارگیری تکنولوژی تولید زبان طبیعی[1] دارند که خود بحث پیچیده‌ای است [17،۲،۱].

خلاصه نیز می‌تواند انواع دیگری داشته باشد که می‌توان به موارد زیر اشاره کرد [2،۱]:

- بر حسب تعداد اسناد ورودی: تک سندی یا چند سندی[2] می‌تواند باشد.
- برحسب تعامل با کاربر: مبتنی بر پرس وجو یا غیر مبتنی بر پرس وجو می‌تواند باشد.
- برحسب عمومیت: عمومی یا محدود به حوزه می‌تواند باشد.

خلاصه‌سازی خودکار متن دارای ضعف‌ها و مشکلاتی است. در زیر این چالش‌ها را دسته‌بندی کرده و مورد بررسی قرار خواهیم داد:

- خلاصه‌سازی درگیر مکانیزم تهیه خروجی موثر و کارآمد است. زیرا ایجاد یک خلاصه یکپارچه و منسجم که عبارات آن دارای یک رابطه منطقی با یکدیگر باشند بسیار حائز اهمیت می‌باشد.

- ارزیابی و قضاوت کردن در مورد خلاصه تولیدشده را می‌توان از دیگر چالش‌های خلاصه‌سازی خودکار نام برد. زیرا رهیافت‌های بکار رفته در ارزیابی این سیستم‌ها مانند قضاوت انسانی بی‌ثبات است. در این رهیافت نتایج وابستگی زیادی به افراد دارد و دیدگاه افراد با هم متفاوت است.

- خلاصه‌سازی درگیر انسجام و تراکم[3] بین عبارات موجود در خلاصه است، زیرا توانایی ارزیابی خلاصه‌ها با نرخ تراکم متفاوت، مقیاس و پیچیدگی ارزیابی را افزایش می‌دهد.

- از سوی دیگر در خلاصه‌سازی غیر استخراجی، خلاصه‌ی تولید شده ممکن است بیرون از محتوای متن باشد، به گونه ای که تفسیری که خلاصه‌ساز از متن اصلی داشته است اشتباه بوده و هیچ‌گونه ارتباطی از لحاظ محتوا با متن اصلی نداشته باشد. در این حالت مسئله‌ی وابستگی یا انسجام[4] خلاصه تولید شده، نسبت به منبع اصلی بوجود می‌آید [2].

نحوه نمایش خلاصه توسط خلاصه‌ساز از دیگر چالش‌های خلاصه‌سازی خودکار متون است، زیرا باید نحوه نمایش به گونه‌ای پاسخگوی نیازهای افراد در زمینه‌ها و حوزه‌های مختلف باشد.

---

1. Natural language generation
2.Multi document
3. Compression
4. Coherence

فصل دوم
خلاصه‌سازی و خلاصه‌سازی خودکار

2-1- مقدمه

در این فصل، ابتدا مفاهیم و تعاریف اولیه مربوط به مبحث بازیابی اطلاعات و خلاصه‌سازی متون مطرح می‌شود. سپس انواع منابع و روش‌های پیشین و مشکلات این روش‌ها را مورد بررسی قرار می‌دهیم.

2-2- خلاصه‌سازی

خلاصه در لغت یعنی خاص، برگزیده، منتخب، کوتاه‌کردن مطلب و ... مصدر آن معنی خلاصه‌کردن، مختصرکردن کلام و روشن ساختن می‌باشد. خلاصه در مقابل واژه چکیده نیز آمده است و چکیده یعنی نوشته‌ای از مهمترین مطالب یک کتاب و یا یک مقاله و متن گزارش که بصورت خلاصه تهیه شده است (شکل 2-1). منظور از خلاصه نمودن در اینجا بیان اصل مطلب در کوتاه‌ترین شکل می‌باشد، بطوریکه به هیچ وجه به اصل مطلب خدشه‌ای وارد نیاید.

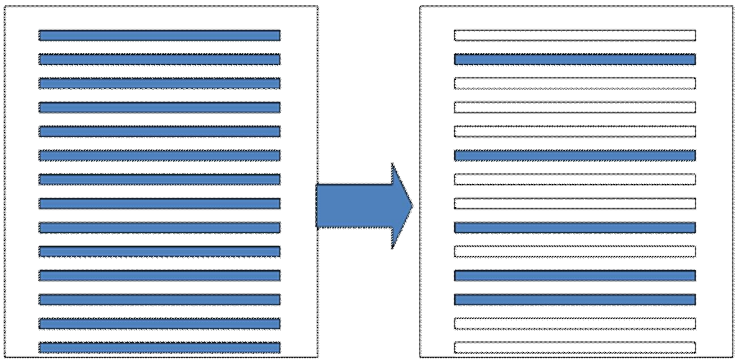
قبل از شرح بیشتر خلاصه‌سازی جایز دانستیم که ابتدا جایگاه خلاصه‌سازی را شرح دهیم و در خلال تعاریف اولیه و مشکلات خلاصه‌ساز را بیاوریم. اکنون بیش از 80 درصد از دانش ما به صورت متن، مستندات و دیگر صورت‌های رسانه نظیر ویدیو و صدا نگهداری می‌شود. اگر از منظر علوم کامپیوتری به این مستندات نگاه کنیم همه‌ی آن‌ها به طبیعتی غیر ساخت یافته وابسته‌اند. یک فرد برای دریافت دانش از اطلاعات یک متن، بایستی ابتدا آنرا درک کند و سپس آنرا پردازش نماید تا بفهمد چه معانی و مفاهیمی در آن موجود است، چه ارتباطی میان مفاهیم وجود دارد و از میان این مفاهیم کدام مهم و مفیداست. چون قضاوت انسان بی‌ثبات است و تحت تأثیر عوامل محیطی، سلیقه فرد و دانش فرد مورد مطالعه می‌باشد، خلاصه‌های مختلفی از افراد مختلف می‌توان داشت. از مهمترین مشکلات مربوط به خلاصه‌سازی توسط انسان این است که هرچه تعداد مستندات بیشتر شود زمان

بیشتری را باید صرف مطالعه و خلاصه‌سازی کرد و از همه مهمتر سردرگمی کاربران در گزینش مطالب را می‌توان اشاره کرد.

2-3- خلاصه‌سازی خودکار

با گسترش روزافزون حجم اطلاعات موجود در وب و افزایش چشم‌گیر مقالات منتشر شده در زمینه‌های مختلف علمی، دسترسی درست و مطالعه مورد نیاز، همواره یکی از مشکلات محققان و پژوهشگران قرن 21 می‌باشد. اینکه چگونه با این حجم می‌توان در روز چندین کتاب را مطالعه نمود. آیا می‌توان سیستمی طراحی نمود که بتواند با داده‌های موجود به تمامی سوالات ما پاسخ دهد، اینها سوالاتی است که پاسخ آنها را می‌توان در یک سیستم خلاصه‌ساز متن جستجو کرد.

خلاصه‌سازی خودکار سند، یعنی تولید یک نسخه مختصرتر از سند اصلی توسط یک برنامه کامپیوتری به نحوی که ویژگی‌ها و نکات اصلی سند اولیه حفظ شود [1]. بنابر تعریف ارائه شده در استاندارد ISO 215 [2] سال 1986، خلاصه "یک بازگویی مختصر از سند" می‌باشد (شکل 2-2).



شکل 2-2- خلاصه‌سازی خودکار سند

همانطور که اشاره شد خلاصه‌سازی خودکار توسط کامپیوتر انجام می‌شود و به همین دلیل تفاوت‌های زیادی با خلاصه‌سازی که توسط انسان تولید می‌شود دارد. انسان‌ها با توجه به هوش و شعور ذاتی خود قادر به درک و فهم مفاهیم موجود در متن و ارتباط بین آنها می‌باشند و این در حالی است که انجام این عملیات توسط ماشین کار بسیار دشوار و پیچیده‌ای می‌باشد. از طرفی دیگر، انسان‌ها با توجه به سطح دانش و پی زمینه‌ی اطلاعاتی که دارند دید متفاوتی از خلاصه‌ی یک متن یکسان دارند. به عنوان مثال کسی که سال‌ها در زمینه شبکه‌های کامپیوتری به تحقیق و مطالعه پرداخته است با کسی که به تازگی قصد تحقیق و مطالعه در زمینه‌ی شبکه‌های کامپیوتری را دارد، متفاوت بوده و خلاصه‌ای که این دو فرد از یک متن در زمینه‌ی شبکه‌های کامپیوتری تولید می‌کنند قطعاً یکسان نخواهد بود.

خلاصه‌سازی خودکار متن خود مشکلات و سختی‌های زیادی دارد اما مزیت‌های نسبت به خلاصه‌سازی دستی دارد که می‌توان به موارد زیر اشاره کرد:

1- ارتباط بین موجودیت‌های متن در خلاصه و موقعیت آنها در متن اصلی به آسانی قابل ایجاد است.

2- اندازه خلاصه قابل کنترل است.

3- محتوای آن قابل پیش‌بینی است.

این مطلب در چند سال اخیر بسیار مورد توجه واقع شده و منجر به ارائه بحث خلاصه‌سازی مبتنی بر کاربر یا خلاصه‌سازی شخصی‌سازی شده، گشته است. در ادامه به این مطلب بیشتر خواهیم پرداخت.

سیستم‌های خلاصه‌ساز در دنیای امروز کاربردهای فراوانی دارند. تولید خلاصه‌های کتب مختلف و مقالات علمی، تولید خلاصه اخبار و انتقال آن از طریق سیستم‌های نظیر تلفن همراه، نمایش خلاصه سند یافته شده توسط موتور جستجو، تولید سیستم‌های پاسخ‌گویی به سوال و ... همگی از کاربردهای این سیستم می‌باشند.

2-4- فرایند خلاصه‌سازی

خلاصه‌سازی یکی از کاربردهای پردازش متن است. پردازش متن شامل چهار سطح است، پردازش لغوی، پردازش ساخت واژی، پردازش نحوی و پردازش معنایی [1]. هر یک از این کاربردهای فراوان پردازش متن، از جمله بازیابی اطلاعات، خلاصه‌سازی، درک، تولید، ترجمه، پرسش و پاسخ، استخراج دانش از متون و موارد دیگر با توجه به گستردگی و پیچیدگی، در یک یا چند سطح فوق به انجام می‌رسد. خلاصه‌سازی، یکی از پیچیده‌ترین کاربردهای پردازش متن است و معمولاً با چند سطح از پردازش متن درگیر است. در ادامه به پردازشهای لغوی و ساخت واژی خواهیم پرداخت و پردازش‌های نحوی و پردازشهای معنایی را به فراخور نیاز بعداً آنها را شرح خواهیم داد.

2-4-1- پردازش لغوی

منظور از پردازش لغوی شناسایی مرز لغات و جملات در یک متن است. این مرز ممکن است به شکل ساده توسط جداکننده‌هایی مانند فاصله، کاما، نقطه، علامت سئوال و ... تعیین شود و یا نیاز به پردازش‌های بیشتری داشته باشد، مانند زمانی که میان بخش‌هایی یک کلمه از فاصله استفاده می‌کنیم (مثل کلمه«می‌توان») و یا وقتی دو کلمه مجزا را بدون فاصله یا پی در پی می‌نویسیم (مثل عبارت در «برابر باد»). تعیین مرز کلمات در زبان فارسی به دلیل گوناگونی رسم الخط و عدم وجود استانداردهای نگارشی و همچنین به دلیل وجود شکل‌های مختلف حروف (اول- وسط- آخر چسبان و غیرچسبان) بیش از زبان انگلیسی مشکل‌ساز است. این مشکل در