۱-۵- اهداف پایان‌نامه

به طور خلاصه اهداف پایان‌نامه را می‌توان به صورت زیر در نظر گرفت:

- مطالعه الگوریتم‌های خلاصه سازی متن
- مطالعه ماشین تورینگ[۱] و کاربردهای آن
- مطالعه اتوماتای یادگیر سلولی[۲] و کاربردهای آن
- مطالعه الگوریتم زنبورعسل مصنوعی[۳] و کاربردهای آن
- مطالعه الگوریتم ژنتیک و کاربردهای آن
- مطالعه الگوریتم بهینه سازی توده ذرات[۴] برای سیستم امتیازدهی به ویژگی های استخراجی متن
- مطالعه سیستم های استنتاج فازی و کاربردهای آن
- ارائه یک روش جدید برای پیش پردازش متن با ماشین تورینگ
- ارائه یک روش جدید برای تعیین شباهت با استفاده از اتوماتای یادگیر سلولی، الگوریتم زنبورعسل مصنوعی و فرمول شباهت
- ارائه یک روش جدید خلاصه سازی مبتنی بر استخراج ویژگی های مهم متن
- ارائه یک روش جدید خلاصه سازی مبتنی بر استخراج ویژگی های مهم متن، وزدن دهی براساس الگوریتم بهینه سازی توده ذرات
- ارائه یک روش جدید خلاصه سازی مبتنی بر استخراج ویژگی های مهم متن، وزدن دهی براساس الگوریتم بهینه سازی توده ذرات و الگوریتم ژنتیک
- ارائه یک روش جدید خلاصه سازی مبتنی بر استخراج ویژگی های مهم متن، وزدن دهی براساس الگوریتم بهینه سازی توده ذرات و الگوریتم ژنتیک
- ارائه یک روش جدید خلاصه سازی ترکیبی مبتنی بر ٤ روش خلاصه سازی مطرح شده
- مقایسه و ارزیابی روش‌های پیشنهادی با روش‌های دیگر خلاصه سازی

---

[۱] Turing Machine (TM)

[۲] Cellular Learning Automata (CLA)

[۳] Artificial Bee colony (ABC)

[۴] Particle Swarm Optimization (PSO)

۱-٦- ساختار پایان‌نامه

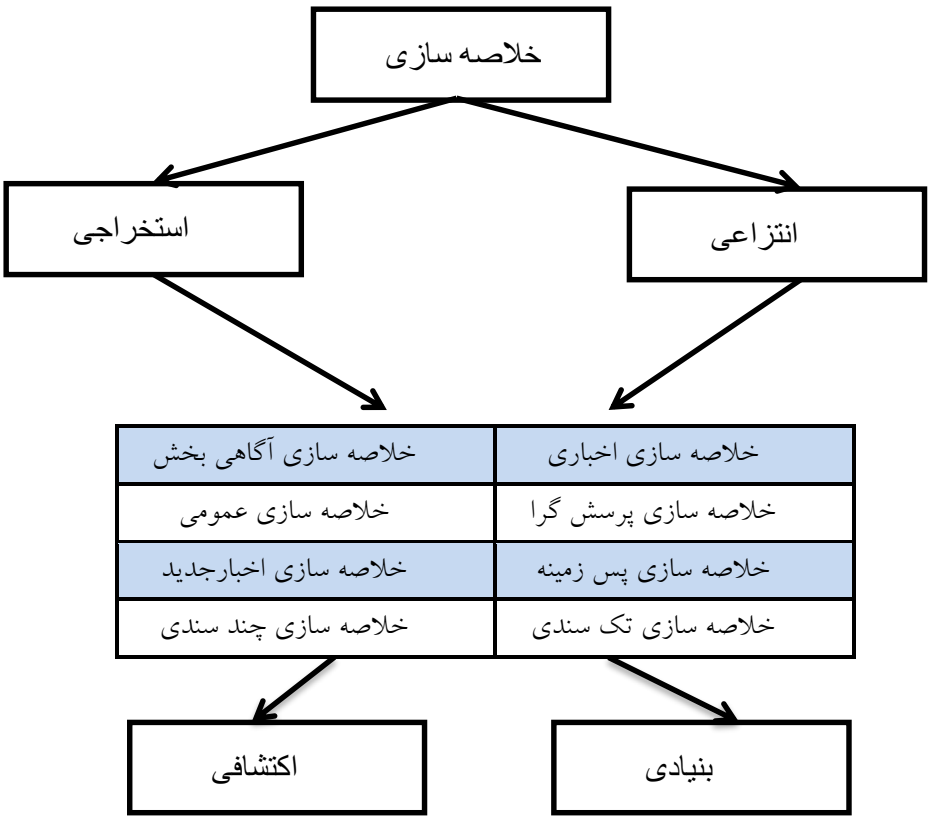
مطالبی که در پایان نامه حاضر بدان خواهیم پرداخت بدین صورت خواهد بود که در فصل دوم این پایان نامه موضوع خلاصه سازی متن بصورت کلی مورد بررسی قرار می گیرد و انواع رویکردها و تاریخچه ای ازآن ها بیان می شود. فصل سوم مفاهیم پایه‌ای بنام ماشین تورینگ، اتوماتای یادگیر سلولی ، الگوریتم بهینه سازی توده ذرات، الگوریتم زنبورعسل مصنوعی، الگوریتم ژنتیک و سیستم های فازی که نقش کلیدی در پایان‌نامه را دارند، مورد بررسی قرارخواهند گرفت. در فصل چهارم، یک سیستم پیش پردازش پیشنهادی با ماشین تورینگ، یک سیستم تعیین شباهت پیشنهادی با الگوریتم اتوماتای سلولی یادگیر و الگوریتم زنبورعسل مصنوعی وهم چنین ۵ روش خلاصه ساز پیشنهادی به صورت سیستم خلاصه سازی مبتنی بر استخراج ویژگی های متنی، سیستم خلاصه سازی مبتنی براستخراج ویژگی های متنی و الگوریتم توده ذرات، سیستم خلاصه سازی مبتنی براستخراج ویژگی متنی، الگوریتم توده ذرات، الگوریتم ژنتیک، سیستم خلاصه سازی مبتنی براستخراج ویژگی متنی، الگوریتم توده ذرات، سیستم های فازی و سیستم خلاصه سازی ترکیبی مبتنی٤ روش خلاصه سازی پیشنهادی معرفی می شوند. در فصل پنجم نحوه ارزیابی و پیاده‌سازی انجام گرفته، همچنین کارایی خلاصه سازهای پیشنهادی در مقایسه با یکدیگر و هم با چنین با خلاصه سازهای پایه H2-H1، Msword، Sys19، Sys30و خلاصه ساز های دیگر بررسی می‌شود. در پایان به نتیجه‌گیری و بیان پیشنهادات می‌پردازیم.

فصل دوم:

متدهاوروش‌های خلاصه سازی متون

۲-۱- مقدمه

خلاصه سازی یک متن را می توان ازجنبه های مختلف بررسی کـرد و بـراین اسـاس خلاصـه سـازی دارای انواع گوناگونی می باشد. نمایی از انواع خلاصه سازی درشکل۱-۲ معرفی شده است. حال به معرفی هریک از انها می پردازیم.



شکل۱-۲: تقسیم بندی انواع خلاصه سازی

۲-۲- تقسیم بندی انواع خلاصه سازی

۲-۲-۱- خلاص سازی استخراجی دربرابر انتزاعی

خلاصه سازی استخراجی، خلاصه‌ای است که عموماً توسط انتخاب بخش‌های مرتبط با موضوع متن بدست می‌آید. خلاصه سازی انتزاعی تفسیری است که محتوای یک مستند را بدون آن که لزوماً از محتوای آن استفاده کرده باشد، شرح می‌دهد. در روش استخراجی، جملات خلاصه معمولاً از حد متوسط طولانی‌تر هستند. اطلاعات مهم معمولاً در طول جملات گسترش یافته‌اند اما جملات استخراج شده این موضوع را در بر نمی‌گیرد مگر اینکه خلاصه به قدر کافی بزرگ باشد که تمام آن جملات را نگه دارد. در این روش اطلاعات ناسازگار ممکن است به دقت نشان داده نشود. با این حال کاربران خلاصه‌های استخراجی را به چکیده ترجیح می‌دهند. زیرا خلاصه‌های استخراجی، اطلاعات را همان‌طور که نویسنده بیان کرده نشان می‌دهند و پردازش آن‌ها ارزان‌تر است. در روش انتزاعی، خلاصه بوسیله ساختن دوباره عبارات و تشکیل مجموعه ای پیوسته که حاوی محتوای مهم متن اصلی است، تهیه می گردد. واضح است که در این روش باید ازانواع روش های گرامری برای ساخت مجدد جمله استفاده نمود که پیچیدگی کار را بالا می برد. به همین دلیل این روش امروزه کمتر مورد توجه می باشد.

۲-۲-۲- خلاصه سازی اخباری دربرابرآگاهی بخش

درخلاصه سازی اخباری، موضوعات اصلی و مهم از متن اصلی استخراج می شود که فرد با خواندن این نوع خلاصه، ازرئوس اصلی مطالب متن اصلی مطلع می شود. ولی برای آگاهی کامل درصورت لزوم باید به اصل متن مراجعه کند. خلاصه سازی آگاهی بخش، خلاصه ای کلی بوده است که سعی دارد مطالب مهم را به همراه توضیح کاملشان استخراج نماید به نوعی که نمایانگر کل متن مطلب اصلی می باشد و تنها به بیان رئوس مطالب مهم اکتفا نمی کند.

۲-۲-۳- خلاصه سازی پس زمینه[۱]دربرابر اخبار جدید[۲]

این دو خلاصه سازی با فرض این که خواننده دانش پیشین درمورد مطلب موردنظربرای خلاصه را ندارد در مقابل کسی که اطلاعاتش بروز است، می باشد.

---

[۱] Background

[۲] Just-the-news

۲-۲-٤- خلاصه سازی عمومی دربرابر پرسش گرا

خلاصه‌سازی عمومی در واقع حس کلی محتوای متن را می‌دهد در حالی که خلاصه‌های مرتبط با پرس و جو خود را به محتوایی محدود می‌کنند که با سوال کاربر در ارتباط است. نوع دوم به طور وسیع در ارتباط با مستنداتی که بزرگ هستند و از لحاظ موضوع متفاوتند بسیار کار آمد تر است.

۲-۲-۵- خلاصه سازی تک سندی[۱]دربرابرچند سندی[۲]

درحالت تک سندی خلاصه تهیه شده ازروی یک متن بوده و درخلاصه چند سندی خلاصه تولیدی بوسیله ترکیب چند متن و استخراج خلاصه می باشد.

۲-۲-٦- خلاصه سازی براساس دیدگاه های اکتشافی[۳]و بنیادی[٤]

به طورکلی متدها و روش های خلاصه سازی را می توان به دودیدگاه کلی تقسیم بندی نمود: دیدگاه های اکتشافی و دیدگاه های بنیادی. تکنیک هایی که براساس نظریات شناخته شده ریاضی عمل می کنند، دردسته دیدگاه های بنیادی تقسیم بندی می شوند. تکنیک هایی که ازتجربیات انسان استفاده می کنند دردسته دیدگاه های اکتشافی قرار می گیرند، ازاین دسته به تحلیل های تجربی زبان طبیعی و تحلیل نحوه نوشتار انسان اشاره کرد. دراکثر مواقع دودیدگاه برای دستیابی به خلاصه سازی بهتر به کمک یکدیگر می آیند. درادامه برخی از این روش هایی که برپایه این دو دیدگاه می باشند را توضیح می دهیم.

۲-۲-۳- انواع روش های خلاصه سازی

۲-۲-۳-۱- روش برپایه موقعیت[۵]

این روش برای اولین بار توسط Edmundson در سال ۱۹٦۹ اعمال گردید و در دسته تکنیک های اکتشافی گنجانده شده است. ایده این روش بدین صورت است که جملات مهم متن در ابتدا یا انتها قرار دارند. تجربه نشان داده که در ۸۵ درصد موارد جملات مهم در مکان های ابتدایی و در۷ درصد موارد در مکان های پایانی قرار دارند. همچنین تنها ۱۳ درصد از پاراگراف های نویسنده های معاصر با جملات مهم شروع می شوند(Baxendale,1958). نقطه ضعف این روش این است که محل جملات مهم درپاراگراف

[۱] Single Document
[۲] Multi Document
[۳] Heuristic approach
[٤] Foundation-based approach
[۵] Position-Based Method

می تواند متناسب با موضوع متون متفاوت باشد. به عبارتی می توان موقعیت جمله مهم را درمتن با استفاده از یک فرآیند یادگیری بدست آورد. نقطه ضعف دیگر این روش این است که برای کلیه پاراگراف های متن ارزش یکسانی در نظر می گیرد. به عبارتی ممکن است پاراگراف های ابتدایی یا انتهایی از اهمیت بیشتری برخوردار باشند.

۲-۲-۳-۲- روش سیاست بهینه مکانی [۱]

این روش اولین بار توسط Lin و Hovey در سال ۱۹۹۷ پیشنهاد گردید (Lin and Hovy, ۱۹۹۷). این روش در دسته تکنیک های اکتشافی گنجانده شده است. ایده این روش بدین صورت است که جملات مهم در متون مختلف با نوع موضوع در مکان مختلفی واقع می شوندو با استفاده از یک الگوریتم یادگیری این مکان به صورت خودکار تعیین می گردد. دراین روش برای هر متن همپوشانی بین هر یک از جملات متن با کلمات کلیدی و جملات چکیده را تعیین می نماییم. براساس این همپوشانی و شباهت یک امتیاز به جمله مذکور تعلق می گیرد. در قدم بعدی با دو مسئله روبرو هستیم :

۱. برای هر پارگراف مجموع میانگین امتیازات جملات انرا حساب می کنیم.بدین ترتیب می توان پار اگرفی که دارای بیشترین اهمیت است را استخراج نمود.
۲. برای هرموقعیت جمله در پاراگراف ، برای مثال جملات اول پاراگراف های امتیاز را حساب می کنیم.بنابراین مشخص می شودکه به طور متوسط در پاراگراف های مختلف چندیم جمله از بیشترین امتیاز برخورداراست.

بنابر دو مسئله گفته شده در بالا، جمله مهم قابل استخراج بوده یا به عبارتی جمله ای که در شرایط بالا صدق می کند از امتیاز بیشتری برخوردار است.

مجموعه داده مورد استفاده دراین روش مجموعه داده ZIFF بوده که برای آموزش ۱۳۰۰۰ مقاله روزنامه در مورد تازه های کامپیوتر و سخت افزار و غیره می باشد. نوع و دسته بندی هر یک از این مقالات نیز در این مجموعه داده مشخص گردیده است. هر مقاله بطور متوسط شامل ۷۱ جمله بوده و همراه با هر مقاله ۳ تا کلمه کلیدی و یک خلاصه که شامل ٦ جمله تولید شده بوسیله انسان می باشد. نتایج بدست امده برای ۲۹۰۰ مقاله از این مجموعه مطابق با جدول۱-۲ بوده است.

جدول ۱-۲: نتایج بدست آمده اززروش سیاست بهینه مکانی

دقت	فراخوانی

[۱] Optimum Position Policy

۳۸٪	۳۵٪

این روش درمقایسه با روش برپایه موقعیت بهتر عمل کرده است زیـرا ایـن روش بـا توجـه بـه اینکـه موقعیت مناسب جملات و پاراگراف را به صورت داینامیکی و متناسب با نوع متن بدسـت مـی آوردنـد از کارایی بهتری برخوردارست.دردروش برپایه موقعیت، فرض براین بود که جملات مهم درپاراگراف اول می باشد ولی این فرض درحالت کلی درست و کارا نمی باشد.

۲-۲-۳-۳- روش برپایه عنوان

این روش نیز برای اولین بار توسط Edmundson درسال ۱۹٦۹ اعمال گردید و دردسته تکنیک هـای اکتشافی گنجانده شده است. ایده این روش بدین صورت است که کلمات واقع شده در عنوان های کلـی و عنوان های بخش های جزئی تر متن ارتباط مستقیمی با خلاصه تولید شده دارند. تجربه نشان داده کـه ۹۹٪ این جملات به صورت آماری از اهمیت برخوردار هستند و در خلاصه کردن مفید واقع می شوند.

۲-۲-۳-٤- روش برپایه عبارات خاص[۱]

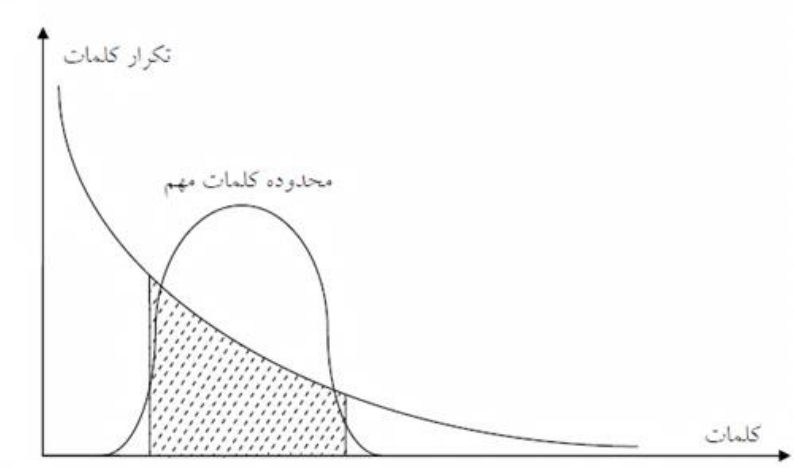
این روش در دسته تکنیک های اکتشافی گنجانده می شود، نیز برای اولین بـار توسـط Edmundson در سال ۱۹٦۹ اعمال گردید. جملات مهم متن اغلب شامل عبـارات امتیـازی ماننـد Significantly، In this paper، We show، In conclusion و غیره می باشند. در مقابل این عبارت ، عبارات بـا امتیـاز منفی مانند Hardly یا impossible می باشد. این عبارات نیز خود می توانند به صورت خودکار از مـتن استخراج شوند.نحوه کارچنین است که با موجه شدن به یک عبارت با امتیازی یک امتیاز مثبت و بـا موجـه شدن با یک عبارت با امتیاز منفی یک پنالتی به جمله مورد نظر اضافه می کنیم و در نهایت جمله بـا امتیـاز بیشتر را استخراج می کنیم.

نقطه ضعف این روش دراین است که عبارات خاص باید از پیش تعیین شوندکه این خود متوجه هزینـه و در ضمن اینکه عبارات خاص در متون مختلف متفاوت هستند که برای حل این مشکل Tuefel درسـال ۹۸ روشی برای استخراج خودکار این عبارات ارائه نموده است.

[۱] Cue-Based Method

۲-۲-۳-۵- روش برپایه بسامد لغوی [۱]

این روش که دردسته تکنیک های اکتشافی گنجانده می شود و بوسیله Luhn درسال ۱۹۵۸ معرفی گردید و بدین صورت بود که جملات مهم حاوی کلماتی هتند که دارای بسامد تکرار متوسطی می باشند. یعنی نه تکرار آنها بسیارزیاد باشد مانند حروف تعریف یا ضمائر و... یا بسیار نادر باشند. بدین منظور جملات را به ازای هر کلمه با ویژگی گفته شده در آن برامتیازش می افزاییم. نمودار بسامد تکرار کلمات در شکل ۲-۲ آورده شده است.



شکل ۲- ۱. نمودار بسامد تکرار کلمات

دراین روش برای شمارش کلمات مشابه از ریشه یابی [۲] نیز استفاده می شود. منظور از ریشه یابی این است که برای مثال دو کلمه system و systems مشابه هستند. در این روش همچنین از یکسری کلمات مانند the ، a ، is و... صرفنظر می شود و Luhn از آنها با نام لغات توقف [۳] یاد کرده است. درقدم بعد باید tf هرکلمه در متن محاسبه گردد.منظور از tf محاسبه توزیع احتمالیوقوع یک کلمه در متن هدف می باشد.همچنین باید idf هر کلمه نیز محاسبه شود. مقدار idf یک عبارت از معادله۱-۲ محاسبه می گردد:

$$idf\,(term) = \log(\frac{NUMDOC}{NUMDOC(term)}) \qquad (۱-۲)$$

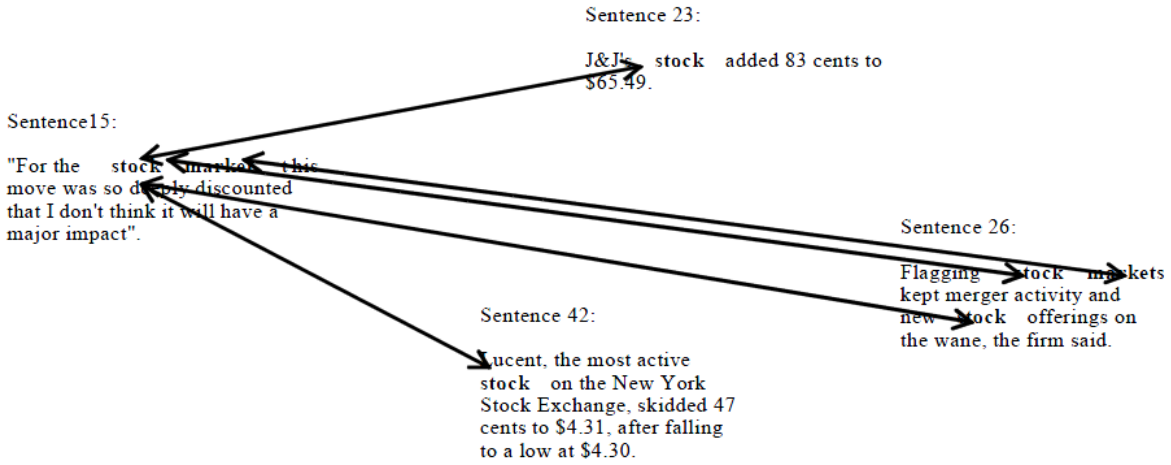
که درآن NOMDOC تعدا متن های موجود در پیکره بوده و NUMDOX(term) تعداد متن هایی است که عبارت term در انها رخ داده است. در نهایت الگوریتم بدین صورت است که عبارتی

[۱] Word-Frequency-Based Method
[۲] Stem
[۳] Stop words

انتخاب می گردد که مقدار $tf \times idf$ آن بیشتر از یک حد آستانه باشد. بنابراین می توان وزن اختصاصی به یک جمله را از مجموع $tf \times idf$ های لغات ان و نرمال کردن آنها بدست آورد.

۲-۲-۳-٦- روش های برپایه پیوستگی[۱]

در ابتدا بوسیله Bembrahim و Ahmad درسال ۱۹۹۵ معرفی گردید و این روش در دسته تکنیک های بنیادی گنجانده شده است. ایده این روش ها بدین صورت است که جملات و عبارات مهم دارای بیشترین اتصال درگراف شباهت خود هستند. این گرف خود می تواندبراساس تکرار یک کلمه یا شباهت معنایی یا هم مرجعی و غیره باشد(Ponte and Croft, 1998). دراین روش چندین دیدگاه وجود دارد که هریک را بصورت مجزا مورد بررسی قرار خواهیم داد.شکل زیر یک نمونه ارتباط بین چند جمله را آورده شده است. همانطوردر شکل ۳-۲ می بینید، می خواهیم ارتباط ٤ جمله فرضی را براساس هم وقوعی کلمات ان بدست اوریم. برای مثال دو جمله ۱۵ و ۲۳ در کلمه stock مشترک هستند، به عبارتی کلمه stock دراین جمله هم وقوع شده است. برای نشان دادن این ارتباط یک یال بین این دوجمله رسم می کنیم. برای مثالی دیگر دو جمله ۲٦ و ۱۵ دردو کلمهstock وmarket مشترک می باشند. بنابراین گراف مورد نظر حاصل می شود.



شکل ۳-۲: یک نمونه ارتباطی لغوی درچند جمله

ازجمله روش هایی که از این متد استفاده کرده اند به صورت زیر می باشد:

- روش هم وقوعی کلمات[۱]

[۱]Cohesion-based methods

- شباهت لغوی (زنجیره لغوی[۲]، WorldNet)
- ترکیبی از روش های بالا

۲-۲-۳-۷- روش هم وقوعی کلمات

این روش بنیادی برای اولین بار بوسیله Mitra درسال ۱۹۹۷ ارائه گردید. دراین روش از تکنیک های بازیابی اطلاعات درسطح متن استفاد می شود. به عبارتی متون را می توان مجموعه ای از پاراگراف ها درنظر گرفت. بااستفاده از متدهای پیشین در بازیابی اطلاعات و با استفاده از معیارهای شباهت بین کلمات ، شباهت بین پاراگراف p_i رابادیگر پاراگراف های متن تعیین می کنیم (Salton, Singhal, Mitra and Buckley, 1997). این روش بدین صورت است که ابتدا باید یک واحد برای متن خود جهت تشابه انتخاب کنیم که Mitra درمقاله خود این واحد را پاراگراف انتخاب کرده است.درقدم بعدی هر پاراگراف را به صورت یک بردار در نظر می گیریم که عناصر ان وزن یا اهمیت ترم (d_{ik}) مورد نظر در پاراگراف است که این ترم می تواند کلمه باشد. درقدم بعدی باید بتوانیم شباهت بین دو پاراگراف رامشخص کنیم.بادراختیار داشتن بردار ۲-۲

$$D_i = (d_{i1},\dots,d_{im}) \qquad (۱-۲)$$

درقدم بعدی باید شباهت بین پاراگراف را مشخص کنیم.بادراختیار داشتن بردار D_i و نرمال کردن آن شباهت دو پاراگراف از فرمول ۳-۲ بدست می آید:

$$\mathrm{Sim}(D_i, D_j) = \sum d_{ik}, d_{jk} \qquad (۳-۲)$$

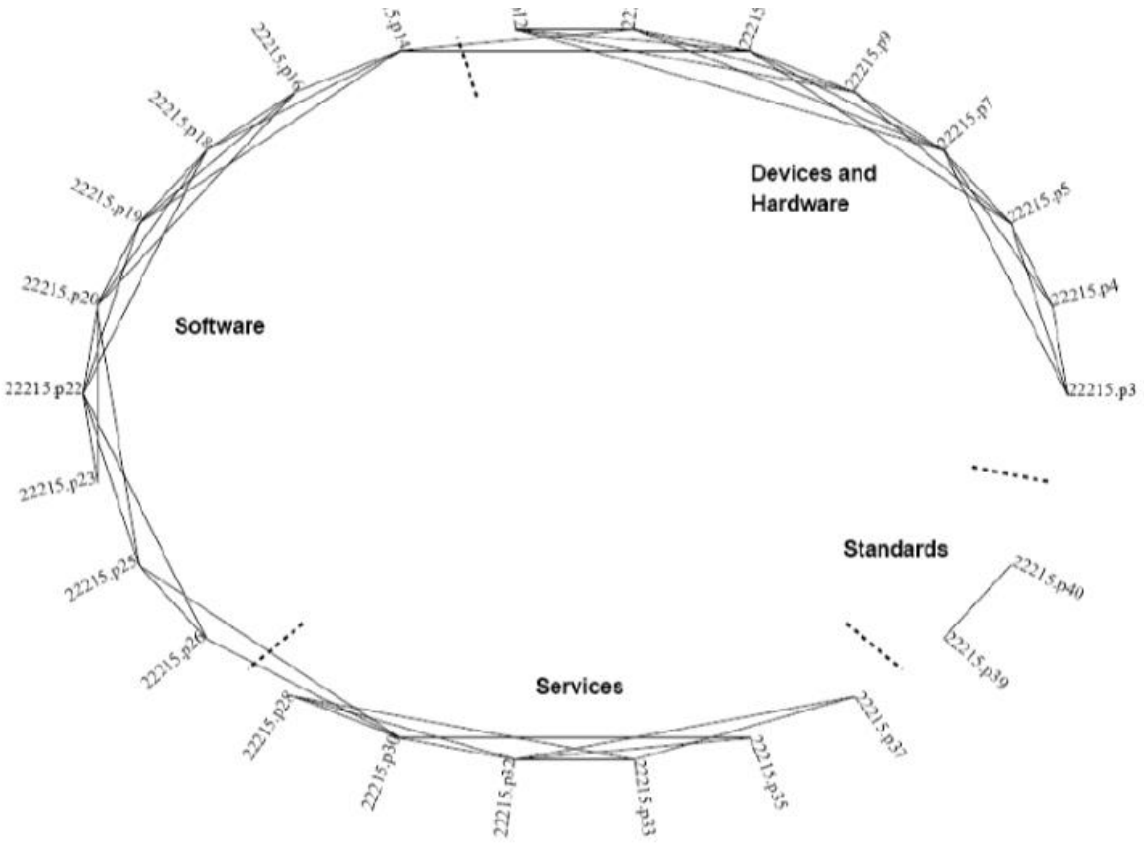
سپس کلیه پاراگراف را به هم متصل نموده و وزن اتصالات را برابر مقدار شباهت آنها قرارمی دهیم. بدین ترتیب برای یک متن به گرافی بدست می آید که به آن نقشه ارتباطی متن گفته می شود.درقدم بعد باید یال هایی که وزن کمتر از ییک مقدارآستانه دارند را حذف کنیم که درنهایت به گرافی مانند شکل زیر می رسیم.

بادقت درگراف بدست امده درشکل٤-۲ متوجه می شویم که برای مثال دو پاراگراف p39 و p40 میزان ارتباطی بیشتر از حد آستانه را نداشته اند. بنابراین تشکیل یک حوزه یا خوشه معنایی را می دهند که در این مثال دربردارنده معنای Standards هستند. بنابراین گراف ما باحذف ارتباطات کمتر از مقدار آستانه، به چند بخش تقسیم می شوندکه اتصالات درون هر بخش بیشینه و اتصال بین بخش ها کمینه

[۱] WORD co-occurrence

[۲] Lexical chains-based method

است.این زیر گراف ها در واقع پاراگراف هایی را نشان می دهند که از لحـاظ معنـایی بیشـترین ارتبـاط را دارند و همگی تقریبا یک مفهوم را دربردارند. برای مثال درگراف مثال قبل این بخش ها نشان داده شده اند.



شکل ٤-٢ : یک نمونه ارتباط بین پاراگراف ها در یک متن

درقدم بعدی باید ازبین پاراگراف های یک بخش، مهمترین انرا انتخاب کنیم. برای ایـن انتخـاب روش های متنوعی ارائه گردیده که برای مثال می توان پاراگرافی را انتخاب نمـود کـه بیشـترین اتصـال را داشـته باشد. برای ارزیابی نیز بدین طریق عمل شده که ابتدا پاراگراف های مهم یک متن توسط دو فرد اسـتخراج گردیده و سپس میزان همپوشانی پاراگراف های استخراجی این دوفرد مشخص می گردد. سپس همپوشانی پاراگراف های استخراج شده بوسیله سیستم را دربهترین حالت و بدترین حالت و اشتراک پـاراگراف هـای