رساله دکتری

گروه مهندسی کامپیوتر

استخراج خودکار مفاهیم از متن مبتنی بر روش های زبان شناسی

نگارنده:

محسن کامیار

استاد راهنما:

دکتر محسن کاهانی

استاد مشاور:

دکتر نادر جهانگیری

مرداد ۱۳۹۳


بسم الله الرحمن الرحيم

تقدیر و تشکر

بدینوسیله بر خود لازم می‌دانم که از زحمات بی‌دریغ استاد گرامی، جناب آقای دکتر محسن کاهانی که راهنمایی‌های ارزشمند ایشان در تمام مراحل انجام این پایان‌نامه راه‌گشای من بوده است تشکر نمایم.

سعی نابرده در این راه به جایی نرسی مزد اگر می طلبی طاعت استاد ببر

همچنین از تلاش‌های بی‌وقفه و پشتیبانی‌های پدر و مادر عزیزم و همسر مهربانم کمال سپاس‌گزاری را دارم.

چکیده فارسی

یکی از مهمترین رویدادها در حوزه علم بشر که تأثیر زیادی نیز بر نحوه زندگی بشر داشته است شبکه جهانی اینترنت و به صورت خاص «وب» می باشد. بسیاری معتقدند که امروزه وب مهمترین ابزار بیان دانش و دیدگاه های بشر محسوب می گردد. این اطلاعات که بدون این فناوری قابل دستیابی نبودند می توانند مورد استفاده افراد مختلفی قرار گیرند. ساده ترین نوع استفاده از این اطلاعات در دنیای تحقیقات، کاملا مشهود است. اگر این فضای تبادل اطلاعات وجود نداشته باشد چگونه می توان از آخرین نتایج تحقیقات مطلع گشت.

اما حجم اطلاعات موجود بسیار زیاد است و مسلما باید ابزارهایی برای دستیابی سریع به اطلاعات نیز موجود باشند. این ابزارها در ابتدا شامل لیست های موضوعی از سایت های مختلف بودند که به سرعت جای خود را به موتورهای جستجو دادند. اما این موتورهای جستجو همگی با مشکل دیگری مواجه هستند. زبان ارائه دانش در وب تنها برای ارائه دیداری اطلاعات ابداع شده است و به غیر از محتوی نمی تواند فرااطلاعات قابل توجهی را جهت جستجوی مناسب در اختیار قرار دهد. به علت این ساختار چنانچه برای فرااطلاعات موجود در این زبان ارائه اطلاعات امتیازی را در نظر بگیریم با مشکلات متعددی مواجه خواهیم شد که علت اصلی آن عدم استفاده صحیح افراد به عمد و یا غیر عمد از این فرااطلاعات می باشد.

برای حل این مشکل مسیری جدید به نام وب معنایی ایجاد گردید. در این مسیر مهمترین جزء را می توان هستان شناسی (Ontology) دانست. هر هستان شناسی در یک حوزه معنایی خاص تعریف می گردد. هستان شناسی عبارت است از مجموعه ی تمام مفاهیم در حوزه معنایی مرتبط، مشخصات هر یک از این مفاهیم، محدودیت های موجود برروی مشخصات مفاهیم و در نهایت ارتباطات موجود بین مفاهیم متفاوت.

ایجاد یک هستان شناسی محتاج آشنایی کامل با تمام دانش در یک حوزه معنایی می باشد. چیزی که تقریبا دست یافتنی نیست. به همین خاطر است که عمده هستان شناسی های تولید شده تا کنون تنها برروی مجموعه ی بسیار کوچکی از یک حوزه ی معنایی خاص متمرکز شده اند. پس برای تحقق هدف اصلی وب معنایی باید بتوانیم به طریقی خودکار و یا نیمه خودکار تولید هستان-شناسی را انجام دهیم.

در این پژوهش روشی جدید برای استخراج خودکار مفاهیم ارائه می گردد که تولید خودکار هستان شناسی را دست یافتنی تر می نماید. مهم ترین مشکل در روش های موجود در این زمینه را می توان در این دانست که این روش ها به فضای زبانی که اسناد در این فضا تعریف می شوند به چشم فضاهای هندسی معمول نگاه می کنند. از این رو علاوه بر عدم دقت مناسب نتایج، دارای پیچیدگی محاسباتی بسیار بالایی می باشند و عملا استفاده از آن ها در محیط وب که حجم اطلاعات بسیار بالا است غیرممکن می باشد. در این پژوهش

نگاهی جدید به فضای زبانی خواهیم داشت و دیدگاه های هندسی موجود به فضای کلمات زبان را کنار خواهیم گذاشت.

راه کارهای متفاوتی برای دستیابی به هدف فوق ارائه خواهد گردید. یکی از گام ها حرکت از فضاهای کاملا هندسی به سمت فضاهای کلی تر مانند فضاهای باناخ می باشد. این تغییر دیدگاه ابزارهای جدیدی را در اختیار خواهد گذاشت که از آن جمله می توان به تعریف معیارهای اهمیت و مشابهت جدید اشاره نمود.

گام دیگر ارائه ی یک روش خوشه بندی مناسب به جای روش های موجود برای استخراج مفاهیم می باشد تا از این طریق محدودیت های موجود در پیچیدگی محاسباتی روش های موجود مرتفع گردد.

در تمام راه کارهایی که در فوق عنوان گردید استفاده از معیارهای زبان شناختی می تواند راه گشایی مهم باشد. از این جمله می توان به عناصر موجود برروی زنجیره گفتار اشاره نمود. این عناصر در یک گفتگوی محاوره ای خود را در قالب تأکید برروی یک کلمه خاص، حرکات بدن و دیگر داده های غیر متنی نشان می دهند؛ اما در یک متن، نویسنده باید سعی نماید تا با استفاده از کلمات و قالب های نوشتاری این عناصر را تبدیل به متن نماید که با دخیل نمودن چنین قالب هایی می توان از این ویژگی ها استفاده لازم را برد.

واژه های کلیدی: وب معنایی، هستان شناسی، استخراج خودکار هستان شناسی، استخراج مفاهیم، موتورهای جستجو

Keywords: Semantic Web, Ontology, Automatic Ontology Extraction, Concept Extraction, Search Engine

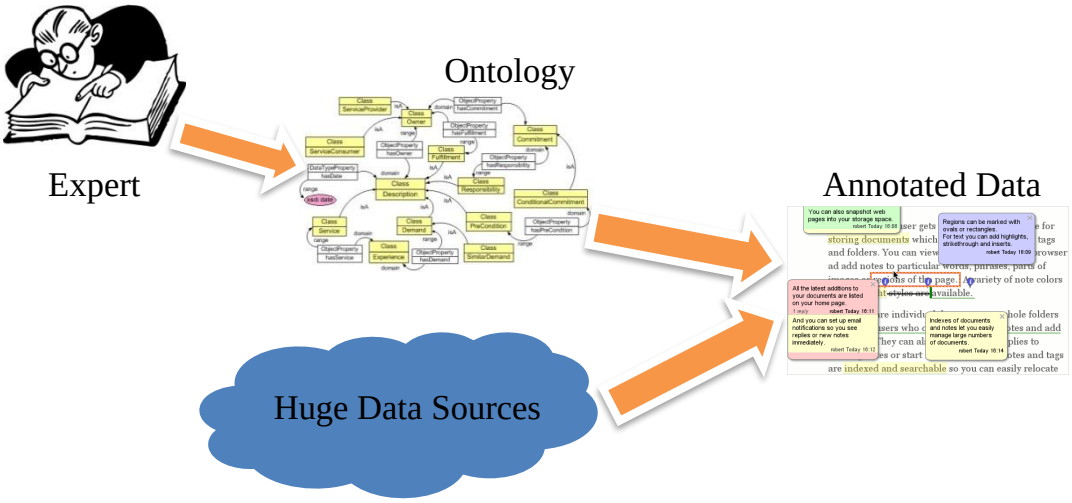
فهرست

فهرست جداول

فهرست شکل ها

۱. مقدمه

روش های جاری در حوزه بازیابی اطلاعات نتوانستند به خوبی نیازهای انسان در دستیابی به اطلاعات مورد نیازش را فراهم آورند. به همین دلیل بود که مسئله ورود معنا در داده ها مطرح شد و به عنوان گام بنیادین نیاز به ساختارهایی بود تا بتوان با استفاده از آن ها معنا را در بازیابی اطلاعات دخیل نمود. این ساختارها را اصطلاحا هستان شناسی[1]نامیدند. در شکل ۱ نمونه ای از ارتباطات موجود در حوزه وب معنایی برای بیان اطلاعات نمایش داده شده است.



شکل ۱ – نمایی کلی از تولید وب معنایی

[1] Ontology

همانطور که واضح است هستان شناسی نقش اساسی در دستیابی به اهداف بیان شده را ایفا می نماید. اما سابقه وی معنایی تا امروز نشان می دهد که ایجاد یک هستان شناسی به صورت غیر خودکار در یک حوزه کوچک از دانش بشر نیز می تواند کار بسیار پر هزینه ای باشد.

کاملا واضح است که ایجاد هستان شناسی برای تمامی حوزه های دانش بشر به صورت غیرخودکار دست-نیافتنی است و یا اینکه با صرف هزینه‌ی بسیار بالایی می‌توان به آن دست یافت. در همین راستا تلاش‌هایی برای تولید خودکار هستان‌شناسی صورت گرفته است [AHM05، TOD01، MAE01، BHA04، RUD07، SUN08، AND02، SLE03]. عمده این روش‌ها دارای روش کار نسبتا مشابهی می‌باشند. در این پژوهش نیز همین بخش هدف قرار گرفته است. جزئیات کارهای انجام شده در بخش‌های بعد مورد بررسی قرار خواهند گرفت.

۱٫۱. استخراج هستان شناسی

همانطور که از توضیحات در پیش آمده مشخص می باشد، در اختیار داشتن یک هستان شناسی نقشی بسیار مهم را در وب معنایی ایفا می نماید و می توان آن را پیش زمینه ورود به این دنیا دانست. البته وجود یک هستان شناسی بدون حضور دیگر مؤلفه های وب معنایی نیز یک امتیاز مهم محسوب می شود که می تواند کاربردهای زیادی را امکان پذیر نماید.

۱٫۱٫۱. ساختار قالب عمومی استخراج هستان شناسی

همان طور که پیش از این نیز گفته شد یک هستان شناسی عبارت است از مجموعه ای از مفاهیم، ویژگی های هر مفهوم و مجموعه‌ی قوانین موجود برروی نحوه‌ی ارتباط مفاهیم گوناگون با یکدیگر و همین‌طور محدودیت‌های موجود برروی مقادیر ویژگی‌های مفاهیم مختلف. با این وصف عمده‌ی کارهای انجام شده برروی استخراج هستان‌شناسی نیز به دو دسته استخراج مفاهیم و استخراج قوانین تقسیم می‌گردند. در ادامه فعالیت‌های صورت گرفته در هر حوزه را به صورت جداگانه مورد بررسی قرار خواهند گرفت.

عمده فعالیت‌های صورت گرفته در زمینه استخراج مفاهیم [PAP98،HOF99،LEA03] دارای ساختاری به شرح زیر می‌باشند. مهم‌ترین روش مورد استفاده در این کارها الگوریتمی به نام LSI[۲][DUM88] می-باشد. روند کلی کار به این صورت است که ابتدا برای یک مجموعه از اطلاعات موجود در یک زمینه‌ی معنایی از پیش مشخص شده، تمام کلمات استخراج شده و ماتریس لغت-صفحه ایجاد می‌گردد. در ایجاد این ماتریس از لغات به صورت مستقیم استفاده نمی‌گردد بلکه ابتدا تمام کلمات با استفاده از یک روش

[2] Latent Semantic Indexing

ریشه‌یابی، به ریشه‌ی خود تبدیل شده و این ریشه است که به عنوان کلمه‌ی مورد نظر در ماتریس لغت-صفحه جای می‌گیرد. ذکر این نکته دارای اهمیت است که در حال حاضر مهم‌ترین الگوریتم‌های مورد استفاده در ریشه‌یابی لغات انگلیسی دو روش [KRO93،POR80] می‌باشند.

نکته‌ی مهم اینکه در این رویکرد فرض بر این است که مجموعه‌ای از صفحات و یا اسناد مرتبط با حوزه‌ی معنایی مورد نظر موجودند. اما برای دست‌یابی به چنین مجموعه‌ای در وب باید از روش‌های خزش متمرکز[۳] استفاده نمود. در این روش‌ها سعی می‌شود تا با استفاده از مجموعه‌ای از پارامترها تنها به استخراج صفحاتی اقدام گردد که مرتبط با حوزه‌ی معنایی مورد نظر می‌باشند. در این فرآیند در ابتدا باید مجموعه‌ی کوچکی از صفحات مرتبط موجود باشند که عملیات خزش از همین مجموعه‌ی اولیه شروع می‌شود. عمده پارامترهایی که در روش‌های مختلف مورد استفاده قرار می‌گیرند را می‌توان فاصله‌ی یک صفحه براساس تعداد لینکهای پیمایش شده نسبت به مجموعه‌ی اولیه، ضریبی از میزان اهمیت پدر یک صفحه، حضور مجموعه‌ای از کلمات کلیدی، حضور مهم‌ترین کلمات موجود در صفحات بازیابی شده تا کنون و بسیاری معیارهای مبتنی بر حدس و یا یادگیری دیگر دانست. البته در بعضی از این روش‌ها از هستان‌شناسی حوزه-ی معنایی مورد نظر نیز استفاده می‌شود که با توجه به هدف این پژوهش چنین استفاده‌ای غیر منطقی می-باشد.

با توجه به آنچه مرور شد پس از تولید ماتریس لغت-صفحه می‌توان گفت که در اصل مجموعه‌ای از نقاط (صفحات) در فضای لغات نتیجه عملیات مذکور می‌باشد. تعداد ابعاد این فضا برابر است با تعداد کل لغات موجود. مقادیر هر نقطه از این فضا در هر بعد می‌تواند براساس معیارهای گوناگونی سنجیده شود.

ساده‌ترین معیار قابل استفاده می‌تواند معیار حضور یا عدم حضور باشد که در صورت حضور یک کلمه در یک صفحه مقدار نقطه‌ی متناظر با صفحه‌ی مورد نظر ۱ و در غیر این صورت ۰ خواهد بود. مشخص است که چنین معیاری با توجه به بی تفاوت بودن نسبت به میزان تکرار یک کلمه نمی‌تواند معیار مناسبی محسوب گردد.

معیار دیگری که مورد استفاده قرار می‌گیرد معیار میزان تکرار کلمه در یک صفحه است. در این معیار چنانچه کلمه‌ی a در صفحه‌ی d_1، x بار تکرار شده باشد و در صفحه‌ی d_2، y بار تکرار شده باشد و $x<y$ باشد اما برای تعداد کل کلمات موجود در هر یک از این دو صفحه داشته باشیم $||d_1||>>||d_2||$ آنگاه کلمه‌ی a را

[3] Focused Crawling

برای صفحه d_2 پراهمیت‌تر از صفحه d_1 تشخیص می‌دهد حال آنکه واضح است چنین نیست. این معیار مشکلات بسیار زیاد دیگری نیز دارد که در اینجا نمی‌توان تمام موارد را ذکر نمود.

با توجه به مشکلات موجود در معیار بالا، معیار دیگری جایگزین گردیده است که به آن TFiDF[4] [JON72] گفته می‌شود. نحوه‌ی محاسبه‌ی این معیار به شرح زیر است.

شبه کد1 - شبه کد محاسبه TFiDF

$$\text{for term } i \text{ and document } j \Rightarrow tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}}$$

$$\text{for tem } i \Rightarrow idf_i = \log \frac{|D|}{|\{d : t_i \in d\}|}$$

$$TFiDF_{i,j} = tf_{i,j} \times idf_i$$

این معیار توانسته است تا بسیاری از مشکلات را به خوبی حل نماید و نتایج بسیار مناسبی ارائه کند، اما همچنان با مشکلاتی مواجه است که در بخش بعد توضیح داده خواهد شد.

با توجه به مطالب ذکر شده در بالا ماتریسی به دست می‌آید که اصطلاحا به آن مدل فضای برداری[5] گفته می‌شود. این ماتریس برای ادامه فعالیت مورد استفاده قرار می‌گیرد. این ماتریس خواصی دارد که باید در انتخاب الگوریتم‌ها مورد توجه قرار گیرند. مهم‌ترین خاصیت این ماتریس خالی[6] بودن آن است که مهم‌ترین تأثیر را در انتخاب الگوریتم‌ها می‌گذارد.

در گام بعد الگوریتم LSI مورد استفاده قرار می‌گیرد. این الگوریتم به این صورت عمل می‌نماید که با استفاده از الگوریتم تقسیم ماتریس SVD [CVE79] ماتریس مدل فضای[7] برداری را به سه ماتریس لغت-مفهوم، بردارهای ویژه و مفهوم-صفحه تقسیم می‌کند.[8] بردارهای ویژه دارای این خاصیت هستند که می‌توانند برای خوشه‌بندی نودهای ماتریس‌ها براساس میزان[9] وابستگی آن‌ها مورد استفاده قرار گیرند.

---

4 Term Frequency inverse Document Frequency
5 Vector Space Model
6 Sparse
7 Singular Value Decomposition
8 Eigen Vectors
9 Clustering

کارهای زیاد دیگری نیز در همین حوزه صورت گرفته است که از جمله‌ی مهم‌ترین آن‌ها می‌توان به جایگزینی SVD با روش دیگری به نام SDD[CVE79][10] در الگوریتم LSI که دارای پیچیدگی محاسباتی کم‌تری نیز می‌باشد و همچنین ارائه‌ی نمونه‌های جدیدی از LSI با توجه به خالی بودن ماتریس مدل فضای برداری [MAV07،SAL75] و همچنین تبدیل الگوریتم LSI به یک الگوریتم برخط[11] از حالت برون‌خط[12] فعلی [TOU08،TOU05،ZHA99] اشاره نمود.

پس از به دست آوردن ماتریس‌های فوق با استفاده از ماتریس لغت-مفهوم و با توجه به مقادیر ورودی‌های ماتریس به ازای هر مفهوم مهم‌ترین لغات شناسایی می‌شوند. با جداسازی این لغات یک فرد خبره خواهد توانست مفهوم مرتبط با هر مجموعه از لغات را نام‌گذاری نماید و مجموعه‌ی مفاهیم استخراج می‌گردند.

در این زمینه کارهای دیگری نیز انجام شده‌اند، اما می‌توان گفت مهم‌ترین رویکرد موجود روشی است که در بالا توضیح داده شده است.

مهم‌ترین گام بعدی در استخراج هستان‌شناسی را می‌توان استخراج قوانین دانست. در یکی از مهم ترین کارهای انجام شده در این زمینه سعی شده است با استفاده از ماتریس لغت-مفهوم که پیش از این توضیح داده شد و تشکیل یک ماتریس دوبخشی بین لغات و مفاهیم روابط بین مفاهیم از قبیل زیرمجموعه‌بودن، اشتراک در بعضی خصوصیات و دیگر ویژگی هایی که عمدتا مرتبط با ویژگی های مجموعه ای بین کلمات و صفحات هستند را شناسایی نماید. اما این قوانین به همین مجموعه منحصر شده‌اند و به قوانین پیچیده‌ای مانند محدودیت برروی مقادیر ویژگی‌های مفاهیم، ارتباط‌های یک به چند بین مفاهیم گوناگون و .. اشاره نشده است و کاملا نیز واضح می‌نماید که استخراج چنین قوانینی با استفاده از چنین مدلی نمی‌تواند قابل انجام باشد. کارهای دیگری در زمینه‌های دیگر مانند پایگاه‌های دانش برای استخراج قوانین وابستگی[13] صورت گرفته است که از یک سو دارای همخوانی با مدل موجود در استخراج هستان‌شناسی نمی‌باشند و از سوی دیگر پا را فراتر از روش ذکر شده در بالا نمی‌گذارند.

با توجه به مجموعه‌ی اطلاعات ارائه شده لازم می‌نماید نگاهی مجدد به حوزه استخراج هستان‌شناسی صورت گیرد. شاید بتوان گفت حرکت به سمت استخراج قوانین در حال حاضر مسیری کاملا نامشخص بنماید و از این رو انتخاب چنین موضوعی ممکن است با مشکلات جدی در مسیر مواجه گردد. از روی دیگر در حوزه استخراج مفاهیم نیز مسائل باز بسیاری موجودند که می‌توان تلاش را در این زمینه متمرکز نمود.

---

[10] Semi Discrete Decomposition
[11] Online
[12] Offline
[13] Association Rules

همچنین در یک نگاه کلی می‌توان گفت که مهم‌ترین دلیل استفاده از الگوریتم‌های ریشه‌یابی و همچنین معیارهایی مانند TFiDF این است که بتوان از این طریق فضای حالت ماتریس لغت-صفحه را تا حد ممکن کوچک نمود. پس چنانچه بتوانیم مفاهیم را در یک حوزه‌ی معنایی استخراج نماییم علاوه براینکه برای استخراج هستان‌شناسی مورد استفاده قرار خواهد گرفت همچنین می‌تواند به عنوان ابزاری کارا در کاهش حجم ماتریس صفحه-لغت مورد استفاده قرار گیرد.

با این اوصاف در بخش بعد تمرکز برروی مسئله‌ی استخراج مفاهیم خواهد بود و با ذکر مشکلاتی که در روش‌های ذکر شده موجودند مسیر حرکت کمی روشن‌تر می‌گردد.

۱،۱،۲.مشکلات مؤلفه های موجود در استخراج هستان شناسی

در بررسی این مشکلات بهتر است تا مهم‌ترین مؤلفه‌های موجود در استخراج مفاهیم بار دیگر ذکر گردند و سپس در هر مورد به توضیح در مورد مشکلات پرداخته شود. مهم‌ترین مؤلفه‌ها را می‌توان به شرح زیر دانست:

- معیارهای مشابهت
- معیارهای اهمیت
- LSI

عمده معیار مشابهتی که در این دسته از روش‌ها مورد استفاده قرار می‌گیرد معیار مشابهت کسینوسی می-باشد. در این معیار فرض بر این است که هر نقطه از فضا معادل با یک بردار در فضا است که از نقطه مبدأ تا نقطه‌ی مورد نظر امتداد می‌یابد. حال برای دو نقطه در فضا دو بردار وجود دارد و میزان شباهت این دو برابر است با فاصله‌ی کسینوسی این دو و از رابطه‌ی زیر محاسبه می‌گردد:

$$dist(v_1, v_2) = \frac{v_1 . v_2}{|v_1| . |v_2|} \qquad (۱)$$

این معیار به خوبی می‌تواند برای دو نقطه که واقعا در فضای اقلیدسی[۱] تعریف شده باشند مورد استفاده واقع شود. اما در استفاده این معیار برای دو صفحه از وب با مشکلاتی مواجه می‌شویم. اولین نکته این است که در مورد صفحاتی که اندازه‌های آن‌ها بسیار متفاوت است به درستی کار نمی‌کند. به عبارت دیگر این معیار در صورتی می‌تواند برای دو صفحه اندازه‌ی درستی را تولید نماید که این دو صفحه از حیث اندازه دارای شباهت باشند. نکته‌ی بعد اینکه چنانچه یکی از این صفحات زیرمجموعه‌ی صفحه‌ی دیگر باشد در

[1] Euclidean Space [4]

این صورت میزان شباهت دو صفحه بسیار کم به دست می‌آید. علت این امر این است که به ازای تعداد زیادی از نقاط صفحه کوچک‌تر مقدار صفر را دارد حال آنکه این دو صفحه در اصل بسیار به هم شبیه می-باشند.

تقریبا می‌توان گفت معیار مشابهت ذکر شده در بالا پراستفاده‌ترین معیار مورد استفاده است. در بخش‌های بعد علت عدم کارکرد مناسب این معیار به شرح ذکر شده در فوق توضیح داده خواهد شد.

پس از معیارهای مشابهت باید به توضیح در مورد مشکلاتی پرداخت که معیارهای اهمیت با آن مواجه می-باشند. همان‌طور که در بخش قبل توضیح داده شد مهم‌ترین معیار اهمیت لغت در صفحه‌ی مورد استفاده معیار TFiDF می‌باشد. اما این معیار نیز با مشکلاتی مواجه است. باز هم دوباره اولین مشکل مربوط به صفحاتی است که دارای اندازه‌ی خارج از حد طبیعی هستند. در چنین مواردی تقریبا معیار فوق برای تمام لغات نزدیک به صفر خواهد بود و نخواهد توانست به درستی میزان اهمیت لغات نسبت به هم در همان صفحه و یا نسبت میزان اهمیت یک لغت در صفحه مورد نظر نسبت به صفحات دیگری که همان لغت را دارند نشان دهد. علاوه بر این انجام هرگونه نرمال‌سازی برروی مقادیر این معیار نیز ممکن است باعث ایجاد نتایج کاملا غیرواقعی گردد.

مشکل بعدی در این معیار این است که اطلاعات مربوط به ترتیب لغات را از دست می‌دهد. حال آنکه نشان داده شده است که در بسیاری از موارد ترتیب استفاده از لغات در نوع مفهومی که به همراه یک صفحه می-باشد بسیار تأثیرگذار می‌باشد. بر همین اساس مفهومی به نام n-gram معرفی شده و به عنوان جایگزین یک اصطلاح مورد استفاده قرار می‌گیرد و نقش یک لغت مجزا را ایفا می‌کنند [BRO92].

یکی دیگر از مشکلات این روش این است که به عنوان مثال چنانچه صفحه‌ای در مورد توضیح دادن مفهومی باشد مسلما از لغت آن مفهوم کمتر استفاده خواهد نمود و از این رو این معیار نخواهد توانست میزان اهمیت لغت مربوطه در صفحه‌ی مورد نظر را به درستی تشخیص دهد. حتی استفاده از کلمات هم‌معنی و دیگر دانش های ئیش زمینه در خصوص معانی کلمات نیز در چنین مواردی نمی‌توانند تأثیر چندانی در نتیجه داشته باشند.

می‌توان یکی از مهم‌ترین مشکلات این روش‌ها را عدم اتکای آنها بر مؤلفه‌های تأثیرگذار از دیدگاه زبان شناختی دانست. این مسئله باعث می شود تا اهمیت کلمات تنها براساس معیارهای آماری سنجیده شوند و از این رو مشکلاتی از جنسی که در بالا آمد را فراهم آورد. همان طور که پیش از این نیز گفتیم نمونه آن را می‌توان عناصر موجود برروی زنجیره گفتار دانست. براساس نظریه‌های زبان شناسی لغات تنها کمتر از ۳۰ درصد در بیان معنا تأثیرگذار هستند و در اصل معنا از طریق همین عناصر موجود برروی زنجیره گفتار

انتقال می یابند. از جمله این عناصر می‌توان به تأکید برروی یک کلمه، مکث، حالت بدن و دیگر داده های غیر متنی اشاره نمود.

اما در یک متن نویسنده باید همین عناصر موجود برروی زنجیره گفتار را به کلمات تبدیل نموده و منظور خود را به خواننده القا نماید. در اصل چنین کلماتی در تعیین معنا به صورت مستقیم تأثیری ندارند اما با نشان دادن سمت و سوی کلی در متن به ما در تشخیص کلمات مهم، تأییدات و نفی‌ها و بقیه داده هایی که در محاوره از طریق متن انتقال نمی یابند کمک می‌نمایند.

آخرین مؤلفه که در اینجا مورد بررسی قرار می‌گیرد روش LSI می‌باشد. همان‌طور که پیش از این نیز توضیح داده شد در این روش از الگوریتم دیگری به نام SVD برای تقسیم ماتریس استفاده می‌شود. مهم-ترین خاصیت مورد استفاده در این روش خاصیت خوشه‌بندی است که با استفاده از بردارهای ویژه صورت می‌گیرد و از این طریق لغاتی که از نظر معنایی دارای همبستگی بالایی نسبت به هم می‌باشند در خوشه-های یکسانی قرار می‌گیرند.

اما این روش نیز با مشکلاتی مواجه می‌باشد. مهم‌ترین و اولین مشکلی که می‌توان برای آن ذکر نمود پیچیدگی زمانی بسیار بالای آن می‌باشد. این الگوریتم دارای پیچیدگی زمانی بالایی است که در حجم وب و با توجه به تعداد زیاد صفحات مورد بررسی و همچنین تغییرات مداوم آن‌ها تحمل چنین پیچیدگی زمانی امکان‌پذیر نیست.

مشکل بعدی این روش اینجاست که میزان کارایی آن تا کنون با استفاده از معیارهای تشابهی مانند آنچه پیش از این توضیح داده‌ایم سنجیده شده است. همچنین در جزئیات عملکرد خود آن نیز از چنین معیارهایی استفاده گردیده است که مسلما با توجه به توضیحات داده شده نمی‌توانند به درستی نشان‌دهنده میزان شباهت دو نقطه در مدل فضای برداری باشند.

یکی دیگر از مشکلات این روش برون‌خط بودن آن است. به این معنی که در این روش به صورت پیش فرض در صورت رخ دادن تغییرات در ماتریس مدل فضای برداری باید کلیه‌ی محاسبات مجددا صورت گیرند تا مفاهیم جدید استخراج گردند. البته فعالیت‌هایی در این زمینه صورت گرفته است که پیش از این نیز به آن اشاره شد [TOU08،TOU05،ZHA99] اما این روش‌ها نیز با پیچیدگی زمانی بالایی مواجه بوده و قابل استفاده برای کاربردهای برخط نمی‌باشند.

یکی دیگر از مشکلات توضیح داده شده این است که این روش مبتنی بر خواصی است که ممکن است در مورد ماتریس مدل فضای برداری صادق نباشند. از این جمله می‌توان به مبتنی بودن این روش بر بعضی

توزیع‌های خاص تصادفی مانند مدل گاوسی و همچنین خالی بودن ماتریس مدل فضای برداری اشاره کرد که باعث می‌شود این روش از حیث خواص متناسب با ماتریس مدل فضای بردار نباشد.

در بخش‌های بعد تمرکز اصلی ابتدا برروی ریشه‌یابی این مشکلات و سپس راهکارهایی برای ادامه این پژوهش و حل مشکلات خواهد بود.

۱,۲. راهکار پیشنهادی

در قسمت های قبل موفق شدیم تا نمایی کلی از مسئله را بیان نماییم و همین طور مشکلات موجود را مرور نماییم. همچنین تا حدودی اصول راهکار پیشنهادی مشخص گردید. در اینجا سعی می شود تا این مطالب سر جمع شده و نمایی کامل از راهکار پیشنهادی ارائه گردد. اما برای ارائه راهکار پیشنهادی باید یک گام جلوتر برویم و به ریشه یابی مشکلات بپردازیم. در اصل باید مشخص گردد که مشکلات بیان شده در قسمت قبل از کجا ناشی می شوند تا بتوانیم راهکار مناسب برای حل آنها بیابیم.

۱,۲,۱.ریشه یابی مشکلات

در یک نگاه دقیق‌تر می‌توان مشکلاتی را که در بخش قبل برای روال معمول در استخراج مفاهیم ذکر شد در عوامل زیر ریشه‌یابی نمود:

- عدم تطابق فضای این مسئله با خواص فضای اقلیدسی: فضای داده‌ای این مسئله عمدتا فضای لغت-صفحه و یا لغت-مفهوم می‌باشد. این فضا به هیچ عنوان از خواصی که در فضای اقلیدسی شناخته شده است تبعیت نمی‌کند. به عنوان مثال مشکلی که برای معیار شباهت کسینوسی ذکر شد قابل توجه است. در فضای اقلیدسی چنانچه دو نقطه کاملا مشابه موجود باشند و سپس تعداد زیادی از بعدهای یکی از این دو نقطه برابر صفر باشند اتفاقی که می‌افتد این است که واقعا این دو نقطه از هم دور خواهند شد، حال آنکه در فضای داده‌ای که ما در مسئله استخراج مفاهیم با آن مواجه هستیم چنین اتفاقی نخواهد افتاد. همین مسئله مجددا خود را در مورد مشکلی که معیار میزان اهمیت TFiDF با اندازه صفحات دارد نشان می‌دهد. در آنجا نیز می‌توان یک لغت را صفحه‌ای دانست که تنها همان یک لغت را در اختیار دارد. در مورد صفحاتی که تعداد زیادی لغت دارند نمی‌توان به درستی میزان شباهت صفحه مورد نظر با صفحه فرضی که تنها یک لغت دارد را تعیین نمود و در نتیجه میزان اهمیت لغت مورد نظر نیز به درستی قابل تشخیص نخواهد بود.
- عدم استفاده از اطلاعات مربوط به ترتیب لغات: یکی از اطلاعاتی که در این روش‌های معمول و مذکور پیشین به آن توجه نمی‌شود اطلاعات مربوط به ترتیب کاربرد لغات می‌باشد. مفهومی که

تحت عنوان n-gram در حال حاضر شناخته شده می‌باشد. باید سعی کرد تا چنین اطلاعاتی را نیز به نحوی با TFiDF ترکیب نمود.

- عدم استفاده از عناصر موجود برروی زنجیره گفتار: این عناصر در یک متن از حالت غیر لغوی به لغات تبدیل می شوند. چنین لغاتی به صورت منفرد هیچ معنایی را در متن با خود حمل نمی‌کنند بلکه میزان اهمیت دیگر لغات را مشخص می‌کنند. عدم توجه به این ساختارها باعث عدم توجه به لغات مهم و همچنین اهمیت‌ها در متن می‌گردند.
- عدم تطابق خواص مورد نظر در SVD با خواص مورد نظر در ماتریس مدل فضای برداری: از جمله این خواص می‌توان به پیروی نکردن ماتریس مدل فضای برداری از خواص توزیع گاوسی، عدم تطابق ماتریس مدل فضای برداری با خواص فضای اقلیدسی و خالی بودن ماتریس مدل فضای برداری اشاره نمود.
- عدم استفاده از اطلاعات عدم رخداد: یکی دیگر از مشکلاتی که این روش‌ها با آن مواجه می‌باشند و ناشی از روش اندازه گیری مشابهت مورد استفاده می‌باشد، توجه نکردن به عدم رخداد داده‌ها می‌باشد. اگر دو کلمه در موقعیت‌های مشابهی رخ ندهند این روش‌ها این مسئله را به عنوان عاملی در تعیین مشابهت مورد استفاده قرار نمی دهند.

۱،۲،۲. مروری بر راهکار پیشنهادی

براساس آنچه در بخش قبل به عنوان مشکلات اساسی روش‌های موجود بیان شد و همچنین ریشه یابی این مشکلات، در این پایان نامه دو حوزه اصلی به عنوان اهداف انتخاب گردیدند. حوزه اول ارائه راهکاری جهت بهبود استخراج دانش نهفته با تکیه بر بهبود معیارهای اندازه گیری میزان شباهت و یا فاصله می‌باشد. همانطور که گفته شد در این بخش یکی از مهم‌ترین مشکلات استفاده از معیارهایی است که تنها برای نوع داده‌های مشخصی قابل استفاده بوده و نمی توانند به عنوان روشی همه گیر به خوبی برای تمامی داده‌ها مورد استفاده قرار گیرند. از این رو راهکار پیشنهادی در این بخش ایجاد عمومیت بیشتر در معیارهای اندازه گیری مشابهت می‌باشد تا بتوانند به خوبی برای تمام انواع داده‌ها مورد استفاده قرار گیرند.

برای فائق آمدن بر این مشکلات یکی از تکنیک‌های مورد آماری مورد استفاده قرار گرفته است که به آن‌ها توابع ارتباط[۵]گفته می‌شود. توابع ارتباط، توابعی هستند که تخمینی از تابع توزیع چند جمله ای برای دو یا چند متغیر تصادفی داده شده و با تابع توزیع حاشیه ای داده شده را ارائه می‌نمایند. همچنین براساس این توابع می‌توان میزان همبستگی بین دو یا چند جمله را محاسبه نمود. همبستگی محاسبه شده از این طریق

---

[1] Copula Functions [5]

دارای اطلاعات زیادی در مورد متغیرهای تصادفی مورد نظر می‌باشد. از جمله اینکه توابع ارتباط نه تنها می‌توانند شباهت در هنگام رخداد متغیرهای تصادفی را نمایش دهند بلکه می‌توانند شباهت را در هنگام عدم رخداد متغیرهای تصادفی نیز نمایش دهند.

شاید بتوان گفت که مهم‌ترین ویژگی توابع ارتباط عدم وابستگی آن‌ها به تابع توزیع متغیرهای تصادفی مورد استفاده می‌باشد. حتی می‌توان از متغیرهای تصادفی با توابع توزیع متفاوت با هم نیز استفاده نمود. این همان خاصیتی است که روش‌های موجود وجود ندارد و همخوانی زیادی با داده‌ها در دنیای واقعی وب دارد. این خاصیت می‌تواند راهگشای بسیاری از مشکلات موجود در روش‌های جاری باشد.

اما یکی از مشکلات استفاده از توابع ارتباط هزینه محاسباتی آن‌ها می‌باشد. محاسبه یک تابع ارتباط مناسب برای n متغیر معمولا مستلزم محاسبه یک انتگرال چندگانه با همان درجه تعداد متغیرها می‌باشد. اما حتی روش‌های تقریبی عددی محاسبه انتگرال دارای هزینه بسیار بالایی می‌باشند. این مسئله باعث شده است تا توابع ارتباط عمدتا در خصوص مسائلی مورد استفاده قرار گیرند که دارای تعداد متغیرهای محدود می‌باشند. به عنوان مثال این توابع در حوزه اقتصاد بسیار کاربرد دارند. اما کمتر در حوزه علوم کامپیوتر مورد استفاه قرار گرفته‌اند.

در این پایان نامه سعی شده است تا علاوه بر استفاده از توابع ارتباط به عنوان راهکاری مناسب جهت مقابله با مشکلات روش‌های موجود در استخراج مفاهیم، مشکلات این توابع نیز حل گردد تا بتوان از این روش‌ها به عنوان راهکارهایی کارا و با زمان اجرای مناسب جهت به دست آوردن مفاهیم استفاده نمود.

رویکرد دیگر مورد توجه در این پایان نامه پرداختن به مشکلات مربوط به وزن دهی لغات درمتن می‌باشد. عمده روش‌های مورد استفاده روش‌های آماری می‌باشند. این روش‌ها از مؤلفه اصلی بسامد تکرار یک کلمه استفاده نموده و به روش‌های مختلف تلاش در بهبود آن دارند. اما همه این روش‌ها به علت مبتنی بودن بر تکرار به راحتی دچار اشتباه می‌شوند. این اشتباه می‌تواند شامل اهمیت دادن به کلمات فاقد اهمیت و یا برعکس اهمیت ندادن به کلمات با اهمیت باشد. با توجه به این رفتار دوگانه بهبود آن‌ها کار بسیار دشواری می‌نماید. چنانچه روشی در یکی از دو جهت دارای مشکل باشد ارائه روشی جهت بهبود آن بسیار ساده‌تر می‌باشد. اما در روش‌هایی با هر دو نوع خطا بهبود کار سخت تری می‌باشد.

می توان گفت که شاید راه بهبود این مشکلات حرکت در ورای خواص آماری باشد. به کار بردن خواص زبان شناختی داده‌ها یک راهکار اساسی در بهبود این روش‌ها می‌باشد. به همین منظور در این پایان نامه سعی

شده است تا با استفاده از نظریه مرکزیت[۶]نسبت به کشف روابط معنایی بین اجزاء مختلف متن پرداخته شود. نظریه مرکزیت در اصل یک تابع حدس[۱۷]می‌باشد که براساس معیارهایی سعی در کشف روابط معنایی بین اجزاء متن را دارد. اما این روش دارای محدودیت‌هایی در اعمال برای انواع کاربردها می‌باشد. همچنین این روش به صورتی کاملا محلی رفتار می‌کند. از این رو سعی شده است تا اولا مدلی کاملا محاسباتی برای این نظریه ارائه گردد و همچنین این نظریه از حالت محلی خارج شود و به صورت عمومی رفتار نماید. همچنین در پایان این روش ارائه شده با استفاده از یک مدل استاندارد درستی یابی می‌گردد. با استفاده از دو رویکرد فوق می‌توان مشکلات مطرح شده در پیش را تا حدودی مرتفع نمود.

پیش از بررسی روش‌های پیشنهادی در این پایان نامه باید مروری بر پیش نیازها نیز داشت. برای درک بهتر ارتباط مباحث مطرح شده در این بخش مجددا مرور مختصری بر رویکرد کلی روش‌های پیشنهادی در این پایان نامه می‌نماییم. برای فائق آمدن بر مشکلات مطرح شده برای روش‌های موجود و همچنین ریشه‌های شناسایی شده برای این مشکلات رویکردهای اصلی این پایان نامه عبارتند از:

- عبور از فضاهای اقلیدسی و شبه آن و مدل کردن داده‌ها در فضاهای جامع‌تر که بتوانند تمام خواص داده را پوشش دهند
- استفاده از روش‌های مستقل از توزیع داده‌ها برای شناسایی میزان شباهت و یا عبارت دیگر تعیین میزان شباهت در تعبیری خارج از تعبیر رایج اقلیدسی و شبه آن
- استفاده از مشخصات زبان شناختی متن در تعیین میزان اهمیت اجزاء معنایی در یک بستر معنایی
- درستی یابی روش‌های ارائه شده با استفاده از یک مدل استاندارد

مجموعه این رویکردها می‌توانند منجر به ارائه راهکاری مناسب جهت استخراج مفاهیم گردند. در بخش ضمائم مروری بر پایه‌های اصلی همین رویکردهای اساسی که می‌توانند در درک بهتر بخش‌های بعد کمک نمایند صورت گرفته است.

۱,۳. نتیجه گیری

می توان این طور نتیجه گرفت که راه ورود به حوزه معنا به دست آوردن هستان شناسی در حوزه معنایی مورد نظر است. از سوی دیگر حوزه های معنایی مورد هدف بشر بسیار گسترده و در حال تغییر هستند.. به همین خاطر نمی توان انتظار داشت با روش های غیرخودکار بتوان به هستان شناسی های لازم در همه حوزه های معنایی دست یافت. واضح است که برای دستیابی به این مهم به روش های خودکار احتیاج داریم.

---

[1] Centering Theory [6]
[1] Heuristic Function [7]

از سوی دیگر حتی در صورت عدم دستیابی به هستان شناسی، دستیابی به مفاهیم یک حوزه معنایی می توانند کاربردهای زیادی در حوزه بازیابی اطلاعات داشته باشند.

مشکلات عمده ای که در این حوزه وجود دارند عدم استفاده از اطلاعات زبانی، زمان اجرای بالای روش های ارائه شده و استفاده از معیارهای کاملا هندسی برای تعیین میزان شباهت موجودیت های مختلف می باشد. این مسائل باعث شده اند تا نتوان به نتایج مناسبی در این حوزه دست یافت و البته با گذر از مشکلات ذکر شده می توان به راهکار مناسبی در این حوزه دست یافت.

۱,۴. ساختار پایان نامه

در ادامه پایان نامه در فصل دوم کارهای گذشته مرور گردیده‌اند. البته عمدتا در خصوص روش های اصلی به مرور پرداخته شده است. علت این امر نیز این است که بقیه روش ها مشتقاتی از این روش های اصلی می باشند و کارهای انجام شده در این پایان نامه گسترشی بر این روش های پایه نیستند و از همین رو بررسی بقیه مشتقات نمی تواند کارساز باشد. در فصل سوم روش ارائه شده برای اصلاح معیارهای مشابهت با استفاده از توابع ارتباط بیان گردیده است. این روش پیشنهادی سعی می نماید که فضای عمومی در حوزه تعیین میزان شباهت ها در حوزه بازیابی اطلاعات را تغییر دهد. عمومی سازی[۱۸]و سراسری سازی[۱۹]نظریه مرکزیت در فصل چهارم مورد بررسی قرار گرفته است. این فصل کاملا مبتنی بر روش های زبان شناختی می باشد و سعی می کند مدلی ریاضی برای بیان روش های شهودی در این حوزه ارائه نماید. در فصل پنجم نتایج عملی رویکردهای فوق مورد بررسی قرار گرفته‌اند. در این بررسی ابتدا هر یک از دو روش ارائه شده به صورت مجزا مورد آزمون قرار می گیرند و پس از آن یک پیاده سازی جامع از هر دو روش صورت گرفته است که نتایج آن در این فصل مرور شده‌اند. براساس تحلیل موجود برروی نتایج و همچنین نقاطی که هنوز امکان بهبود دارند، مجموعه کاملی از نتایج و کارهای آتی تهیه شده‌اند که در فصل ششم ارائه شده‌اند.

---

[1] Generalization [8]
[1] Globalization [9]

۲. مرور کارهای گذشته

با توجه به مطالب بیان شده در فصل قبل در خصوص اجزاء اصلی در استخراج هستان شناسی به صورت خودکار، کارهای مشابه باید در سه دسته مورد بررسی قرار گیرند. دسته اول روش های موجود در حوزه استخراج مفاهیم می باشند. در این دسته عمدتا توجه برروی روش های آماری برای تعیین مفاهیم در یک حوزه معنایی می باشد. البته علاوه بر مشکلاتی که ممکن است این روش ها به لحاظ ساختاری داشته باشند یک وابستگی مهم در حوزه تعیین مشابهت دارند. به عبارت دیگر این روش ها عمدتا از روش های مرسوم در تعیین مشابهت که عمدتا نیز هندسی می باشند استفاده می نمایند.

با توجه به توضیحات فوق لازم است تا در دسته دوم روش های تعیین مشابهت مورد بررسی قرار گیرند. به جهت روش نمایش اطلاعات در حوزه بازیابی اطلاعات که عمدتا روش های برداری می باشند، روش های هندسی تعیین مشابهت مورد استفاده قرار می گیرند. حتی بعضا روش های صرفا احتمالی به کار گرفته می شوند اما در نهایت برای دستیابی به پاسخ های نهایی مسائل بهینه سازی تعریف می شوند که عمدتا تابع هدف آن ها مبتنی بر روش های تعیین مشابهت هندسی هستند.

اما در آخرین دسته روش های تعیین میزان اهمیت لغت در صفحه مورد بررسی قرار می گیرند. عمدتا روش های مورد استفاده کنونی صرفا براساس اطلاعات آماری اهمیت لغت را تعیین می نمایند و همچنین همه این روش ها مبتنی بر فرض عدم اهمیت ترتیب لغات در صفحه هستند. همچنین از هیچگونه اطلاعات زبان شناختی در این حوزه استفاده نمی شود. مرور این روش ها می تواند به صورتی دقیق تر نقاط ضعف و نقاط مناسب برای تغییر را نشان دهد.

۲,۱. روش های استخراج مفاهیم

روش های متنوعی تا کنون در حوزه استخراج هستان شناسی ارائه گردیده اند. در این پایان نامه نیز هدف اصلی فعالیت برروی مؤلفه های تأثیرگذار در دقت اطلاعات می باشد و فعالیت به صورت مستقیم برروی روش ها نیست. از این رو ورود به جزئیات روش های مورد استفاده برای استخراج هستان شناسی دارای اهمیت چندانی در این پایان نامه نیست بلکه باید کلیات روش ها مورد بررسی قرار گیرند تا اصول مورد استفاده شناسایی شده و کلیات روش ها به لحاظ همخوانی با قالب کاری کلی در حوزه استخراج مفاهیم دارای تناقضات جدی نباشند. با این وجود سعی شده است تا حدودی نیز به توضیح در خصوص تغییراتی که برروی روش های اصلی انجام شده است پرداخته می شود. در حالت کلی این تغییرات در اصول روش ها نیست و تنها در بعضی از جزئیات تغییراتی صورت گرفته است و یا روش ها برای اجرای موازی و دیگر ملزومات محاسبات حجم بالا اصلاح گشته اند.

تقریبا تمام روش های این حوزه تحت عنوان روش های استخراج مدل نهان معنایی شناخته می شوند. می توان مهم ترین این روش ها را روش های LSI[۲۰]، PLSI[۲۱] و LDA[۲۲] دانست. در ادامه سعی می شود که به بررسی این روش ها و معرفی بعضی توسعه ها برروی این روش ها پرداخته شود.

LSI.۲,۱,۱

LSI یک روش مبتنی بر تحلیل طیفی[۲۳]می باشد. تقریبا می بتوان گفت که در ابتدای معرفی این روش هیچگونه تحلیل مشخص و مشروح در خصوص دلایل موفقیت آن وجود نداشت. اما پس از استفاده از این روش در کاربردهای مختلف مشاهده شد که این روش می تواند کارایی بسیار مناسبی در بسیاری از کاربردهای داده کاوی از خود نشان دهد.

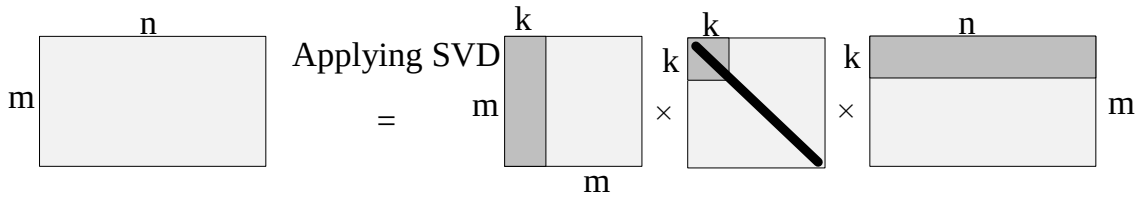
می توان گفت که مهم ترین عوامل ایجاد پیچیدگی در کاربردهای بازیابی اطلاعات ناشی از دو مفهوم هم معنایی[۲۴]و هم آیی[۲۵]می باشد. به عبارت دیگر به لحاظ محاسباتی شناسایی این دو مفهوم بسیار سخت است. در خصوص هم معنایی ها تقریبا مهم ترین فعالیت های انجام شده استفاده از روش های ریشه یابی[۲۶] و همچنین شبکه های معنایی لغات[۲۷]می باشد. اما هر دوی این روش ها تقریبا تنها می توانند هم معنایی

---

[0] Latent Semantic Indexing
[1] Probabilistic Latent Semantic Indexing
[2] Latent Dirichlet Allocation
[3] Spectral Analysis
[4] Synonymy
[5] Polysemy
[6] Stemming
[7] WordNet

ها را شناسایی نمایند و در خصوص هم آیی ها که نقش بسیار مهم تری در حوزه شناسایی روابط معنایی دارند با استفاده از روش های فوق نمی توان نتایج مناسبی را دریافت نمود.

LSI از یک روش تقسیم ماتریس به نام SVD[۲۸] استفاده می نماید. در این روش برای یک ماتریس داده شده بردارها و مقادیر ویژه ماتریس همبستگی محاسبه می شوند. یکی از مهم ترین ویژگی های مناسب این روش تولید مقادیر ویژه با ترتیب نزولی مقادیر است. به تبع این ویژگی بردارهای ویژه متناظر نیز به همین ترتیب تولید می شوند. از سوی دیگر در هر گام یک بردار ویژه تولید شده و در گام های بعد مقادیر آن ها تغییر نمی کند. از این رو می توان به لحاظ محاسباتی تنها تعداد معدودی از مهم ترین مقادیر را تولید نموده و بقیه مقادیر دارای اهمیت پایین تر تولید نشوند. این ویژگی کمک زیادی به کاهش زمان اجرا در مسائل واقعی می نماید. نمای کلی این روش را می توان در زیر مشاهده نمود



شکل ۲ – نمایی کلی از الگوریتم SVD

از دیدگاه ریاضی این بردارهای ویژه اجزاء متعامد ماتریس اولیه را نمایش می دهند. به لحاظ معنایی برای یک ماتریس داده شده این اجزاء متعامد در اصل حوزه های معنایی مستقل را نمایش می دهند. استقلال به معنای عدم ارتباط نیست بلکه منظور از آن حوزه های معنایی است که می توانند نمایندگی یک معنای کامل را انجام دهند. به عبارت دیگر این حوزه های معنایی هر یک نماینده یک مفهوم هستند. مسلما این مفاهیم دارای ارتباطاتی با هم هستند اما به تنهایی دارای معنای مشخصی نیز هستند.

همان طور که در بالا توضیح داده شد است مهم ترین وجه تسمیه روش LSI را می توان توانایی آن در شناسایی روابط معنایی پنهان در متن دانست. این ویژگی کمک می نماید تا به خوبی بتوان رفتارهای پویای متن شناسایی شوند. به عبارت دیگر ابزارهایی مانند ریشه یاب و شبکه معنایی لغات و دیگر ابزارهای تولید شده توسط انسان تنها ویژگی های ایستای زبان را شناسایی می نمایند. اما در حوزه داده کاوی و در کاربردهای عمومی که در حوزه های معنایی گوناگونی عمل می نمایند زبان ویژگی های بسیار پویایی را از خود نشان می دهد که در هر بار اجرای الگوریتم ها باید بررسی گردند.

---

[2] Singular Value Decomposition

یکی دیگر از ویژگی های مهم LSI دسته بندی همزمان اسناد و لغات می باشد. به عبارت دیگر پس از اجرای LSI هم لغات و هم اسناد به لحظا معنایی دسته بندی می گردند. البته این دسته بندی در کاربردهای خوشه بندی اسناد کارایی مناسبی از خود نشان نداده است اما در کاربردهای پاسخ گویی به پرس و جو این روش توانسته است کارایی بسیار مناسبی از خود نشان دهد.

در مقابل این ویژگی ها این روش دارای ضعف هایی نیز هست که مهم ترین مورد آن وابستگی به تابع توزیع داده ها می باشد. این روش برای داده هایی که توزیع گاوسی ندارند دارای میزان انتشار خطای بالایی بوده و دقت مناسبی ندارد [HOF99]. مهم ترین علت این وابستگی به تابع توزیع را می توان استفاده از ضرب داخلی برای محاسبه همبستگی دانست. رفتار ضرب داخلی را می توان بسیار شبیه به ضریب همبستگی Pearson دانست که یک ضریب همبستگی وابسته به تابع توزیع می باشد [NEL98].

یک مسئله مهم دیگر که در استفاده از LSI باید مورد توجه قرار گیرد تعیین تعداد مفاهیم است. به عبارت دیگر تعداد مفاهیم مستقل در یک مجموعه از اسناد محدود است و باید تنها تعداد مشخصی از بردارهای یکه شناسایی شده را به عنوان مفاهیم استفاده نماییم. برای تعیین تعداد این مفاهیم روش های مختلفی ارائه گردیده اند. یکی از این روش ها مبتنی بر اندازه مقادیر یکه تولید شده می باشد [GAO11]. در این روش نقطه ای که یک تغییر عمده بین مقادیر یکه رخ می دهد شناسایی شده و تنها مقادیری بزرگ تر از آن نقطه به عنوان نمایندگان مفاهیم انتخاب می شوند. این روش توانسته است نتایج خوبی را از خود نشان دهد. روش های دیگری نیز ارائه شده اند که بعضا مبتنی بر روش های یادگیری می باشند []. اما روش ارائه شده در [GAO11] به لحاظ سادگی و کارایی مناسب یکی از بهترین روش های ارائه شده است. در عین حال بعضی از تحقیقات نشان داده اند که مقادیر نسبتا ثابتی برای انتخاب تعداد مفاهیم براساس اندازه داده ها و ماهیت آن ها قابل تعیین هستند. این اعداد بین ۲۰۰ تا ۸۰۰ بوده و نشان داده شده است که مقادیری در این بازه بهترین کارایی را برای انواع مجموعه های داده نمایش می دهند [].

توسعه های مختلفی برروی LSI تا کنون اعمال شده اند. به عنوان مثال در [KOL98] به جای SVD از SDD[۲۹] جهت تقسیم ماتریس استفاده شده است. این روش توانسته است رفتارها و کارایی مناسب تری را نسبت به SVD از خود نشان دهد. همچنین در توسعه دیگری برروی LSI سعی شده است تا روشی مناسب جهت ماتریس های تنک[۳۰] ارائه گردد تا با حداقل زمان بتوان به کارایی مورد نظر دست یافت [MAV07].

---

[2] Semi-Discrete Matrix Decomposition
[3] Sparse Matrices

PLSI .۲,۱,۲

همانطور که گفته شد یکی از مهم ترین مشکلات موجود در LSI وابستگی آن به تابع توزیع می باشد. این مشکل در [HOF99] مورد توجه قرار گرفته و بر همین اساس روشی با نام PLSI[۳] ارائه شده است که مهم ترین ویژگی آن عدم وابستگی به تابع توزیع داده ها می باشد. نگاه PLSI کاملا مبتنی بر احتمالات است. از این رو برای هر لغت یک متغیر تصادفی به نام t_i، برای هر سند یک متغیر تصادفی به نام d_j و برای هر مفهوم نیز یک متغیر تصادفی به نام z_k را تشکیل می دهد. در این دیدگاه میزان ارتباط یک لغت و یا سند با یک مفهوم با استفاده از احتمال های شرطی $P(t_i|z_k)$ و $P(d_j|z_k)$ نمایش داده می شوند. بر همین اساس و با استفاده از قوانین بیز مسئله بهینه سازی زیر ساخته شده است که حل آن منجر به تعیین مقادیر احتمال های شرطی فوق می باشد. در این مسئله بهینه سازی هدف حداکثر نمودن تابع Log-likelihood زیر است

$$\mathcal{L} = \sum_{d \in D} \sum_{t \in T} n(d,t) \log P(d,t)$$

که در آن

$$P(d,t) = \sum_{z \in Z} P(z) P(d|z) P(t|z)$$

با استفاده از ماتریس لغت-صفحه می توان مقدار $P(t_i|d_j)$ را به دست آورد اما مقدار $P(t_i,d_j)$ مشخص نیست. در نهایت این مسئله با استفاده از روش EM[۳۲] حل می شود. مقادیر $P(d|z)$ و $P(t|z)$ نیز با استفاده از قوانین بیز و براساس مقادیر موجود $P(t_i|d_j)$ محاسبه می شوند. با استفاده از این مقادیر محاسبه شده میزان تعلق هر لغت و یا سند به هر مفهوم تعیین شده و می توان دسته بندی معنایی برروی داده ها را انجام داد.

همچنین برای بهبود زمان اجرا تغییراتی برروی مسئله صورت گرفته است و با استفاده از Tempered EM نسبت به حل مسئله اقدام گردیده است. مهم ترین ویژگی PLSI را می توان عدم وابستگی به تابع توزیع دانست. این روش توانسته است به لحاظ پارامترهای ارزیابی precision و recall به میزان قابل توجهی نسبت به LSI از خود بهبود نشان دهد. همچنین این روش دارای Perplexity بسیار مناسب تری نسبت به LSI می باشد. این معیار نشان می دهد که نتایج به دست آمده برای یک نمونه داده تا چه حد قابل تعمیم به شرایط و مجموعه داده های دیگر می باشند.

---

[3] Probabilistic Latent Semantic Indexing
[3] Expectation Maximization [2]

اما در این بین مشکلاتی نیز برای این روش قابل بیان است. از جمله اینکه تعداد متغیرهای تولید شده توسط این روش در حدود $|D| \times |Z| + |T| \times |Z|$ می باشد که برای مسائل واقعی بسیار زیاد است. این تعداد متغیر مشکلات جدی را در حوزه پیچیدگی محاسباتی ایجاد می نمایند و از این رو یکی از مهم ترین مشکلات این روش می باشد. از سوی دیگر EM نیز یک روش تکراری[۳۳] می باشد و از این رو امکان پیاده سازی های موازی و توزیع شده موفق برای آن موجود نیست. که این مسئله نیز می تواند مانعی بر سر راه دستیابی به کارایی مناسب محسوب گردد.

یکی دیگر از مشکلات این روش تعیین تعداد مفاهیم است. مانند LSI ما باید تخمینی در خصوص تعداد مفاهیم داشته باشیم. در LSI روش هایی مورد استفاده قرار می گرفتند که در حین محاسبه الگوریتم امکان تعیین تعداد مفاهیم مورد نیاز وجود داشت. به عبارت دیگر در LSI می توان شرط توقف مشخصی را برای الگوریتم تعیین نمود. اما در خصوص PLSI چنین امکانی در دست نیست و باید صرفا براساس مقادیر تجربی و همچنین روش های مبتنی بر یادگیری اقدام به تعیین تعداد مفاهیم نمود که می تواند یکی از نقاط ضعف روش محسوب گردد.

LDA .۲,۱,۳

در این روش سعی شده است تا با استفاده از همان مفاهیم مورد استفاده در PLSI نسبت به شناسایی داده های مخفی اقدام گردد. اما در این خصوص سعی شده است تا حد ممکن نسبت به کاهش تعداد متغیرهای تصادفی اقدام گردد. در اصل می توان گفت LDA جزء سازوکارهای مولد می باشد. این سازوکارها عمدتا به دنبال ایجاد مدل هایی هستند که با استفاده از آن ها بتوان مشخصات غیرقابل مشاهده از داده ها را با استفاده از مشخصات قابل مشاهده تولید نمایند. در خصوص LDA این مشخصات غیرقابل مشاهده در اصل همان دسته های داده هستند.

به صورت کلی روش در مفروضات اولیه شبیه به PLSI است. به عبارت دیگر در LDA نیز فرض می شود که لغات و اسناد و مفاهیم متغیرهایی تصادفی هستند و هدف یافتن احتمالات شرطی این متغیرها نسبت به هم می باشد. این احتمالات شرطی تعیین کننده میزان ارتباط این متغیرها با هم خواهند بود. در خصوص فرضیات اولیه LDA فرض می کند که متغیرهای تصادفی مفاهیم دارای توزیع Dirichlet هستند. یکی از مفروضات این روش که در اکثر روش ها نیز مورد استفاده قرار می گیرد این است که هر سند شامل مجموعه ای از کلمات است و ترتیب رخداد در بین آن ها دارای اهمیت نمی باشد. به عبارت دیگر کلمات در یک سند جابجاپذیر هستند. یک تفاوت دیگر این روش با PLSI در این است که در PLSI هم لغات و هم

---

[3] Iterative [3]

اسناد را متغیرهایی ترکیبی بر اساس مفاهیم فرض می کند اما LDA اسناد را ترکیبی از متغیرهای تصادفی مفاهیم و مفاهیم را ترکیبی از متغیرهای تصادفی لغات می داند. الگوریتم دارای شکل کلی زیر می باشد:

شبه کد ۲ - الگوریتم LDA

1. Choose $N \sim \text{Poisson}(\xi)$.
2. Choose $\theta \sim \text{Dir}(\alpha)$.
3. For each of the N words w_n:
   (a) Choose a topic $z_n \sim \text{Multinomial}(\theta)$.
   (b) Choose a word w_n from $p(w_n \mid z_n, \beta)$, a multinomial probability conditioned on the topic z_n.

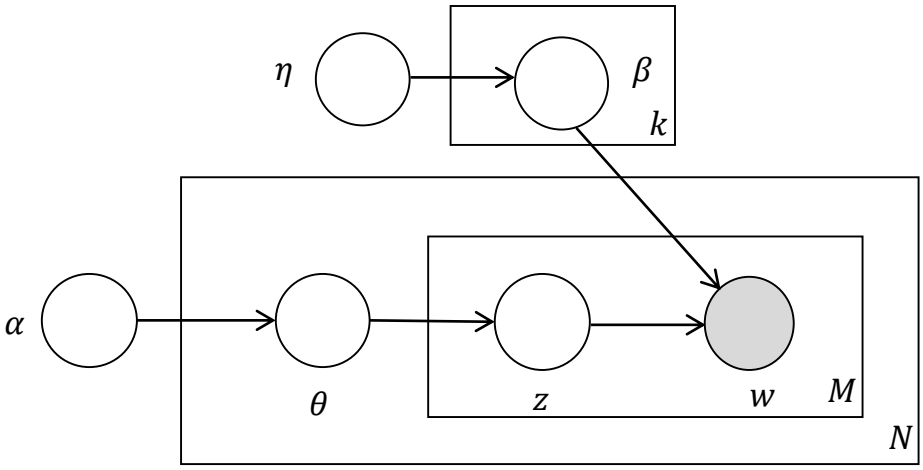
این الگوریتم باید به ازای هر سند اجرا شود. در این الگوریتم طول یک سند با N نشان داده شده است و خط اول الگوریتم تنها برای تعیین این نکته است که طول اسناد دارای توزیع پوآسون است. البته این فرض هیچ تأثیری در الگوریتم ندارد. تعداد مفاهیم k فرض شده است و البته تعیین این تعداد در اینجا مورد بحث نیست. همچنین ماتریس β در اصل ماتریس احتمال شرطی رخداد هر لغت به شرط رخداد هر مفهوم است که جداگانه باید این ماتریس محاسبه شود و به عنوان پارامتر ورودی این الگوریتم مورد استفاده قرار می گیرد.

از سوی دیگر بردار θ که دارای k عضو می باشد براساس یک بردار ورودی به نام α محاسبه می شود. θ در اصل پارامتر ورودی تابع توزیع چندجمله ای مفاهیم می باشد و باید براساس ماتریس لغت صفحه محاسبه شود.

براساس این پارامترهای ورودی احتمال دوجمله ای لغات و مفاهیم نسبت به هم به صورت زیر است.

$$p(\theta, z, w|\alpha, \beta) = p(\theta|\alpha) \prod_{n=1}^{N} p(z_n|\theta) p(w_n|z_n, \beta) \qquad (۲)$$

مدل کلی استنتاج در LDA را می توان به صورت شکل ۳ ارائه نمود.



شکل ۳ - ارائه مدل کلی استنتاج در LDA براساس روش ارائه گرافی

برای محاسبه α و β نیز به این صورت عمل می شود که معادله Log-Likelihood نوشته می شود و با استفاده از روش EM این متغیرهای تصادفی تخمین زده می شوند. این معادله بهینه سازی به صورت زیر است

$$l(\alpha,\beta)=\sum_{d=1}^{M}\log p(w_d|\alpha,\beta) \qquad (۳)$$

آنچه مسلم است این روش علیرغم تمامی بهبودهایی که در نتایج نسبت به روش های مشابه از خود نشان می دهد اما دارای ضعف زمان اجرا می باشد و دارای زمان اجرای بسیار بالایی است.

۲,۱,۴. نتیجه گیری

در این بخش به مرور روش های اصلی در حوزه استخراج دانش نهان در ماتریس های لغت صفحه پرداخته شد. آنچه به صورت خاص به آن توجه شد سه عامل اصلی زیر هستند:

- زمان اجرای الگوریتم ها
- عدم وابستگی به توزیع داده ها
- حجم پارامترهای ورودی روش برای اجرا

مقایسه کاملی از این جنبه ها در فصل ۵ صورت خواهد گرفت. روش ارائه شده در فصل ۳ نیز عمدتا بر همین عوامل تکیه دارد و سعی می کند تا این زوایا را به خوبی پوشش دهد. البته مسائل دیگری نیز در بررسی روش ها باید مورد نظر قرار گیرند. از آن جمله می توان به سادگی پیاده سازی و همچنین امکان پیاده سازی توزیع شده و همچنین موازی روش ها اشاره نمود. واضح است که هر چه یک الگوریتم ساده تر باشد و بیشتر مبتنی بر اجزاء شناخته شده موجود باشد می تواند از این منظر مناسب تر باشد. البته به علت

وسعت فعالیت ها جهت مقایسه از منظر پیاده سازی موازی این بخش در پایان نامه پوشش داده نشده است و خود می تواند به عنوان یک حیطه تحقیق مناسب مورد بررسی قرار گیرد.

۲,۲.روش‌های موجود در تعیین مشابهت

توابع محاسبه مشابهت کاربردهای فراوانی در دسته بندی، خوشه بندی و کاربردهای بازیابی اطلاعات دارند. همچنین می توان گفت که این روش ها تأثیر به سزایی نیز در نتیجه حاصل خواهند داشت. علت این امر به دو وجه مختلف باز می گردد. اولین وجه این است که در هر زمینه کاربردی ویژگی های تأثیرگذار در مشابهت متفاوت هستند. به عنوان مثال در فصل قبل در خصوص ویژگی های مورد نیاز در حوزه متن کاوی توضیح داده شد. از سوی دیگر در هر حوزه کاری نیز هر تابع مشابهت می تواند برای دسته ای خاص از کاربردها مفید واقع گردد.

با توجه به این توضیحات آشنایی کامل با انواع توابع مشابهت و همچنین ویژگی های هر یک می تواند راهگشای انتخاب تابع مناسب برای کاربردی خاص گردد. در این بخش قصد داریم تا مروری جامع برروی انواع روش های مشابهت داشته باشیم و در یک جمع بندی به بررسی مشکلاتی که پیش از این برای روش های مشابهت جاری مطرح کردیم پرداخته شود.

پیش از بررسی این روش ها باید به نکته ای اشاره گردد. معیارهای مورد بررسی به دو دسته اصلی تقسیم می شوند. دسته اول معیارهای فاصله[۳۴]و یا به عبارت دیگر همان سنجه[۳۵]ها هستند. این دسته از معیارهای دارای خواص مورد نظر برای سنجه ها می باشند. دسته دوم که دارای خواص مورد نیاز برای یک سنجه نمی باشند را عمدتا معیارهای شباهت[۳۶]و یا معیارهای واگرایی[۳۷]می گویند.

دو مدل مورد وثوق در نمایش داده ها در معیارهای فاصله/شباهت مدل های نمایش احتمالی و برداری می باشند. تقریبا به راحتی می توان هر یک از این دو روش نمایش را به دیگری تبدیل نمود که از این نظر نگرانی وجود نخواهد داشت. در ادامه براساس پایه تولید هر یک از این معیارهای فاصله/مشابهت به ارائه آن ها پرداخته می شود.

---

3 Distance Measure 4
3 Metric 5
3 Similarity Measure 6
3 Divergence Measure 7

۲,۲,۱. خانواده L_p

این دسته از معیارها ایده اصلی خود را از معیار فاصله اقلیدسی اخذ نموده اند. البته در حوزه رفتار ممکن است کاملا متفاوت از معیار اقلیدسی رفتار نمایند. می توان گفت که عامل مشترک در این خانواده از معیارها $|P_i - Q_i|$ می باشد. اما عموما معیارهای مبتنی بر خواص هندسی تا حدود زیادی رفتارهای مشابهی را نمایش می دهند. در ضمائم لیستی از این معیارها آورده شده اند.

۲,۲,۲. خانواده L_1

دسته ای از معیارها بر پایه قدر مطلق تفاضل مؤلفه های بردارهای مورد بررسی ایجاد گردیده اند. به عبارت دیگر عامل اصلی در این معیارها $|P_i - Q_i|$ می باشد. به عنوان مثال معیار Sørensen به صورت گسترده در حوزه بوم شناسی مورد استفاده قرار می گیرد. معیار Gower نیز بر اساس قدر مظلق طراحی شده است با این تفاوت که بردارها نورمال نموده و بر این اساس فاصله را محاسبه می نماید. معیار Canberra نیز شکلی تغییر یافته از Sørensen می باشد که نسبت به تغییرات در نزدیکی مقدار صفر بسیار حساس می باشد. در ضمائم لیستی از این معیارها آورده شده اند.

۲,۲,۳. خانواده اشتراک

این دسته نیز بسیار شبیه به دسته L_1 می باشند و دارای عامل مشترک مشابهی می باشند. در ضمائم لیستی از این معیارها آورده شده اند.

۲,۲,۴. خانواده ضرب داخلی

این دسته از معیارها مبتنی بر ضرب داخلی بردارها می باشند. در بیان ساده تر عامل اصلی در این دسته از معیارها $P_i.Q_i$ می باشد که به لحاظ رفتاری بسیار شبیه به معیارهای اقلیدسی می باشد. در ضمائم لیستی از این معیارها آورده شده اند.

۲,۲,۵. خانواده مجذور وتر

این دسته از معیارها مبتنی بر جذر ضرب مؤلفه ها می باشند و همگی را می توان به صورت تابعی از معیار مجذور وتر محاسبه نمود. در این خانواده نیز عامل $P_i.Q_i$ نقش اصلی را ایفا می نماید. در ضمائم لیستی از این معیارها آورده شده اند.

۲,۲,۶. خانواده مجذور L_2

این دسته مبتنی بر مجذور فاصله L_2 هستند و همگی دارای این مؤلفه می باشند. در این دسته نیز عامل $(P_i - Q_i)^2$ نقش بسیار مهمی دارد که بسیار شبیه به عامل $|P_i - Q_i|$ در معیارهای اقلیدسی رفتار می

نماید. در بین این معیارها، معیارهای نامتقارن نیز موجود می باشند. در ضمائم لیستی از این معیارها آورده شده اند.

۲,۲,۷. خانواده بی نظمی شانون[۳۸]

این دسته از معیارها همگی دارای مؤلفه بی نظمی می باشند و از این مؤلفه برای تعیین میزان فاصله استفاده نموده اند. به عبارت دیگر عامل اصلی در این دسته از معیارها $P_i \ln P_i$ می باشد. این دسته از معیارها نیز دارای شباهت رفتاری بسیار بالایی با خانواده مجذور وتر می باشند. در ضمائم لیستی از این معیارها آورده شده اند.

همچنین در [DEZ09] از میزان شباهت نتیجه فشرده سازی Ziv-Lempel به عنوان معیاری برای مقایسه دو رشته استفاده شده اطول رشته حاصل از این نوع فشرده سازی دارای ارتباط مستقیمی با بی نظمی شانون می باشد و از این رو این روش به عنوان یکی از روش های مبتنی بر بی نظمی شناخته می شود.

۲,۲,۸. معیارهای ترکیبی

در این دسته از معیارها ترکیبی از مؤلفه های مختلف مورد استفاده قرار گرفته اند. با توجه به مؤلفه های از پیش تعریف شده می توان دریافت که هر یک از این معیارهای ترکیبی شامل کدام مؤلفه ها می باشند. در ضمائم لیستی از این معیارها آورده شده اند.

۲,۲,۹. نتیجه گیری

در این بخش لیستی کامل از انواع روش های تعیین شباهت ارائه شد. در این بررسی انواع مؤلفه های مختلف قابل استفاده برای تعریف شباهت آورده شده اند که هر یک می توانند کاربردهای منحصر به خود را داشته باشند. اما بررسی رفتاری این معیارها نیز می تواند خالی از لطف نباشد. این بررسی می تواند به خوبی نشان دهد که میزان شباهت رفتاری این معیارها به چه میزان است و اینکه کدام معیارها از مشکلات مشابهی رنج می برند. در [DEZ09] چنین بررسی صورت گرفته است. در این بررسی یک مسئله ساخته شده و سپس همبستگی بین نتایج به دست آمده از توابع فاصله مختلف در مورد این مسئله محاسبه شده اند. نتایج زیر از این بررسی قابل استخراج هستند

- معیارهای مبتنی بر ضرب داخلی دارای همبستگی بسیار بالایی با معیار فاصله اقلیدسی و معیارهای هم خانواده آن می باشند.

---

[3] Shannon's Entropy [8]

- متوسط عدم شباهت کلیه معیارهای ذکر شده با یکدیگر در حدود ۰,۱۵ می باشد. این عدد نشان دهنده رفتارهای بسیار مشابه بین این معیارها می باشد. در هر مورد نیز سعی شده تا شباهت های رفتاری نزدیک تر بیان گردند.

با توجه به این نتایج می توان گفت که اکثر معیارهای مشابهت ارائه شده دارای مشکلاتی که پیش از این در خصوص معیارهایی مانند فاصله کسینوسی و اقلیدسی بیان شدند هستند. از این رو لزوم توجه به روش هایی که بتوانند فارق از چنین مشکلاتی به تعیین میزان مشابهت بین متغیرهای تصادفی کمک نمایند کاملا واضح می نماید.

۲,۳. روش های تعیین میزان اهمیت

برای ساده سازی قالب کاری مورد استفاده در وزن دهی لغات معمولا وزن یک لغت را حاصل سه عامل اصلی می دانند. به عبارت دیگر وزن یک لغت مانند t_i در یک متن مانند d_j با استفاده از ضرب این عوامل به دست می آید که برابر خواهد بود با $LW_{i,j}.GW_i.NF_j$. این فاکتورها عبارت خواهند بود از

- وزن محلی: این وزن براساس اطلاعات محلی حضور یک لغت در یک متن محاسبه می گردد.
- وزن سراسری: این وزن براساس اطلاعات رخداد یک لغت در متون مختلف در یک پیکره مورد استفاده محاسبه می شود.
- عامل نرمال سازی: این عامل وظیفه نرمال نمودن وزن های محلی محاسبه شده را برعهده دارد تا وزن های محاسبه شده برای یک لغت در متون مختلف قابل مقایسه با هم باشند.

این تقسیم بندی تقریبا در تمام روش های وزن دهی به لغت مشترک هستند. روش های وزن دهی به لغات به یک دسته بندی اصلی تقسیم می شوند و آن هم تقسیم به روش های آماری و غیر آماری است. در روش های آماری پایه اصلی استفاده از اطلاعات خداد یک لغت در یک متن است. به عبارت دیگر بسامد نقش اصلی را در این روش ها برعهده دارد. در عوض روش های غیرآماری سعی می کنند از منابع دیگر اطلاعات علاوه بر بسامد برای محاسبه وزن استفاده نمایند. این منابع می توانند اطلاعات نهفته دیگر داخل متن و یا یک مجموع دانش پیش زمینه باشد که فراهم گردیده است.

یکی از معمول ترین روش های وزن دهی به لغات را می توان روش TFiDF دانست. تعریف این روش براساس قالب کاری کلی ذکر شده در فوق به شکل زیر می باشد:

- $LW_{i,j}$ در TFiDF عبارت است از تعداد رخداد یک لغت مانند t_i در یک متن مانند d_j.

- GW_i در TFiDF عبارت است از $\log\frac{|D|}{|d_k:d_k\in D,t_i\in d_k|}$ که در اینجا منظور از D مجموعه پیکره مورد استفاده است
- NF_j در TFiDF عبارت است از $\frac{1}{L_j}$ که L_j طول متن d_j می باشد.

به علت مشکلاتی که در روش های مبتنی بر آمار وجود دارد تمرکز اصلی این پایان نامه برروی این روش ها نیست. اما از جهت کامل بودن مطلب در ادامه به مروری بر روش های آماری موجود خواهیم پرداخت. علاوه بر آن در روش های غیرآماری نیز دو دسته روش های مبتنی بر دانش پیش زمینه و روش های مستقل از دانش پیش زمینه را بررسی خواهیم نمود.

۲,۳,۱. روش های آماری

در ابتدا باید برای توضیح در مورد این روش ها یک قرارداد نمایشی را ارائه نمود. در این قرارداد فرض کنید کلمه ای مانند t_k را در اختیار داریم. در چنین وضعیتی تمام کلمات دیگر در مجموعه داده ها با $\overline{t_k}$ نمایش داده می شوند. همچنین فرض کنید که خوشه هایی از اسناد را که کلمه t_k متعلق به آن ها می باشد با c_i نمایش دهیم. به صورت قبل تمام خوشه های دیگر اسناد را که کلمه مذکور متعلق به آن ها نیست با $\overline{c_i}$ نمایش خواهیم داد. در چنین حالتی اسناد براساس رخداد کلمه مورد نظر و یا کلمات دیگر در هر یک از این دو نوع خوشه ها به صورت زیر تقسیم می شوند و تعداد هر یک با متغیرهای ذکر شده در جدول زیر نمایش داده می شوند [LAN06].

جدول ۱ – متغیرهای اصلی در محاسبه وزن لغات در روش های متداول

	t_k	$\overline{t_k}$
Positive clusters c_i	a	b
Negative Clusters $\overline{c_i}$	c	d

به عنوان مثال عامل idf براساس مقادیر بالا به صورت زیر محاسبه می شود:

$$\log(\frac{a+b+c+d}{a+c}) \qquad (۴)$$

عمدتا این روش ها شامل دو دسته می شوند: روش های بدون نظارت، روش های دارای نظارت. روش های بدون نظارت عمدتا روش های معمول وزن دهی می باشند. اولین نمونه آن وزن دهی دودویی می باشد که در کاربردهایی مانند درخت تصمیم و بسیاری زمینه های دیگر کاربرد دارد اما در دیگر کاربردهای متن کاوی عمدتا از آن استفاده نمی شود. روش دیگر استفاده از بسامد کلمه می باشد. البته روش های مشابه دیگری نیز مانند $\log(1+tf)$ و یا ITF [LEO02] نیز ارائه شده اند که در [LAN06] نشان داده شده

است که هر سه این روش ها نتایج مشابهی را در خوشه بندی مبتنی بر SVM تولید می نمایند. در این بین روش بسیار معروف *tfidf* نیز جزء همین دسته از روش ها قرار می گیرد. نمونه های تغییر یافته ای از این روش نیز ارائه شده اند که از آن جمله می توان به $\log(tf) \cdot idf$، *tf.idf-prob* و *term relevance* اشاره نمود [WU81]. در [LAN06] نشان داده شده است که این نمونه های تغییر یافته تفاوت زیادی با نسخه اصلی *tfidf* ندارند و نتایج حاصل بسیار نزدیک به هم می باشد.

دسته بعدی، روش های دارای ناظر می باشند. به عنوان مثال اگر از یک روش خوشه بندی برای وزن دهی استفاده کنیم، روش را دارای ناظر می دانیم، زیرا روش از یک دانش پیش زمینه استفاده نموده است. این دسته خود عمدتا به دو قسمت تقسیم می شوند. روش هایی که این دانش پیش زمینه را با روش های بدون نظارت ترکیب می کنند و روش هایی که تنها با استفاده از دانش پیش زمینه وزن دهی را انجام می دهند.

در دسته اول روش های مبتنی بر نظارت عمدتا فاکتور *idf* با یکی از روش های انتخاب ویژگی[۳۹] که در روش های خوشه بندی کاربرد دارند جایگزین می شود. در یک نگاه متفاوت می توان *idf* را به نوعی یک معیار انتخاب ویژگی دانست. در اصل *idf* تعیین می نماید که از بین کلمات موجود کدامیک می توانند در کاربردهای مختلف اطلاعات مناسبی در اختیار قرار دهند و کدام کلمات حضور و یا عدم حضورشان اطلاعات زیادی در اختیار نخواهد گذاشت. از این رو جایگزین نمودن معیارهای پیشرفته تر انتخاب ویژگی شاید بتوانند مفیدتر واقع شوند. از این جمله می توان به [DEN04] اشاره نمود که از معیار χ^2 به جای *idf* استفاده نموده است و نشان داده است که با این جایگزینی نتایج بهتری در حوزیه خوشه بندی با استفاده از SVM به دست می آید. همچنین [DEB03] از هر سه معیار χ^2، *Information gain* و *gain ration* استفاده نموده است و نشان داده است که در کاربردهای متنوع نمی توان برتری قابل توجهی برای این روش ها نسبت به *tf.idf* متصور بود. البته [MOR02] از *gain ration* در کاربرد خلاصه سازی متن استفاده نموده است و توانسته است نتایج بسیار بهتری نسبت به حالت معمول به دست آورد.

در دسته دوم از روش های مبتنی بر ناظر مستقیما از روش های خوشه بندی عمدتا خطی برروی یک مجموعه داده یادگیری استفاده می شود. براساس نتایج این خوشه بندی کلمات دارای اهمیت شناخته شده و به آن ها وزن بیشتری در داده های عملیاتی داده می شود. به عنوان مثال [MLA04] سه روش *Information Gain*، *Odds Ration* و استفاده از خوشه بندی خطی را با هم مقایسه نموده است و نشان داده است که استفاده از روش خوشه بندی برای وزن دهی در کاربردهای خوشه بندی نتایج بهتری را به همراه خواهد داشت. همچنین [LAN06] روشی را به نام *tf.rf* ارائه نموده است که *rf* برابر است با

---

[3] Feature Selection [9]

$\log(2+\frac{a}{c})$ و توانسته است نتایج بهتری را نسبت به دیگر روش های توضیح داده شده در این دسته در حوزه کاربردهای خوشه بندی به دست آورد.

۲,۳,۲. روش های مبتنی بر دانش پیش زمینه

دلیل عمده استفاده روش های وزن دهی از دانش پیش زمینه، امکان تقویت معنایی واحدهای معنایی مانند جمله های و یا حتی خود لغات به لحاظ دایره لغات می باشد. عموما واحدهای معنایی مورد استفاده چنان کوچک هستند که نمی توانند دانش زیادی را در خصوص معنای مورد نظر با خود حمل نمایند از این رو لازم است تا این اجزاء معنایی از طرقی تقویت گردند تا بتوانند دانش بیشتری را در اختیار قرار دهند.

در این دسته، روش وزن دهی ممکن است به دو دلیل احتیاج به دانش اولیه داشته باشد. روش های وزن دهی در این دسته، یا از طریق دانش اولیه به غنی سازی مجموعه داده اولیه و بسط مجموعه کلمات به مجموعه بزرگتری از کلمات می پردازند و یا روابط میان کلمات را از طریق جستجوی این کلمات در مجموعه دانش اولیه استخراج می کنند. یکی از اولین نمونه های این دسته کاری است که در [MOR88] و [MOR91] انجام شده است. در این پژوهش از زنجیره های لغوی به عنوان محور اصلی شناسایی میزان ارتباط یک کلمه با محتوای معنایی متن استفاده نموده اند. یک زنجیره ی لغوی در اصل به بیان ارتباطات معنایی لغات در چهار دسته یکسان، هم معنی، ابر معنی و زیر معنی می پردازد. روش تعیین این ارتباطات استفاده از یک فرهنگ لغات دارای لینک بین کلمات[۴۰]است. [MOR91] از Roget به عنوان چنین فرهنگ لغتی در کار خود استفاده کرده اند. البته فرهنگ لغتی مانند Roget امکان خودکارسازی انجام فرآیندها را ندارد (بر خلاف WordNet). از سوی دیگر [HIR98] برای خودکار سازی این فرآیند از wordnet به عنوان فرهنگ لغت برخط استفاده کردند. با این تفاوت که به علت تمایز در نوع روابط تعریف شده در آن با قوانین تشکیل زنجیره های لغوی، این قوانین را تغییر داده و فقط بر اساس اسامی، زنجیره های لغوی خود را تشکیل دادند.

همچنین [CAR02] از همین روش استفاده نمود و نشان داد که سیستم تولید شده با نام Lex Track نتایج بهتری را در زمینه ی پیگیری اخبار نسبت به سیستم KeyCos از خود نشان می دهد. همچنین در حوزه ی خلاصه سازی نیز [BAR97] با ترکیب همین روش با wordnet، یک الگوریتم برچسب زنی نحوی، پارسر کم عمق[۴۱]و یک الگوریتم قسمت بندی به نتایج خوبی دست یافتند.

---

[4] Thesaurus [0]

[4] - Shallow parser [1]

در [KAN05] نیز از زنجیره های لغوی برای خوشه بندی متن و سپس وزن دهی به لغات استفاده شده است. روش کار به این صورت است که ابتدا با تشکیل زنجیره های لغوی، دسته های کلمات تشکیل می شوند. این دسته ها در اصل کلمات را به لحاظ معنایی خوشه بندی می نمایند. پس از تشکیل این زنجیره ها و روابط بین لغات، هر خوشه با روشی دقیقاً مشابه با رتبه بندی صفحات[۴۲]و با استفاده از مفهوم hub رده بندی می شوند. تفاوت در این است که بر خلاف لینکهای صفحات وب، روابط بین لغات بر اساس نوع آنها دارای تأثیر متفاوتی در انتشار وزن لغات در بین یکدیگر دارند. در نهایت خوشه هایی که مجموع وزن لغات آن ها از میانگین وزن دیگر خوشه ها با اعمال ضریب تجربی α بیشتر باشد، به عنوان خوشه های اصلی انتخاب و لغات بر اساس میزان ارتباط با این خوشه ها رده بندی می شوند. این روش نسبت به استفاده از TF-IDF به میزان ۱۵٪ افزایش دقت و ۸۰٪ کاهش حجم در نگهداری اندیس صفحات دارد.

همچنین در [BAD09] روشی مبتنی بر توزیع کلمات در اسناد برای وزن دهی به کلمات ارائه شده است. [CHA08] نیز روشی ارائه کردند که میزان اهمیت کلمات را بر اساس دسته بندی موجود بر روی آنها تعیین می نماید. [WAN08] نیز روشی ارائه کردند که با استفاده از دانش پیش زمینه موجود در Wikipedia یک کرنل معنایی[۴۳]ایجاد می گردد و از این طریق اسناد غنی می گردند. سپس از این طریق وزن دهی به لغات با دقت بالاتری صورت می گیرد. روش های مشابهی با استفاده از wordnet برای غنی سازی اسناد ارائه گردیده اند که می توان از آن جمله به [BLO04] و [BLO06] اشاره نمود.

همچنین [BAD09] روشی ترکیبی با استفاده از غنی سازی مبتنی بر wordnet و خوشه بندی مبتنی بر LSA ارائه نموده اند.

در تحقیق دیگری در [LUO11] سعی شده است تا به هر دو جنبه زمینه و معنای کلمات توجه شود. به این منظور فرض می شود که اسناد در دسته هایی قرار دارند و برای هر دسته می توان مجموعه ای از کلمات را در اختیار داشت که نماینده آن دسته هستند. سپس با استفاده از wordnet، این مجموعه کلمات افزایش داده می شوند. در نهایت هر کلمه نیز با توجه به کلمات زمینه ی آن و با کمک wordnet گسترش داده می شود. وزن کلمه در سند عبارت خواهد بود از ترمیب میزان ارتباط سند با دسته ای که در آن قرار دارد و میزان ارتباط لغت با دسته ای که سند مربوطه در آن قرار دارد. این ارتباط با استفاده از معیارهای شباهت معنایی ارائه شده در [LIN98] محاسبه می گردد.

---

[4] - Page rank [2]
[4] - Semantic kernel [3]

همچنین در [WIT09] نیز روش مشابهی مبتنی بر گسترش کلمات با استفاده از wordnet و محاسبه وزن بر اساس میزان شباهت معنایی ارائه شده است. عمده ی تفاوت اینگونه از روش ها در معیار شباهت معنایی و همچنین روش گسترش یک کلمه می باشد.

۲،۳،۳. روش های مستقل از دانش پیش زمینه

این دسته از روش ها سعی دارند تا تنها با توجه به ماهیت اولیه داده ها و استفاده از روش های زبانی و ایده های الگوریتمیک، به استخراج روابط معنایی میان کلمات، و وزن دهی و خوشه بندی و دسته بندی کلمات بپردازند.

از این دست روش ها می توان به کاری که در [ERN07] ارائه شده است، اشاره کرد. در این تحقیق به کلمات به عنوان افرادی از یک شبکه ی اجتماعی نگریسته شده است و اهمیت هر کلمه ناشی از اهمیت کلماتی است که با آن در ارتباط هستند. این نگاه در رابطه ۲-۱۱ نمایش داده شده است.

$$S(w) = (1-\lambda)d_w + \lambda \sum_{u \,\epsilon\, r(w)} S(u)C_{w,u} \qquad (۵)$$

در رابطه ی بالا d_w وزن اولیه هر کلمه است که با استفاده از معیاری مانند TF-IDF مصاحبه می شود و در یک مدل یادگیری سعی می شود تا $C_{w,u}$، یعنی وزن ارتباط بین دو کلمه w و u محاسبه شود. ایده اصلی در [ERN07] برای محاسبه ی $C_{w,u}$، این است که میزان ارتباط دو کلمه تنها در جایی محاسبه می شود که زمینه[۴]مشترکی قرار گرفته باشند. این زمینه ی مشترک عبارت است از یک قطعه متن با اندازه ی محدود که پیوسته باشد. به عنوان مثال یک لینک و یا یک پاراگراف و یا کلماتی که با فاصله مشخصی از یک کلمه قرار دارند تشکیل یک زمینه را می دهند. در نهایت محاسبه C(w,u) با استفاده از رابطه ۲-۱۲ و ۲-۱۳ انجام می شود.

$$C_{w,u} = \sum_{\hat{w}\,\epsilon\, occ(w)} \sum_{\hat{u}\,\epsilon\, ict(\hat{w},u)} C_p(\hat{w},\hat{u}) \qquad (۶)$$

به طوریکه

$$C_p(\hat{w},\hat{u}) = \prod_{i=0}^{i=n} \sigma(\alpha_i x_i + \beta_i) \qquad (۷)$$

[4] - Context [4]

می باشد، که در آن α_i و β_i پارامترهای ویژگی x_i بین دو کلمه w و u هستند و باید طی یک فرآیندیادگیری تعیین شوند و σ تابع سیگموید[۴۵]تحت رابطه ۲-۱۴ می باشد.

$$\sigma(x) = {}^{1}/_{(1 + e^{-x})} \qquad (۸)$$

این ویژگی ها می توانند هر نوع ویژگی شامل ویژگی های مبتنی بر تئوری اطلاعات مانند Information gain و یا ویژگی های morphological مانند قواعد نحوی و گرامری و یا لغوی مانند wordnet و یا آماری باشند، که با توجه به پارامترهای α_i و β_i هر یک تأثیر متفاوتی در نتیجه خواهند داشت. در نهایت برای یادگیری α_i و β_i از روش های یادگیری مانند شبکه عصبی[۴۶]با نگاه به دقت به عنوان تابع خطا عمل شده است. این روش توانسته است نسبت به TF-IDF به صورت متوسط ۱۵٪ افزایش در دقت را به دست آورد [ERN07].

همینطور در این دسته می توان به کاری که در [DAS09] انجام شده است، اشاره کرد. در این تحقیق با استفاده از نظریه ی مرکزیت ویژگی هایی برای هر کلمه در جمله بر اساس مقادیر CB جمله مشخص شده است. سپس با استفاده از یک مدلLDA[۴۷] مبتنی بر ویژگی های بدست آمده یک کاربرد خاص، که خلاصه سازی متون می باشد مورد پیگیری قرار گرفته است.

۲،۳،۴. نتیجه گیری

روش های غیر آماری گفته شده عمدتا مبتنی بر استفاده از دانش پیش زمینه هستند و تعداد کمی از آن ها مستقل از دانش پیش زمینه می باشند. در مجموع می توان گفت که روش های غیز آماری توانسته اند تا حدودی مشکلات زیر را در روش های آماری پوشش دهند

- در نظر نگرفتن معنا
- در نظر نگرفتن ویژگی های زبان شناختی متن و یا لغات
- تأثیر مستقیم ویژگی های آماری متن برروی وزن لغات
- دقت پایین

اما با این حال چند چالش جدی در این روش ها موجود است که از آن جمله می توان به موارد زیر اشاره نمود:

---

4 - Sigmoid 5
4 - Nueral Network 6
4 - Latent Dirichlet Allocation7

- عدم توجه به ویژگی های فرا لغتی در متن. از جمله این ویژگی ها می توان به پیوستگی متن و یا وجود هم آیی ها در بین لغات اشاره نمود.
- عدم وجود یک مدل محاسباتی برای این روش ها
- عدم وجود عمومیت جهت استفاده این روش ها در انواع کاربردهای مختلف
- نگاه کیف لغت به اسناد در اکثر این روش ها و عدم توجه به اهمیت ترتیب لغات در متن
- انتخاب لغت به عنوان جزء پایه در سند

هدف اصلی در این پایان نامه پوشش دادن همین چالش های ذکر شده می باشد. به همین منظور هدف اصلی در این روش ارائه شده در این حوزه در این پایان نامه استخراج اطلاعات بیشتر از متن و استفاده از این اطلاعات به عنوان عوامل تأثیرگذار در وزن دهی به لغات می باشد. علاوه بر این سعی می شود تا یک قالب کاری عمومی تر برای شناسایی اجزاء معنایی ارائه گردد. به همین منظور در این قالب کاری مفهوم قطعه معنایی[۴۸]معرفی می گردد و علاوه بر آن مفهوم دانش پیش زمینه گسترش داده می شود تا امکان استفاده از انواع دانش های پیش زمینه فراهم گردد. همچنین در این قالب کاری برای تمامی این مفاهیم یک روش ارائه مبتنی بر ریاضی ارائه گردیده است که این روش ارائه کمک می نماید تا بتوان یک مدل محاسباتی برای الگوریتم های ارائه شده در اختیار داشت و به راحتی به محاسبه پیچیدگی های زمانی و مکانی الگوریتم ها پرداخت.

پس از ارائه این قالب کاری در بخش ۴,۱ به توضیح در خصوص الگوریتم ارائه شده در بخش ۴,۲ پرداخته خواهد شد. این الگوریتم سعی دارد تا با استفاده از اطلاعات ترتیب لغات و جملات در متن به وزن گذاری لغات بپردازد. به همین منظور مفهومی به نام بردار زمینه[۴۹]تعریف می گردد. در نگاه یک انسان به معنای یک متن، معنا عبارت است از موجودیت متحرک که در طول متن با استفاده از جملات تلاش در تکمیل موجودیت خود دارد. به عبارت خلاصه تر معنا یک حرکت پیوسته در طول متن است. همچنین توجه به این نکته که متون برای استفاده انسان ها نوشته می شوند باعث می شود تا دریابیم چنانچه ماشین نیز بتواند از همین نگاه استفاده نماید آنگاه می توان برداشت بهتری از معنا را در اختیار ماشین قرار داد. در نگاه معمول ماشین به متن، یک متن به عنوان یک موجودیت بدون تغییر و به صورت یکپارچه دیده می شود. بردار زمینه در هر نقطه از متن نشان دهنده معنا از ابتدای متن تا آن نقطه می باشد. بر همین اساس از پیوستگی به عنوان منبعی برای تعیین میزان اهمیت جملات و لغات استفاده می گردد.

---

[4] Semantic Chunk [8]
[4] Context Vector [9]

۳. روش جدیدی در محاسبه مشابهت دو موجودیت

یکی از اجزاء اصلی در تشخیص مفاهیم روش‌هایی هستند که برای استخراج مدل نهان معنایی[۵۰] مجموعه‌های داده مورد استفاده قرار می‌گیرند. اولین نمونه موفق از این روش‌ها را می‌توان LSI دانست. این روش از یک روش تقسیم ماتریس به نام[۵۱]SVD برای تقسیم ماتریس استفاده می‌نماید. اما همانطور که گفته شد این روش دارای مشکلاتی است که لزوم تغییر در آن را به خوبی نمایان می‌سازد. مهم‌ترین این مشکلات وابستگی این روش به داده‌هایی با توزیع گاوسی می‌باشد. از این رو این روش تنها برای داده‌هایی با توزیع گاوسی می‌تواند بهترین نتایج را داشته باشد و برای دیگر انواع داده به علت انتشار بالای خطا نمی تواند نتایج خوبی را به همراه داشته باشد. اما در اصل این مشکل از معیاری که LSI برای تعیین مشابهت بین داده‌ها استفاده می‌نماید حاصل می‌گردد [HOF99]. این معیار در اصل همان معیار همبستگی معمول مورد استفاده در روش‌های مختلف می‌باشد که ارتباط نزدیکی با معیارهای دیگر مانند فاصله کسینوسی نیز دارد.

در عین حال برای حل این مشکل دسته‌ای از روش‌ها مانند PLSI و LDA از طریق تعریف متغیرهای احتمال شرطی و با استفاده از روش[۵۲]EM اقدام به حل مسئله و کشف مدل معنایی نهان می‌نمایند. اما این روش‌ها دارای زمان اجرای بالایی می‌باشند و از این رو مورد توجه ما نمی باشند. در عین حال LSI علیرغم مشکلاتش در حوزه تعیین مدل معنایی دارای زمان اجرای قابل قبولی می‌باشد. از این رو به نظر می‌رسد

---

[50] Latent Semantic Model
[51] Singular Value Decomposition
[52] Expectation Maximization

ارائه راهکاری که دارای شباهت عملکرد با LSI باشد و در عین حال با اصلاح معیار مشابهت بتواند مشکلات LSI را مرتفع سازد، می‌تواند گزینه مناسبی محسوب گردد.

از سوی دیگر توابع ارتباط که پیش از این در خصوص آن‌ها توضیح دادیم دارای خواص بسیار مناسبی هستند که آن‌ها را تبدیل به گزینه‌های مناسبی جهت جایگزینی معیارهای شباهت می‌نماید. اما از مهم‌ترین مشکلات این روش‌ها زمان اجرای آن‌ها است که دارای پیچیدگی بسیار بالایی هستند. این پیچیدگی محاسباتی باعث می‌شود که این روش‌ها را نتوان برای حجم بالایی از داده‌ها مورد استفاده قرار داد. همچنین استفاده از توابع ارتباط دارای جزئیات بسیار بالای دیگری نیز می‌باشد که باید در نظر گرفته شوند.

آنچه در ارائه یک راهکار جدید در حوزه تعیین مفاهیم و یا به عبارت دیگر تعیین مدل معنایی نهفته باید مورد توجه قرار گیرند عبارتند از:

- روش‌های ارائه شده باید توانایی ایجاد ارتباط بین یک لغت با بیش از یک مفهوم را داشته باشند. همان طور که واضح است مفاهیم موجودیت‌هایی مجزا از هم نیستند و دارای اشتراکاتی به لحاظ معنایی می‌باشند و از این رو ممکن است یک لغت در چندین مفهوم مورد استفاده قرار گیرد. در نتیجه تنها روش‌هایی قابل استفاده می‌باشند که بتوانند به صورت همزمان ارتباط یک لغت با چندین مفهوم را شناسایی نمایند.
- روش ارائه شده نباید نسبت به تابع توزیع داده‌ها حساس باشد. برای آشکار شدن بیشتر مطلب فرض کنید هر لغت را بتوان به صورت یک متغیر تصادفی نمایش داد. فضای رخداد چنین متغیرهایی مجموعه اسنادی خواهد بود که به عنوان مجموعه مرجع مورد استفاده قرار می‌گیرند. براین اساس می‌توان برای هر متغیر تصادفی یک تابع توزیع مفروض دانست. در مجموعه‌های داده بزرگ مانند وب که مجموعه متغیرهای بسیار گسترده و همچنین حوزه‌های معنایی بسیار متنوعی را پوشش می‌دهند، نمی توان هیچ تابع توزیع مشخصی را یافت که بتواند خواص اکثر متغیرهای تصادفی تعریف شده را پوشش دهد. همچنین حتی نمی توان خانواده‌ای از توابع توزیع و یا تعداد محدودی از توابع توزیع را یافت که بتواند چنین خاصیتی داشته باشند. بنابراین استفاده از روش‌هایی که دارای وابستگی به تابع توزیع داده‌ها می‌باشند نمی تواند به هیچ عنوان قابل توجیه باشد و مسلما با همان مشکلاتی که برای روش‌هایی مانند LSI مطرح گردید مواجه خواهند بود.
- پیچیدگی محاسباتی روش ارائه شده یکی از مسائل اصلی مد نظر می‌باشد. در مجموعه داده‌های بسیار وسیع مانند وب استفاده از روش‌هایی که دارای پیچیدگی محاسباتی بالا می‌باشند امکان پذیر نمی باشد. در چنین مجموعه داده‌هایی حتی استفاده از الگوریتم‌هایی که در دسته مسائل P قرار می‌گیرند اما پیچیدگی محاسباتی آن‌ها چند جمله ای‌هایی با درجه بالا می‌باشند نیز قابل توجیه

نیست. پر واضح است که استفاده الگوریتم‌هایی که در دسته P قرار نمی گیرند به هیچ وجه قابل قبول نیست.

- حجم پارامترهای ورودی یکی دیگر از مسائل مهم در این روش‌ها می‌باشد. بعضی از روش‌ها حجم بسیار بالایی از پارامترهای ورودی را دریافت می‌کنند که باید در هر بار اجرای الگوریتم محاسبه شده و به عنوان پارامترهای ورودی در اختیار الگوریتم قرار گیرند. این پارامترهای ورودی عمدتا مقادیری هستند که باید براساس روش‌های بهینه سازی و یا یادگیری محاسبه شوند و محاسبه آن‌ها دارای پیچیدگی زمانی بسیار بالایی می‌باشد. ساده‌ترین نوع از این پارامترهای ورودی تعداد مفاهیم است. اما در روش‌هایی مانند LDA حجم بسیار بزرگی از پارامترهای ورودی که مبتنی بر بهینه سازی محاسبه می‌شوند باید در اختیار الگوریتم قرار گیرند.

تقریبا هیچیک از روش‌های ارائه شده تا کنون نتواسته است تمام این مزایا را به صورت همزمان پوشش دهد. اما روش ارائه شده در این فصل سعی بر آن دارد تا بتواند تمامی این مزایا را در اختیار داشته باشد.

با توجه به آنچه گفته شد در ادامه اصول راهکار پیشنهادی ارائه می‌گردد. مهم‌ترین مسائل مطرح در ارائه راهکار انتخاب تابع ارتباط مناسب، پیوستگی و همچنین ارائه یک راهکار محاسباتی برای روش ارائه شده می‌باشند. در نهایت نتایج ارزیابی عملی این روش مورد بررسی قرار گرفته است.

۳,۱. راهکار پیشنهادی

براساس آنچه در این فصل گفته شد می‌توان نتیجه گرفت که چنانچه روش بتواند رویکرد کلی روش LSI را اتخاذ نماید و در عین حال خاصیت عدم وابستگی به تابع توزیع داده‌ها را نیز در اختیار داشته باشد، می‌تواند به عنوان یک روش کاملا کاربردی مورد استفاده قرار گیرد. LSI به عنوان یکی از روش‌های خوشه بندی طیفی[۵۳]شناخته می‌شود. این روش‌های خوشه بندی مبتنی بر محاسبه بردارهای ویژه[۵۴]ماتریس مشابهت می‌باشند. می‌توان رویکرد کلی این روش‌ها را در قالب شبه الگوریتم زیر بیان نمود [DEE90, PAP98].

چنانچه معیار مشابهت مناسبی در اختیار داشته باشیم می‌توانیم بر مشکل وابستگی این دسته از روش‌ها به تابع توزیع داده‌ها فائق آییم. معیار مشابهتی که توسط LSI مورد استفاده قرار می‌گیرد ماتریس همبستگی بردارهای لغات می‌باشد. این معیار شباهت همانطور که پیش از این گفته شد دارای وابستگی به تابع توزیع داده‌ها می‌باشدو تنها برای داده‌های با تابع توزیع گاوسی بهترین نتایج را تولید می‌نماید.

---

[5] Spectral Clustering [3]
[5] EigenVector [4]

Input: Similarity matrix $S \in R^{m \times m}$, number of clusters (k)

- Construct a similarity graph
- Compute the normalized LaplacianL_{sym}
- Compute first k eigenvectors $u_1, \dots, u_k$ of L_{sym} and construct $U_{m \times k}$ using them
- From matrix T from U by normalizing the rows to norm 1, that is set to

$$t_{ij} = \left(\sum_k u_{ik}^2 \right)^{1/2}$$

- Cluster the rows of T using k-means into k clusters

Output: Clusters $A_1, \dots, A_k$with $A_i = \{j | t_j \in C_i\}$.

یکی از ایده‌هایی که براساس این توضیحات به نظر می‌رسد استفاده از ضرایب همبستگی مستقل از توزیع می‌باشد. از این دست ضرایب همبستگی می‌توان به Spearman و Kendall's τ اشاره نمود. اما اعمال این ضرایب نمی تواند بهبود قابل توجهی در نتایج ایجاد نماید (نتایج عملی استفاده از این معیارها در فصل نتایج عملی مورد بررسی قرار گرفته است). یکی از دلایل مهم این است که هر چند این ضرایب نسبت به تابع توزیع داده‌ها مستقل می‌باشند اما نمی توانند حجم اطلاعات قابل توجهی را در خصوص شباهت متغیرهای تصادفی در اختیار قرار دهند. همچنین این ضرایب خواص خود را برای داده‌های گسسته حفظ نمی نمایند و از این رو در این نوع از داده‌ها دارای دقت مناسب در تعیین شباهت بین متغیرهای تصادفی نیستند [DEN05]. همچنین یکی دیگر از مشکلات این ضرایب در داده‌های گسسته این است که بازه برد این ضرایب در داده‌های گسسته بسیار محدودتر از بازه تعریف شده یعنی بازه ۰ و ۱ می‌باشد. که این مسئله تأثیر بسیار زیادی بر دقت نتایج و همچنین نحوه تفسیر نتایج می‌گذارد [DEN05].

براساس همین مسائل روش ارائه شده مبتنی بر استفاده از توابع ارتباط برای ایجاد مدل معنایی نهان می‌باشد. از این رو این روش[5]COSIN نام گذاری شده است. مهم‌ترین خاصیت توابع ارتباط که آن‌ها را تبدیل به ابزاری مناسب برای این منظور می‌نماید استقلال آن‌ها نسبت به تابع توزیع داده‌ها می‌باشد. همچنین یکی دیگر از ویژگی‌های آن‌ها امکان کاربرد آن‌ها برای داده‌هایی با چندین نوع تابع توزیع می‌باشد [NEL98]. این خواص دقیقا شرایط لازم برای اعمال برروی داده‌هایی مانند محیط وب را فراهم می‌نمایند.

[5] COpula based Semantic INdĕxing

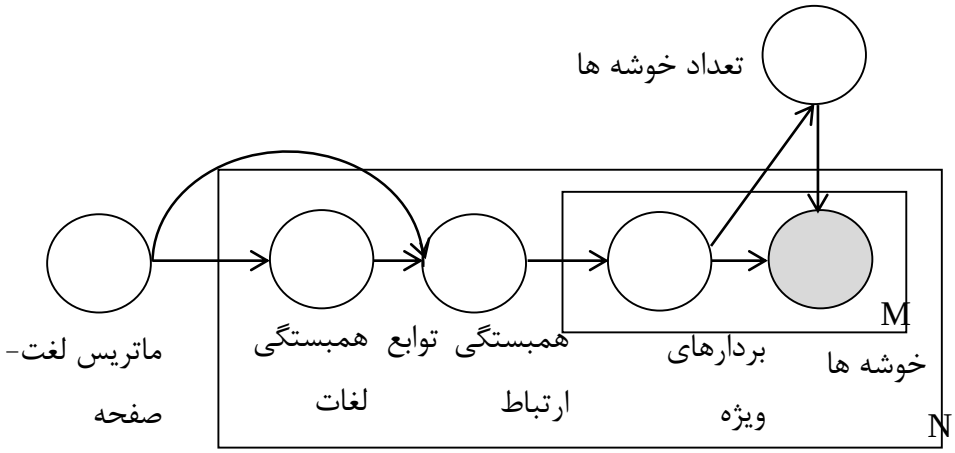
در COSIN با استفاده از توابع ارتباط شباهت بین متغیرهای تصادفی محاسبه گردیده و سپس مشابه روش‌های خوشه بندی طیفی عملیات خوشه بندی صورت می‌گیرد. بر همین اساس می‌توان نمای کلی COSIN را به صورت زیر ارائه نمود.

شبه کد ۴ - الگوریتم کلی محاسبه همبستگی مبتنی بر روش پیشنهادی

1. Compute *TFiDF* for Term-Document Matrix
2. Normalize *TFiDF* values
3. **for** each term like t_i **do**
    a. **for** each term like t_j **do**
        i. compute *Spearman*'s sign correlation between t_i and t_j
    b. end for
4. end for
5. **for** each term like t_i **do**
    a. **for** each term like t_j **do**
        i. compute *Copula* for t_i and t_j
    b. end for
6. end for
7. **for** each term like t_i **do**
    a. **for** each term like t_j **do**
        i. compute *Correlation*s similarity between t_i and t_j based on copula
    b. end for
8. end for
9. compute *Spectral_Clustering* for similarity matrix of terms

برای اینکه درکی واضح‌تر از نحوه استنتاج این الگوریتم در تشخیص مفاهیم و تشکیل مدل معنایی نهان در اختیار داشته باشیم در شکل ۴ مدل گرافی این الگوریتم ارائه گردیده است. این مدل براساس مدل گرافی استاندارد که در علوم احتمالات مورد استفاده قرار می‌گیرد رسم شده است. این مدل به عنوان ابزاری برای بیان نحوه استنتاج در احتمالات شرطی مورد استفاده قرار می‌گیرد.

براساس الگوریتم ارائه شده و همچنین شکل ۴ واضح است که تنها پارامترهای ورودی این روش ماتریس لغت-صفحه و همچنین تعداد خوشه‌ها و یا همان تعداد مفاهیم می‌باشد. همچنین محاسبه ضریب همبستگی Kendall و همچنین همبستگی مبتنی بر توابع ارتباط وابسته به تعداد صفحات می‌باشد و محاسبه بردارهای ویژه تنها وابسته به تعداد لغات می‌باشد.



شکل ۴ - مدل استنتاج الگوریتم پیشنهادی

پیچیدگی محاسباتی این الگوریتم $O(m^2(c^2+k)+m.poly(k/\epsilon)+2^{\tilde{O}(k/\epsilon)})$ می‌باشد که در بخش ۳,۱,۳ به صورت کامل توضیح داده شده است. این زمان اجرا از زمان اجرای LSI بهتر است. البته این پیچیدگی محاسباتی با فرض این است که هر بار محاسبه تابع ارتباط بتواند در زمان $O(1)$ انجام پذیرد. این فرض در حالت کلی درست نمی باشد، اما در این پایان نامه مدلی ارائه شده است که بتوان با استفاده از آن به این فرض دست یافت. همچنین مسائل بسیار زیاد دیگری نیز در استفاده از توابع ارتباط مد نظر قرار گیرند که در ادامه در مورد آن‌ها توضیح داده خواهد شد.

۳,۱,۱. انتخاب تابع ارتباط مناسب

همانطور که در بخش ۸,۵ بیان خواهد شد روش‌های مختلفی برای انتخاب و یا ساخت یک تابع ارتباط موجود می‌باشند. در نگاهی متفاوت می‌توان این روش‌ها را به دسته‌های زیر تقسیم نمود:

- استفاده از توابع ارتباط معمول و عمومی
- ایجاد توابع ارتباط مبتنی بر خواص داده‌ها و یا توابع توزیع آن ها
- تخمین توابع توزیع مبتنی بر خواص داده ها

کاملا واضح است که با توجه به عدم دسترسی به خواص داده‌ها و یا توابع توزیع آن‌ها نمی توان اقدام به ایجاد توابع ارتباط مناسب نمود. از این رو دسته دوم از روش‌ها که در فوق ذکر شده است قابل استفاده برای مجموعه داده‌هایی مانند وب نیستند. همچنین همانطور که گفته شد روش‌های تخمین توابع ارتباط دارای پیچیدگی محاسباتی بالایی می‌باشند و از این رو این روش‌ها نیز نمی توانند مورد استفاده قرار گیرند. پس

تنها استفاده از توابع ارتباط عمومی روشی است که به عنوان راه حل برای استفاده در داده‌هایی مانند وب باقی می‌ماند.

برهمین اساس توابع ارتباط معمول مورد مقایسه قرار گرفتند و براساس مقایسه نتایج عملی یکی از آن‌ها انتخاب گردید. این نتایج در بخش نتایج عملی قابل مشاهده می‌باشد. براساس این نتایج تابع ارتباط Student's t به عنوان بهترین تابع از بین توابع موجود انتخاب گردید. به همین علت این تابع به عنوان تابع ارتباط مورد استفاده در روش ارائه شده در این پایان نامه مورد استفاده قرار گرفته است. یکی دیگر از خواص بسیار مهم این تابع شناسایی وابستگی هایی[۶]انتهایی می‌باشد [CHE06]. منظور از وابستگی انتهایی، وابستگی‌هایی است که در بازه‌های انتهایی مقادیر متغیرهای تصادفی رخ می‌دهند. منظور از مقادیر انتهایی مقادیر احتمال نزدیک به صفر و یک می‌باشند. به عبارت دیگر این تابع به خوبی می‌تواند وابستگی بین دو متغیر تصادفی را چه در زمان رخدادهای یکسان و چه عدم رخدادهای یکسان شناسایی نماید که خاصیتی بسیار مهم در کاربرد مد نظر این پایان نامه می‌باشد.

۳,۱,۲. پیوستگی

به صورت معمول توابع ارتباط برای متغیرهای پیوسته مورد استفاده قرار می‌گیرند. اما در کاربردهایی مانند آنچه در این پایان نامه مد نظر می‌باشد متغیرها از نوع گسسته هستند. به همین منظور باید روش‌هایی مورد استفاده قرار گیرند تا داده‌های گسسته موجود به داده‌های پیوسته تبدیل شده و سپس توابع ارتباط برروی آن‌ها اعمال گردند و یا اینکه نتایج حاصل از داده‌های گسسته به نحوی تصحیح شود تا بتواند ویژگی‌های داده‌ها را به خوبی نمایش دهد. البته در بسیاری از موارد بدون چنین تصحیحاتی داده‌های گسسته مورد استفاده قرار می‌گیرند. اشکالاتی که در اثر گسستگی داده‌ها به وجود می‌آیند در [DEN05] به خوبی شرح داده شده و مورد بررسی قرار گرفته‌اند. در یک جمع بندی می‌توان گفت استفاده از داده‌های گسسته ممکن است بتواند تا حدی پاسخگوی نیازها باشد اما در حالت کلی با وجود داده‌های گسسته نمی توان انتظار داشت که میزان دقت روش از حدی بیشتر گردد. حتی ضرایب همبستگی که مستقل از تابع توزیع داده‌ها می‌باشند نیز مانند Spearman و Kendall's τ نمی توانند در داده‌های گسسته به خوبی نمایانگر ویژگی‌های وابستگی بین داده‌ها باشند. می‌توان مشکلات حاصل از گسستگی داده‌ها در موارد زیر خلاصه نمود:

- عدم امکان کشف صحیح روابط وابستگی بین داده‌ها به نحوی که روش‌های مختلف کشف همبستگی بین داده‌ها نخواهند توانست این وابستگی‌ها را به خوبی نمایش دهند.

---

[5] Tail Dependence [6]

- یکتایی وجود توابع ارتباط در شرایط گسستگی داده‌ها نقض می‌شود. این ویژگی پایه‌ای اساسی برای بسیاری از ویژگی‌های توابع ارتباط است و از این رو عدم وجود آن منجر به نقض بسیاری از ویژگی‌ها و عدم کاربرد توابع ارتباط می‌گردد.

در [DEN05] راه حلی برای این مشکل ارائه شده است. برای این منظور متغیرهای تصادفی از طریق اضافه کردن یک متغیر تصادفی دیگر به آن‌ها به گونه‌ای گسترش می‌یابند که علاوه بر تبدیل شدن به داده‌های پیوسته بتوانند ویژگی‌های احتمالی خود را نیز حفظ نمایند. در اینجا به شرحی مختصر از این روش می‌پردازیم.

فرض کنید متغیری مانند X در اختیار داریم که یک متغیر گسسته می‌باشد. حال متغیر پیوسته X^* با استفاده از آن ساخته می‌شود به نحوی که

$$X^* = X + (U - 1) \qquad (۹)$$

در اینجا U یک متغیر تصادفی در بازه صفر و یک است که دارای یک تابع توزیع اکیدا صعودی می‌باشد. به عنوان مثال اگر فرض کنیم U یک متغیر تصادفی یکنواخت است دارای تابع توزیع تجمعی برابر با U و تابع توزیع چگالی برابر با ۱ خواهد بود. واضح است که چنین متغیری دارای یک تابع توزیع اکیدا صعودی است. پس از این توسعه بر روی متغیر تصادفی X، گفته می‌شود که X از طریق U پیوسته شده است و حال می‌توان از متغیر X^* به عنوان یک متغیر پیوسته با همان ویژگی‌های احتمالی استفاده نمود. جزئیات اثبات این ویژگی در [DEN05] آمده است که در اینجا از ذکر جزئیات خودداری می‌نماییم.

براساس این تغییر فرمول محاسبه همبستگی با استفاده از توابع ارتباط که در بخش ۸,۵ ارائه شده است پس از جایگزاری به شکل زیر در خواهد آمد

$$correlation = \frac{1}{4}\sum_{x=0}^{+\infty}\sum_{y=0}^{+\infty}\Pr(X = x, Y = y) \times \{C(F_x, G_y) + C(F_x, G_{y-1}) + C(F_{x-1}, G_y) + C(F_{x-1}, G_{y-1})\} \qquad (۱۰)$$

که در اینجا F و G توابع توزیع تجمعی متغیرهای X و Y هستند. با استفاده از این رابطه می‌توان همبستگی بین متغیرهای تصادفی را با حفظ تمام ویژگی‌های احتمالی آن‌ها قابل محاسبه خواهد بود. همچنین کارهایی در [LEO11] صورت گرفته است تا بتوان توابع ارتباط را برای انواع متغیرهای گسسته و پیوسته به صورت همزمان مورد استفاده قرار داد.

۳,۱,۳. ایجاد مدل محاسباتی

در ابتدای این بخش پیچیدگی زمانی الگوریتم ارائه شده با فرض امکان محاسبه توابع ارتباط در $O(1)$ بیان شد. اما این فرض در حالت کلی درست نیست و در اصل مهم‌ترین ضعف توابع ارتباط پیچیدگی محاسباتی بالای محاسبه آن‌ها می‌باشد. عمدتا توابع ارتباط مناسب شامل محاسبه انتگرال‌های دو گانه (برای توابع ارتباط دو متغیره، مسلما برای تعداد متغیرهای بیشتر درجه انتگرال‌ها نیز افزایش می‌یابد) می‌باشند. روش‌های عددی محاسبه انتگرال‌های دوگانه دارای پیچیدگی محاسباتی بالایی می‌باشند و حتی روش‌هایی که این روش‌های عددی را تقریب می‌زنند نیز دارای پیچیدگی محاسباتی بالایی هستند [CHE06]. در این پایان نامه نیز تابع ارتباط *Student's t* به عنوان تابع ارتباط انتخاب گردیده است که جزء توابع ارتباط پیچیده و دارای انتگرال‌های دوگانه محسوب می‌گردد. همانطور که گفته شد فرض محاسبه این تابع در $O(1)$ درست نمی باشد. راه حلی که در این پایان نامه برای این مشکل ارائه گردیده است گسسته سازی داده‌ها و سپس ارائه جدول مقادیر از پیش محاسبه شده توابع ارتباط به عنوان یک دانش پیش زمینه به الگوریتم می‌باشد که امکان محاسبه تابع ارتباط در $O(1)$ را فراهم می‌آورد.

در یک نگاه کلی به داده‌ها می‌توان گفت که کل تعداد محاسبه‌های تابع ارتباط برابر با $O(m^2c^2)$ می‌باشد. که در آن m برابر است با تعداد لغات موجود در مجموعه داده‌ها و c نیز برابر است با تعداد متوسط اسنادی که یک لغت در آن‌ها تکرار شده است. همچنین پارامترهای ورودی تابع ارتباط *Student's t* عبارتند از دو متغیر تصادفی و همچنین دو پارامتر تنظیم کننده رفتار تابع که براساس خواص همبستگی بین دو متغیر تصادفی باید تنظیم گردند و همان طور که در بخش ۸,۵ گفته شد بر اساس ضرایب همبستگی ساده‌تر که مستقل از تابع توزیع می‌باشند مانند Spearman و Kendall's τ می‌توان محاسبه نمود. براین اساس چنانچه این محاسبات حتی بدون گسسته سازی یک بار صورت بگیرند در محاسبات بعدی می‌توانند به عنوان دانش پیش زمینه مورد استفاده قرار گیرند.

اما برای کاهش این محاسبات حتی به عنوان دانش پیش زمینه در اینجا از روش‌های گسسته سازی استفاده گردیده است تا علاوه بر کاهش محاسبات دقت تا حد ممکن حفظ گردد. روش‌های گسسته سازی متنوع و پرکاربردی مانند ChiMerge [KER92]، گسسته ساز Holte's 1R [HOL93]، تقسیم ساز بازگشتی مبتنی بر کمترین آشفتگی[۵۷] [FAY93]، K-Means [FEL07] و بسیاری روش‌های دیگر وجود دارند. روش‌های نام برده شده در اینجا در اصل نمایندگان خانواده‌ای از روش‌ها هستند که رویکرد یکسانی دارند و تنها تغییرات کوچکی در جزئیات داده شده تا میزان کمی در دقت آن‌ها بهبود داده شود و یا آن‌ها را برای

[5] Recursive Minimal Entropy Partitioner

یک کاربرد خاص مناسب سازد. در [KOT06, DOU95] مقایسه مشروحی بین دسته‌های مختلف روش‌های گسسته سازی صورت گرفته است. بر این اساس دو روش [FAY93] و K-Means به عنوان روش‌های منتخب در این پایان نامه برگزیده شدند و گسسته سازی برای هر دوی این روش‌ها صورت گرفته است که نتایج آن در بخش نتایج عملی مورد بررسی قرار گرفته‌اند. براساس این نتایج روش K-Means توانسته است نتایج بهتری را نشان دهد و همین روش به عنوان روش گسسته سازی در این پایان نامه انتخاب گردید. با استفاده از گسسته سازی یک کاهش فضا در جهت تعداد نمونه‌ها داریم که می‌تواند تعداد محاسبات را تا حد بسیار مناسبی کاهش دهد.

با اعمال K-Means حجم داده‌ها کاهش یافته و بر این اساس یک جدول از مقادیر از پیش محاسبه شده تابع ارتباط *Student's t* محاسبه می‌گردد که به عنوان دانش پیش زمینه برای الگوریتم استفاده خواهد شد. این راهکار فرض محاسبه تابع ارتباط در زمان $O(1)$ را عملی می‌نماید.

تنها چند نکته در اینجا باقی است که باید مورد بررسی قرار گیرند. نکته اول این است که تأثیر این گسسته سازی برروی دقت چه مقدار است. برای بررسی این موضوع داده‌های دقت الگوریتم ارائه شده با اعمال گسسته سازی و بدون آن در بخش نتایج عملی مورد بررسی قرار گرفته‌اند. همچنین نکته دوم حجم این دانش پیش زمینه است. چنانچه حجم این داده بسیار بالا باشد نمی توان انتظار داشت که بتوان با ساختار داده مناسبی به زمان دسترسی $O(1)$ دست یافت. در این خصوص نیز حجم داده‌ها پیش از گسسته سازی و پس از آن در بخش نتایج عملی به صورت کامل بیان شده‌اند. این نتایج نشان می‌دهند که گسسته سازی تأثیر قابل توجهی در دقت نگذاشته است و در مقابل حجم داده‌ها به میزان قابل توجهی کاهش یافته است به صورتی که می‌توان انتظار داشت تا بتوان این داده‌ها را در ساختارهای داده مناسب در حافظه اصلی ذخیره نمود تا زمان دسترسی $O(1)$ محقق گردد.

۳,۲. ارزیابی روش

در این بخش به بررسی ویژگی‌هایی می‌پردازیم که در بخش ۱,۲ به عنوان ویژگی‌های مورد نیاز برای یک روش استخراج مفاهیم بیان نمودیم. این بررسی می‌تواند توانایی‌های COSIN را به عنوان روشی برای استخراج مفاهیم و تشکیل مدل معنایی نهان برروی داده‌های با حجم بسیار بالا نمایش دهد. به همین منظور بار دیگر به صورت خلاصه ویژگی‌های مد نظر را مرور می‌نماییم.

۱. امکان مشارکت یک لغت در چندین مفهوم
۲. عدم وابستگی روش به تابع توزیع داده ها
۳. حجم پارامترهای ورودی کم

۴. پیچیدگی محاسباتی پایین

اولین ویژگی وظیفه ایجاد امکان تشخیص همپوشانی‌های معنایی بین مفاهیم مختلف را برعهده دارد. بدون این ویژگی مفاهیم به عنوان موجودیت‌هایی مستقل از هم فرض می‌شوند که در حوزه معنا یک فرض کاملا غلط است. برای دستیابی به این مهم روش کار به صورتی است که در ادامه می‌آید. همانطور که می‌دانید بردارهای ویژه حاصل از الگوریتم ارائه شده در اصل نماینده مفاهیم هستند. هر لغت در این بردارها دارای میزان اهمیتی است. برای تعیین اینکه کدام لغات وابسته به یک مفهوم می‌باشند از یک مقدار آستانه استفاده می‌شود که براثر آزمایش قابل محاسبه است. لغاتی که دارای یزان اهمیتی بیش از حد آستانه باشند لغاتی هستند که در مفهوم مورد نظر دارای نقش می‌باشند. از این طریق امکان مشارکت لغات در بیش از یک مفهوم فراهم می‌آید و COSIN می‌تواند اولین ویژگی را به دست آورد.

البته باید خاطر نشان نمود که دو رویکرد برای استفاده از این مقدار آستانه وجود دارد. رویکرد اول همان است که در بالا ذکر شد. در رویکرد دوم از این حد آستانه برای رتبه یک لغت در مفهوم استفاده می‌شود. در این رویکرد لغاتی را که در بیشترین میزان اهمیت در مفهوم دارند را به تعداد ثابتی انتخاب می‌نمایند. به دلیل اینکه دامنه تغییرات مقادیر اهمیت لغات در مفاهیم در روش‌های مبتنی بر مقادیر ویژه بسیار بالا است از این رو استراتژی دوم، استراتژی مناسب تری است و به همین دلیل به عنوان استراتژی منتخب در COSIN مورد استفاده قرار گرفته است. علت این دامنه تغییرات این است که مقادیر بردارهای ویژه دارای تناسبی نسبی با مقادیر ویژه می‌باشد و هر چه یک مقدار ویژه بزرگ‌تر باشد تقریبا به همان نسبت مقادیر بردار ویژه متناظر نیز بزرگ‌تر خواهند بود. از این رو نمی توان مقادیر بردارهای ویژه مختلف را به صورت مستقیم با هم مقایسه نمود.

دومین ویژگی مورد نظر برای این دسته از روش‌ها عدم وابستگی به تابع توزیع داده‌ها می‌باشد. همانطور که گفته شده مهم‌ترین امتیاز توابع ارتباط عدم وابستگی آن‌ها به تابع توزیع داده‌ها می‌باشد. در روش COSIN برای تعیین میزان همبستگی دو لغت به یکدیگر از همین روش‌ها استفاده گردید. از سوی دیگر در دیگر اجزای روش نیز از قسمت‌هایی که وابستگی به یک تابع توزیع خاص داشته باشند استفاده نگردیده است. از این رو می‌توان استنتاج نمود که COSIN توانسته است دومین خاصیت مد نظر برای روش‌های استخراج مفهوم را نیز به دست آورد.

سومین ویژگی مد نظر حجم پارامترهای ورودی الگوریتم می‌باشد. در COSIN تنها پارامتر ورودی تعداد مفاهیم می‌باشد. این پارامتر نیز دارای دو روش محاسبه می‌باشد. روش اول استفاده از روش‌های آموزش است که از طریق اجرای روش به تعداد زیاد برروی داده‌های تست به دست می‌آید. اما این روش به علت

کوچک بودن داده‌های تست روشی عملیاتی نیست. از سوی دیگر در روش‌های مبتنی بر بردارهای ویژه روش بهتری وجود دارد که در [GAO11] ارائه شده است. در این روش براساس مقادیر ویژه تصمیم گرفته می‌شود که چه تعداد از بردارهای ویژه دارای ارزش اطلاعاتی می‌باشند. واضح است که این روش هیچ بار محاسباتی بیشتری را به الگوریتم تحمیل نمی نماید. جدول مقادیر از پیش محاسبه شده تابع ارتباط نیز به عنوان یک دانش پیش زمینه مورد استفاده قرار می‌گیرند و لازم نیست که در هر بار اجرای الگوریتم محاسبه شوند. از این رو جزء پارامترهای ورودی محسوب نمی شوند. با این توضیحات واضح است که COSIN دارای تعداد پارامترهای ورودی محدودی است که ویژگی بسیار مهمی برای این روش محسوب می‌گردد.

آخرین ویژگی روش‌های استخراج مفهوم زمان اجرای آن‌ها می‌باشد. براساس ساختار الگوریتم COSIN می‌توان اجزاء زیر را به عنوان اجزاء اصلی محاسبه پیچیدگی زمان COSIN در نظر گرفت:

- محاسبه ماتریس ورودی تابع ارتباط: دو گام ابتدایی COSIN وظیفه آماده سازی ماتریس لغت-صفحه را برای استفاده در تابع ارتباط برعهده دارند. در این محاسبه تنها اعضای غیر صفر ماتریس دخالت دارند از این رو می‌توان زمان اجرای آن را $O(m \times c)$ دانست که در آن c میانگین تعداد صفحاتی است که یک لغت در آن‌ها حضور دارد. در این گام در اصل بردار یک لغت که بردار مقادیر TFiDF می‌باشد به بردار رخداد احتمالی تبدیل می‌گردد. ورودی تابع ارتباط باید مقادیر احتمال رخداد لغات باشند و مقادیر TFiDF کاربردی ندارند.
- محاسبه ضریب همبستگی *Spearman*: در [CHR05] الگوریتمی برای محاسبه این ضریب ارائه گردیده است. این الگوریتم می‌تواند در زمان $c \log c$ همبستگی بین دو بردار لغت را محاسبه نماید. با توجه به لزوم محاسبه این ضرایب همبستگی برای تمام جفت لغت ها، زمان لازم برای این گام $m^2 c \log c$ خواهد بود.
- محاسبه تابع ارتباط: با توجه به تعریف توابع ارتباط، در صورتی که یکی از مقادیر متغیرهای ورودی برابر صفر باشد آنگاه مقدار تابع ارتباط صفر خواهد بود. از این رو محاسبه تابع ارتباط باید تنها برای ورودی‌های غیر صفر صورت گیرد. همچنین با فرض امکان یک بار محاسبه تابع ارتباط در زمان $O(1)$، پیچیدگی زمان این گام برابر با $O(m^2 c^2)$ خواهد بود.
- محاسبه ماتریس شباهت لغات: این ماتریس در اصل ورودی روش خوشه بندی طیفی خواهد بود. فرمول محاسبه شباهت یا همان همبستگی با استفاده توابع ارتباط در بخش ۸,۵ ارائه شده است. واضح است که یک بار محاسبه این فرمول به زمانی برابر با c^2 احتیاج دارد. در نتیجه برای محاسبه وابستگی‌ها بین تمام جفت لغات به زمان $O(m^2 c^2)$ برای این گام احتیاج خواهد بود.

- خوشه بندی طیفی: در آخرین گام باید دو عمل محاسبه بردارهای ویژه و همچنین خوشه بندی براساس این بردارها صورت گیرند. محاسبه بردارهای ویژه را می‌توان با استفاده از روش ارائه شده در [PAP98] در زمان $O(mc\log m)$ انجام داد. همچنین برای خوشه بندی طیفی از روش K-Means در این پایان نامه استفاده شده است. در [FEL07] الگوریتمی برای K-Means ارائه شده است که می‌تواند خوشه بندی را در زمان

$$O(nkd + d.poly(k/\epsilon) + 2^{\tilde{O}(k/\epsilon)}) \qquad (۱۱)$$

انجام دهد. این الگوریتم در اصل یک PTAS برای مسئله K-Means می‌باشد. با جایگزاری شرایط مسئله استخراج مفاهیم که در این پایان نامه مورد بررسی قرار گرفته است به زمان اجرای

$$O(m^2k + m.poly(k/\epsilon) + 2^{\tilde{O}(k/\epsilon)}) \qquad (۱۲)$$

دست می‌یابیم. پس از این مرحله مفاهیم برای کاربردهای بعدی آماده می‌باشند.

براساس این تحلیل می‌توان زمان اجرای COSIN را برابر با

$$O(m^2(c^2 + k) + m.poly(k/\epsilon) + 2^{\tilde{O}(k/\epsilon)}) \qquad (۱۳)$$

دانست. همانطور که گفته شد این زمان اجرا با فرض استفاده از جدول مقادیر از پیش محاسبه شده برای توابع ارتباط می‌باشد.

۳,۱. نتیجه گیری

مباحث این فصل به خوبی نشان دادند که کارکرد توابع ارتباط در فائق آمدن بر مشکلات مطرح شده به چه صورت است. نکته مهم این است که در روش ارائه شده سعی شده است علیرغم اینکه در نگاه اول به نظر می رسید روش مورد استفاده با مشکلاتی در پیاده سازی مواجه می شود، اما سعی شد تا ابتدا از مزایای آن استفاده شود و سپس با ارائه راهکار نگاه اولیه اصلاح شده و به روشی مناسب برای فائق آمدن بر مشکلات مطرح شده دست یافت.

البته راه حل ارائه شده سعی نمود در هر دو حوزه روش های استخراج مفاهیم و معیارهای مشابهت نسبت به پوشش مشکلات مطرح شده اقدام نماید. این مشکلات عبارت بودند از زمان اجرای مناسب، عدم وابستگی به تابع توزیع داده ها، حجم کم پارامترهای ورودی و استفاده از روش های تعیین شباهت غیر هندسی. براساس کلیه توضیحات مطرح شده در این فصل می‌توان مقایسه زیر را برروی روش‌های مختلف ارائه نمود.

جدول2 - جدول مقایسه روش پیشنهادی با روش های موجود

نام روش	COSIN	LDA	PLSI	LSI

حجم پارامترهای ورودی	1	n	1	1
پیچیدگی محاسباتی آماده سازی پارامترهای ورودی	Try and Error	$O(n^3)$	Try and Error	Try and Error
پیچیدگی محاسباتی الگوریتم	$\sim O(m^2c^2)$	NP-hard	NP-hard	$O(nc \log n)$
آیا روش مبتنی بر تکرار است	Yes (partially)	Yes	Yes	Yes
آیا روش به تابع توزیع داده‌ها وابسته است	No	No	No	Yes

براساس این توضیحات مشخص است که COSIN روشی است با ویژگی‌های مثبت LSI و همچنین روش‌های مستقل از توزیع مانند LDA و PLSI. این هدفی است که در ابتدای فصل به عنوان هدف اصلی به آن اشاره شد. از این رو می‌توان گفت که روشی مناسب برای استفاده در مجموعه داده‌های بزرگ در اختیار داریم.

۴. وزن دهی معنایی

در این فصل رویکردی کاملا متفاوت به مقوله پردازش متن نسبت به فصل ۳ خواهیم داشت. در این رویکرد جدید هدف اصلی به کارگیری اصول زبان شناختی و همچنین ایجاد عمومیت لازم در این روش ها جهت استفاده در کاربردهای متفاوت متن کاوی می باشد. علاوه بر آن مسئله ای که به عنوان هدف در این فصل در نظر گرفته شده است، مسئله تعیین وزن لغات در صفحه می باشد.

همانطور که واضح است به کاربردن رو شهای ریاضی برروی داده های متنی مسلزم ارائه متون در قالب زبان ریاضی می باشد. در چنین ارائه ای که معمول ترین آن ارائه برداری است باید ویژگی های اساسی شناسایی شده و برای هر یک از این ویژگی ها وزنی در نظر گرفته شود. در یک متن ویژگی ها به صورت معمول لغات انتخاب می گردند و سپس لغات براساس معیارهایی در هر صفحه وزن دهی می شوند. این معیارها عمدتا معیارهایی آماری بوده و پایه اصلی آن ها را بسامد رخداد لغت در صفحه تشکیل می دهد.

روش های آماری اگر چه تا کنون توانسته اند کمک زیادی در جهت متن کاوی نمایند و راه های زیادی را هموار سازند اما به علت ضعف های ذاتی خود قادر به افزایش دقت از حدی بیشتر نیستند. این ضعف ها ناشی از عدم توجه این معیارها به بعضی از ویژگی های متن می باشد. در زیر سعی شده است تا اهم این مشکلات به صورت مختصر بیان گردند:

- **عدم توجه به اطلاعات با اهمیت در متن:** یکی از مشکلات اصلی این روش ها عدم توجه به بعضی اطلاعات بسیار مهم می باشد که به راحتی قابل استخراج از متن می باشند. یکی از انواع این اطلاعات ترتیب رخداد لغات در یک متن می باشد. پرواضح است که اطلاعات مربوط به ترتیب

رخداد لغات می تواند اطلاعات زیادی را در خصوص اهمیت لغات و همچنین روابط معنایی بین لغات در اختیار قرار دهد. این اطلاعات به سادگی توسط معیارهای آماری در نظر گرفته نمی شوند.
البته اطلاعات ترتیب لغات دارای دو وجه متفاوت می باشد که در بالا نیز تا حدودی به آن اشاره شد. در وجه اول همین ترتیب ظاهری لغات مورد نظر است که بعضا در قالب اصطلاحات و یا هم‌آیی بین لغات نمایان می شود. در این شکل البته روش هایی مانند n-gram ها [BRO92, PER95] ابداع شده اند تا بتوانند تا حدودی ترتیب لغات را مد نظر قرار دهند. اما حتی همین روش ها نیز نمی توانند هم‌آیی بین لغات را شناسایی نمایند. از این رو ضعفی عمده در این حیطه در روش های آماری موجود است.
در وجه دوم ترتیب لغات، منظور ترتیب های پر کاربرد مانند وجه قبلی نیست. بلکه در یک متن برای درک مطلب باید یک تجمیع معنا بین عناصر معنایی متن صورت گیرد. به عبارت دیگر معنایی که از یک متن برداشت می شود حاصل تجمیع معنای قابل برداشت از اجزای مختلف آن می باشد. حال فرض کنید متنی داریم شامل دو جمله S_1 و S_2. در روش های آماری حاصل تجمیع معنایی در چنین متنی جمع برداری این دو جمله است. حال آنکه پر واضح است چنانچه ترتیب این دو جمله در متن تغییر نماید معانی کاملا متفاوتی از متن برداشت خواهد شد. به عبارت دیگر در روش های معنایی ترتیب رخداد لغات تأثیری در معنایی که در کل از یک متن برداشت می شود ندارد و این روش ها به متن به عنوان یک کیف از لغات[۵۸] نگاه می کنند. کاملا واضح است که چنین نگاهی به معنا و به متن کاملا اشتباه است. می توان نتیجه گرفت تجمیع معنایی عملگری فراتر از تجمیع برداری معمول است. در تجمیع برداری ترتیب تأثیری در حاصل نخواهد داشت و یک عملگر جابجایی پذیر[۵۹] است. حال آنکه تجمیع معنایی عملگری جابجایی پذیر نیست.
در یک جمع بندی باید گفت ترتیب، دو نوع اطلاعات را در اختیار می گذارد: اطلاعات مربوط به ترکیب های پرکاربرد از لغات و اطلاعات مربوط به نحوه تجمیع معنایی اجزاء متن. روش های مبتنی بر آمار هیچیک از این دو حوزه اطلاعات را نمی توانند پوشش دهند. در یک نگاه زبان شناختی، در یک متن و یا گفتار ما به غیر از عناصر لغوی، عناصر بالای زنجیره گفتار را داریم که وظیفه اصلی بیان مفهوم بر عهده این عناصر است. یکی از این نمونه عناصر ترتیب است. عناصر دیگر مانند لحن نیز وجود دارند که در متن پوشش داده نمی شوند و نویسنده سعی می کند از طریق عناصر دیگر مانند ترتیب لغات و جملات و یا نشانگرها مفهوم را منتقل سازد.

---

[5] Bag of Word
[5] Commutative
[8]
[9]

- **عدم انتخاب صحیح اجزاء معنایی:** روش های آماری به صورت معمول لغات را به عنوان عناصر معنایی پایه در نظر می گیرند. به عبارت دیگر میزان اهمیت یک لغت در یک متن بدون درنظر گرفتن انواع دیگر کلمات به کار برده شده در متن محاسبه می شود. اما در حوزه زبانشناختی و فلسفه زبان دیدگاه های کاملا متفاوتی نسبت به این مسئله وجود دارد. در دیدگاه فلسفی به زبان لغات تنها نشانه هایی هستند برای ایجاد ارتباط و به تنهایی حامل معنا نیستند، بلکه مجموعه لغات هستند که می توانند نمایندگی معنا را برعهده داشته باشند. به عنوان مثال Wittgenstein در [WIT53] می گوید: «معنا تنها الگوی استفاده است». به عبارت دیگر این لغات نیستند که نمایانگر معنا هستند. بلکه نحوه استفاده از آن ها می تواند نشان دهنده معنایی باشد که در ذهن فرد قرار دارد. و یا در جایی دیگر Firth [FIR68] می گوید: «شما یک کلمه را براساس مجموعه لغاتی که به همراه آن هستند درک می کنید». براساس این گزاره ها که در حوزه فلسفه زبان و زبانشناسی کاربرد فراوانی دارند واضح است که لغات تنها نشانه هایی هستند و این الگوی استفاده از آن ها است که معنا را تحمیل می نماید. در نتیجه انتخاب لغات به عنوان اجزاء پایه معنایی می تواند منجر به ایجاد خطا در نتایج گردد.
- **متن های طولانی:** یکی دیگر از مشکلات روش های آماری نحوه برخورد با متن های با طول زیاد است []. در چنین متن هایی میزان اهمیتی که برای لغات محاسبه می شود به صورت عمومی بسیار کاهش می یابد. در نتیجه ای کاهش مقادیر متن مورد نظر به هیچ عنوان نمی تواند معنای مورد نظر را نمایان سازد و کلمات کلیدی نیز مانند دیگر کلمات فاقد اهمیت بالا فرض می شوند. این مشکل ناشی از دیدگاه روش های آماری به متن است. در دیدگاه روش های آماری یک متن مجموعه ای همگن از لغات است. حال آنکه چنین فرضی درست نیست و در یک متن ما بخش های دارای اهمیت بالا و بخش های دارای اهمیت کم را داریم. بر هیمن اساس می توان بین لغات حاضر در بخش ها تفاوت قائل شد و وزن بیشتری را برای بعضی لغات در نظر گرفت و از این طریق تا حد زیادی بر مشکل متون با حجم بالا فائق آمد.

همین مسائل انگیزه های اصلی برای ارائه راهکاری جدید که بتواند مشکلات مطرح شده را مرتفع سازد فراهم می سازند.

۴,۱. ارائه یک قالب کاری

راه حلی که در اینجا ارائه خواهد شد برگرفته از یک تابع حدس جدید در حوزه روابط معنایی در متن می باشد. این تابع حدس تحت عنوان نظریه مرکزیت ارائه گردیده است [GRO95]. این نظریه که پیش از این در مورد آن توضیح داده شده است مبتنی بر یک قاب ۳ جمله ای می باشد. اما علیرغم کاربردهایی که این

نظریه پیدا کرده است و توانسته است جوای های مناسبی نیز در اختیار بگذارد دارای مشکلاتی نیز می باشد. یکی از این مشکلات محلی بودن این نظریه می باشد. همانطور که گفته شد این نظریه تنها سه جمله متوالی را در هر بررسی مورد توجه قرار می دهد. این مسئله کاملا نشان می دهد که این روش بسیار محلی عمل می نماید. مشکل بعدی این است که در قالب جاری تعریف این نظریه تنها می توان آن را برای مجموعه سندهای معمول مورد استفاده قرار داد، در حالی که در بسیاری از موارد با مجموعه داده هایی مواجه هستیم که دارای شکل معمول اسناد نیستند و داده هایی با فرم های مختلف در یک پیکره ممکن است موجود باشند. در نهایت می توان آخرین مشکل را عدم وجود یک مدل محاسباتی مناسب برای این نظریه دانست. در کل سعی شده است تا در این فصل هر سه مشکل تا حد ممکن پوشش داده شوند. به عبارت دیگر روشی ارائه خواهد شد که تنها به داده های محلی توجه نداشته باشد، دارای عمومیت بیشتری باشد و در نهایت بتوان یک مدل محاسباتی قابل قبول برای آن در اختیار داشت.

برای دستیابی به چنین هدفی لازم است تا پیش از ارائه روش به ارائه یک قالب کاری بپردازیم. در این قالب کاری سعی خواهد شد تا با ارائه چند تعریف و رسمی سازی در تعاریف مصطلح موجود راه برای روشی که در بالا توضیح دادیم هموار گردد. در خصوص نظریه مرکزیت نیز توضیحات کافی در بخش ۸,۶ ارائه گردیده است.

۴,۱,۱.گسترش مفهوم جمله

در بسیاری از کاربردهای داده کاوی و متن کاوی مجموعه داده های ورودی دارای شکل استاندارد مورد انتظار برای یک جمله و یا متن نیستند. به عنوان مثال یک موتور جستجو را در نظر بگیرید. در چنین کاربردی یک پرس و جوی ارسال شده توسط یک کاربر در اصل می تواند به عنوان یک جمله که دارای بار معنایی مشخصی می باشد محسوب گردد. به عبارت دیگر هدف یک کاربر از یک پرس و جو کاملا مشخص می باشد و می تواند ارزش معنایی برای آن قائل گردید. چنین عبارتی که مجموعه ای از کلمات کلیدی است به هیچ وجه دارای ساختار معمول یک جمله نیست و نمی توان بسیاری از الگوریتم ها و یا روش های مبتنی بر خواص ساختاری متن را برای آن به کار برد. اما همین عبارت دارای کلیه خواص معنایی یک جمله می باشد و حتی ترتیب قرارگیری لغات در چنین عبارتی در نتیجه جستجو باید تأثیر داده شود.

به همین منظور باید بتوان مفهوم جمله را از آن چه در حال حاضر شناخته می شود کمی عمومی تر نمود تا بتواند بسیاری ساختارهای دیگر مانند آنچه در بالا مطرح شد را نیز پوشش دهد. در این فصل این مفهوم جدید قطعه معنایی[6] نام گذاشته شده است. در کاربردهای متداول یک قطعه معنایی در اصل همان جمله

---

[6] Semantic Chunk

است. اما در کاربردهای دیگر که دارای داده هایی با ساختارهای نامتعارف و یا بدون تطابق با ساختارهای دستوری زبان هستند، یک قطعه معنایی کوچک ترین جزء متن محسوب می شود که بتواند نمایندگی معنایی را بنماید. به عبارت دیگر در قالب کاری که در اینجا پیشنهاد می شود این لغات نیستند که حامل معنا هستند، بلکه قطعات معنایی حاملین معنا در متن محسوب می شوند و همانطور که پیش از این گفته شد، لغات تنها نشانه هایی هستند که به تنهایی فاقد معنا می باشند.

تعریف ۱ (قطعه معنایی)

در یک نگاه رسمی یک **قطعه معنایی** توسط یک سه گانه مانند $<\mathcal{A}, \mathcal{S}, \preccurlyeq>$ نمایش داده می شود که در آن

- $\mathcal{A}$ مجموع الفبای زبان محسوب می شود. منظور از مجموعه الفبا در اینجا همان مجموعه نشانه ها و یا همان لغات می باشد و نه حروف.
- $\mathcal{S}$ یک زیرمجموعه از مجموعه تمام زیرمجموعه های $\mathcal{A}$ می باشد و یا به عبارت دیگر $\mathcal{S} \in P(\mathcal{A})$.
- $\preccurlyeq$ یک ترتیب جزئی برروی مجموعه اعضای $\mathcal{S}$ می باشد.

مثال ۱

به عنوان مثال مجموعه الفبای {گربه، شیر، خوردن، پستاندار} را در نظر بگیرید. در نتیجه یک مجموعه مانند {گربه، شیر، خوردن} و یک ترتیب جزئی مانند گربه $\preccurlyeq$ شیر $\preccurlyeq$ خوردن می توانند نشان دهنده یک جمله باشند.

از این مثال می توان برداشت کرد که ترتیب جزئی مورد نظر نشان دهنده اهمیت لغات در قطعه معنایی می باشد و ارتباطی با ترتیب لغات در قطعه معنایی ندارد. علاوه بر تعریف فوق باید دنیای معنایی نیز تعریف گردد که در قسمت های بعد مورد استفاده قرار می گیرد.

تعریف ۲ (دنیای معنایی)

دنیای معنایی مجموعه تمام قطعات معنایی ممکن است و با Σ نشان داده می شود.

مثال ۲

یک عبارت پرس و جو مانند «java service daemon create» را در نظر بگیرید. در چنین نمونه ای که فاقد ساختار معمول جمله می باشد می توان انواع توابع حدس را برای تعیین ترتیب جزئی بین لغات

استفاده نمود. ساده ترین حدس می تواند ترتیب کاربرد لغات باشد و هر کلمه ای که زودتر حضور پیدا کرده است دارای اهمیت بالاتر فرض شود. یک حدس دیگر رتبه یک لغت در کل دنیای معنایی مسئله است. یک حدس دیگر می تواند رتبه بندی لغات به صورت محلی و براساس روابط معنایی بین آن ها که توسط یک شبکه دانش تعیین می شود باشد. همچنین رتبه بندی براساس نمایه[۶]گاربر و علاقه مندی های وی نیز می تواند یک گزینه مناسب برای تعیین ترتیب جزئی مورد نظر می باشد.

براساس این دو مثال می توان دریافت که مفهوم قطعه معنایی تنها یک تعریف علمی صرف نیست بلکه یک واقعیت و مفهوم کاملا کاربردی است که دارای کاربردهای بسیار متنوعی می باشد. همچنین این تعریف امکان ارائه الگوریتم های داده کاوی مبتنی بر جمله را در قالبی کاملا محاسباتی فراهم می آورد. یکی دیگر از مزایای آن نیز امکان جدا کردن بخش های مختلف الگوریتم های داده کاوی از یکدیگر می باشد. به عبارت دیگر می توان در ارائه یک الگوریتم جدید تنها به جنبه های مهم آن پرداخت و با فرض وجود یک ترتیب جزئی کمتر درگیر سطوح پایین پردازش زبان شد. معمولا همین الگوریتم های سطح پایین قسمت های غیرقطعی در پردازش زبان را تشکیل می دهند و جدا کردن آن ها از سطوح پردازشی بالاتر بسیار مفید می باشد.

۴،۱،۲. گسترش مفهوم زمینه

یک مفهوم مهم دیگر در پردازش متن، مفهوم زمینه می باشد. معمولا زمینه و سند با یکدیگر هم ارز گرفته می شوند. اما در حالت کلی این فرض درستی نیست. ارائه یک تعریف رسمی از زمینه شاید بتواند کمک خوبی جهت تشخیص و جداسازی مفاهیم زمینه و سند از یکدیگر باشد. به همین منظور در این بخش سعی شده است تا چنین تعریفی ارائه گردد. البته تعیین دقیق حوزه زمینه کار دشواری است. یکی از کاربردهای روش ارائه شده در این فصل می تواند همین تشخیص زمینه با توجه به پیوستگی هایی که شناسایی می شوند می باشد.

تعریف ۳ (زمینه)

فرض کنید Σ^* مجموعه تمام رشته هایی است که با استفاده از به هم چسباندن اعضای Σ (دنیای معنایی) به دست می آیند. آنگاه یک **زمینه** را می توان با $\mathcal{C}$ نمایش داد که $\mathcal{C} \in \Sigma^*$.

در این تعریف یک ترتیب به صورت ضمنی بین قطعات معنایی مورد استفاده در یک زمینه وجود دارد و آن هم ترتیب اتصال این قطعات در زمینه می باشد. این ترتیب تنها برای مرور یک زمینه مورد استفاده واقع می

---

[6] Profile [1]

شود و نشان دهنده اهمیت معنایی این قطعات نیست. یکی از نتایج روش ارائه شده در این فصل رتبه بندی قطعات معنایی براساس اهمیت معنایی آن ها می باشد. مشابه با همین رویکرد در نظریه مرکزیت در حوزه جملات دیده می شود. اما در روش ارائه شده در این فصل سعی شده است تا مشکلات مطرح شده در نظریه مرکزیت پوشش داده شوند. در اینجا مفهوم زمینه محدود به کل یک سند نشده است و معیار کرانه یک زمینه می توان توسط کاربرد و براساس نیاز تعیین گردد. از سوی دیگر همانطور که گفته شد روش ارائه شده در این فصل به عنوان معیاری برای تعیین محدوده زمینه می تواند مورد استفاده قرار گیرد.

۴،۱،۳. گسترش مفهوم دانش پیش زمینه

یک قطعه معنایی مجموعه بسیار کوچکی از لغات را پوشش می دهد. از همین رو به تنهایی نمی تواند معنای مورد نیاز را در اختیار قرار دهد. البته از دیدگاه یک انسان قطعه معنایی در زمینه مشاهده می شود اما یک ماشین در هنگام مشاهده یک قطعه معنایی نمی تواند به صورت مستقیم متوجه زمینه گردد از این رو باید زمینه قطعه معنایی را از طریق گسترش دایره لغات در اختیار آن قرار داد. در بخش ۲،۳،۲ نمونه هایی از این رویکرد معرفی گردیدند که مجموعه های دانش مانند WordNet برای غنا بخشی به معنای یک قطعه معنایی استفاده می نمودند. در اصل WordNet یک مجموعه از لغات است که روابط معنایی مستقیم بین آن ها نیز تعیین شده است. از جمله این روابط می توان به رابطه هم معنایی و دیگر روابط مستقل از متن بین لغات اشاره نمود.

مثال ۳

به عنوان مثال برای کلمه ای مانند «get» می توان کلمات { become, let, acquire, receive, find, obtain, …} را در مجموعه کلمات مرتبط آن در WordNet مشاهده نمود [WRDNT].

اما یک مشکل موجود در WordNet عدم پوشش روابط معنایی غیرمستقیم مانند روابط هم آیی می باشد. برای پوشش چنین موارد بعضا از روابط معنایی شناسایی شده توسط روش های خوشه بندی که عمدتا نیز روابط معنایی غیرمستقیم هستند استفاده می شود [BAD09].

مثال ۴

به عنوان مثال برای کلمه «search» نمی توان کلمه «query» را در بین روابط معنایی آن در WordNet یافت [WRDNT]. در حالی که این دو کلمه دارای رابطه معنایی قدرتمندی با یکدیگر هستند و روش هایی مانند LSI می توانند چنین روابط معنایی را کشف نمایند.

با توجه به موارد گفته شده در فوق ضرورت تقویت معنایی قطعات معنایی کاملا واضح است. اما برای اینکه بتوانیم به جامعیت لازم دست یابیم لازم است که به تعریفی رسمی از دانش پیش زمینه بپردازیم تا با استفاده از آن بتوان روش های مختلف گسترش معنایی قطعات معنایی را پوشش داد. همچنین از این طریق می توان وظیفه انتخاب روش مناسب را برعهده کاربرد مورد نظر قرار داد و جامعیت لازم را برای روش ارائه شده در این پایان نامه فراهم نمود.

تعریف ۴ (دانش پیش زمینه)

یک **دانش پیش زمینه** را می توان با $\mathcal{B}$ نشان داد که عبارت است از یک سه گانه مانند $< \mathcal{A}, \mathrm{E}, \mathcal{L} >$ که

- $\mathcal{A}$ مجموعه الفبای زبان است.
- E زیر مجموعه از $\mathcal{A}^2$ است و یا به عبارت دیگر $\mathrm{E} \subseteq \mathcal{A}^2$.
- $\mathcal{L}$ یک تابع از E به $[0,1)$ است.

برای ایجاد نگاشت بین یک مجموعه دانش مانند WordNet و این تعریف می توان به عنوان مثال یک وزن دهی برروی انواع روابط معنایی موجود در WordNet انجام داد. در مجموعه های حاصل از خوشه بندی چنین وزن دهی به صورت پیش فرض وجود دارد.

با استفاده از چنین تعریف رسمی می توان الگوریتم ها را به صورت کاملا رسمی بیان نمود و حتی به بررسی زمان اجرای آن ها نیز پرداخت. در بخش ۴,۲ دقیقا از همین طریق به معرفی یک الگوریتم، اثبات درستی آن و محاسبه زمان اجرای آن می پردازیم.

۴,۲.وزن دهی لغات

همانطور که در ابتدای فصل اشاره شد، هدف ارائه روشی برای وزن دهی لغات است که بتواند مشکلات مطرح شده برای روش های آماری را مرتفع سازد. ایده اصلی هم از نظریه مرکزیت جهت وزن دهی براساس داده های زبان شناختی گرفته شده است. اما این روش به تنهایی دارای مشکلاتی است که عمدتا ناشی از دیدگاه محلی آن به متن می باشد. به همین مظور و جهت رفع این مشکلات قالب کاری جدید در بخش ۴,۱ ارائه گردید که می تواند راهگشای حرکت به سمت روشی باشد که دارای مشکلات مورد نظر نیست. این راه کار جدید دارای سه بخش اصلی می باشد. در گام اول باید بردار قطعات معنایی ساخته شود. این بردار برای هر قطعه معنایی نشان دهنده معنای آن می باشد. سپس در گام دوم بردار زمینه برای متن ساخته شود. بردار زمینه، برداری است که در هر نقطه از متن می تواند نشان دهنده مفهوم مورد نظر متن تا آن نقطه باشد. به عبارت دیگر بردار زمینه نمایانگر همان سیر تکاملی معنا در زمینه می باشد. برای محاسبه

بردار زمینه از دو الگوریتم گسترش یافته مبتنی بر نظریه مرکزیت استفاده می گردد. در نهایت و در گام پایانی با استفاده از بردار زمینه کار وزن دهی به لغات صورت می گیرد. البته در اینجا هدف تعیین وزن محلی می باشد و برای وزن سراسری و همین طور عامل نرمال سازی می توان از روش های معمول موجود استفاده نمود.

۴،۲،۱. بردار قطعه معنایی

همان طور که گفتیم واحد معنایی در روش ارائه شده در این فصل قطعات معنایی می باشند. پس اولین گام شامل ایجاد بردار قطعات معنایی است. به عبارتی باید به ازای هر قطعه معنایی مانند $\mathcal{S}_i$ یک نمایش برداری در زمینه $\mathcal{C}$ فراهم گردد. معمول ترین روش به این صورت است که به ازای هر قطعه معنایی برداری که تعداد ابعاد آن برابر با اندازه مجموعه الفبا می باشد تشکیل می شود. تمام اعضای این بردار صفر هستند مگر برای مجموعه اعضایی از الفبا که عضو قطعه معنایی مورد نظر نیز هستند. این نحوه نمایش همانطور که در بخش های قبل گفتیم نمی تواند برای کاربرد ما مناسب باشد. با این وجود به دلیل کاربردهای بعدی باید تعریفی از آن ارائه نمایید. اما به دلیل فوق آن را نمایش برداری کهاد نام می گذاریم و به صورت رسمی در قالب زیر تعریف می گردد.

تعریف ۵ (نمایش برداری کهاد قطعه معنایی)

یک قطعه معنایی مانند $\mathcal{S}_i$ از یک زمینه مانند $\mathcal{C}$ را می توان با استفاده از برداری مانند $\overrightarrow{\mathcal{s}_i}$ نمایش داد که به این بردار **نمایش برداری کهاد قطعه معنایی**[۶۲] گفته می شود و داریم:

- $\overrightarrow{\mathcal{s}_i}$ عضوی از یک جبر یکایی[۶۳] می باشد.
- تعداد ابعاد $\overrightarrow{\mathcal{s}_i}$ برابر است با $|\mathcal{A}|$ و یا به عبارت دیگر $\dim(\overrightarrow{\mathcal{s}_i}) = |\mathcal{A}|$.
- $\overrightarrow{\mathcal{s}_{i,j}} = 1 \Leftrightarrow t_j \in \mathcal{S}_i, t_j \in \mathcal{A}$ $\boldsymbol{otherwise}$ $\overrightarrow{\mathcal{s}_{i,j}} = 0$

همانطور که گفته شد مجموعه لغات موجود در یک قطعه معنایی بسیار محدود هستند و نمی توانند به خوبی نمایش دهنده معنای مورد نظر از قطعه معنایی مورد نظر باشند. به همین منظور علاوه بر نمایش برداری کهاد که در بالا ارائه گردید یک نمایش برداری مهاد[۶۴] نیز برای قطعه معنایی تعریف می گردد. الگوریتم زیر می تواند این نمایش برداری را در اختیار قرار دهد. این نمایش برداری جدید از طریق گسترش نمایش برداری کهاد با استفاده از یک دانش پیش زمینه صورت می گیرد.

---

[6] Minor Vector Representation of Semantic Chunk
[6] Unital Algebra
[6] Major Vector Representation

الگوریتم ۱ (نمایش برداری مهاد)

شبه کد ۵ – الگوریتم محاسبه نمایش برداری مهاد

Input: Semantic Chunk $\mathcal{S}_i$, background knowledge $\mathcal{B} \sim< \mathcal{A}, \mathrm{E}, \mathcal{L} >$, stop threshold δ.

1. $h = CreateFibonacciHeap()$
2. **create vector** $\overrightarrow{\mathcal{S}_i}$ **with** $\dim(\overrightarrow{\mathcal{S}_i}) = |\mathcal{A}|$ **and fill it with zeros**
3. **for all** elements t_j **in** $\mathcal{S}_i$ **do**
4. $\quad t_j \to Weight = 1$
5. $\quad t_j \to Visited =$ **False**
6. $\quad h \to Insert(t_j)$
7. **do**
8. $\quad t_j = h \to DeleteMax()$
9. $\quad \overrightarrow{\mathcal{S}_{i,j}} = t_j \to Weight$
10. $\quad t_j \to Visited =$ **True**
11. $\quad$ **for all** elements $e_k \in \mathrm{E}$ **and** $t_j \in e_k$ **do**
12. $\quad\quad t' = e_k \to GetPair(t)$
13. $\quad\quad$ **if not** $t' \to Visited$ **then**
14. $\quad\quad\quad$ **if** $t_j \to Weight \times \mathcal{L}(e_k) > \delta$ **then**
15. $\quad\quad\quad\quad$ **if** $t' \to Weight == 0$ **then**
16. $\quad\quad\quad\quad\quad t' \to Weight = t_j \to Weight \times \mathcal{L}(e_k)$
17. $\quad\quad\quad\quad h \to Insert(t')$
18. $\quad\quad\quad$ **elseif** $t' \to Weight < t_j \to Weight \times \mathcal{L}(e_k)$ **then**
19. $\quad\quad\quad\quad h \to DecreaseKey(t_j \to Weight \times \mathcal{L}(e_k), t')$
20. **while** (**not** $h \to IsEmpty()$)

Output: Major vector representation $\overrightarrow{\mathcal{S}_i}$ for a given Semantic Chunk $\mathcal{S}_i$.

قضیه ۱: الگوریتم ۱ بزرگ ترین وزن های ممکن برای لغات را محاسبه می نماید.

اثبات: به عبارت دیگر زمانی که در خط هشت الگوریتم یک لغت به عنوان پروزن ترین لغت انتخاب می شود، آنگاه وزن این لغت در آن لحظه بیشترین وزن ممکن برای آن لغت می باشد. به برهان خلف فرض کنیم که این طور نیست. به عبارت دیگر کلمه ای مانند t_j وجود دارد که قبلا مشاهده شده است و سپس برای یک کلمه جدیدا مشاهده شده مانند t_k که دارای ارتباطی با t_j نیز هست و این ارتباط با e_m نمایش داده می شود مقدار $t_k \to Weight \times \mathcal{L}(e_m)$ بیش از مقدار $t_j \to Weight$ می باشد. اما چون تابع $\mathcal{L}$ اعضای مجموعه E را به بازه $[0,1)$ نگاشت می کند از فرض خلف خواهیم داشت

$$t_k \to Weight \times \mathcal{L}(e_m) > t_j \to Weight, 0 < \mathcal{L}(e_m) \leq 1 \Rightarrow \\ t_k \to Weight > t_j \to Weight \qquad (۱۴)$$

این رابطه برای تمام لغاتی که پیش از t_k انتخاب شده اند نیز برقرار خواهد بود. بنابراین t_j نمی توانسته است پیش از t_k انتخاب شود که این با فرض در تناقض است.

قضیه ۲: پیچیدگی زمانی الگوریتم ۱ برابر است با $O(|\mathrm{E}| + |\mathcal{A}|\log|\mathcal{A}|)$.

اثبات: براساس الگوریتم ۱ و همچنین قضیه ۱ می توان نتیجه گرفت که هر لغت تنها یک بار انتخاب می شود. بنابراین این الگوریتم به $O(|\mathcal{A}|)$ تعداد عمل *Insert* احتیاج دارد. همچنین تعداد اعمال *DeleteMax* مورد نیاز حداکثر $O(|\mathcal{A}|)$ و حداکثر تعداد اعمال *DecreaseKey* مورد نیاز برابر $O(|\mathrm{E}|)$ خواهد بود. با توجه به اینکه در این الگوریتم از Fibonacci heap [FRE87] استفاده شده است می توان گفت که پیچیدگی زمانی مورد نیاز برای یک عمل *Insert* و یا *DecreaseKey* برابر $O(1)$ خواهد بود. همچنین پیچیدگی زمانی یک عمل *DeleteMax* برابر $O(\log|\mathcal{A}|)$ می باشد. با توجه به این اعداد می توان نتیجه گرفت که پیچیدگی زمانی کل الگوریتم در بدترین حالت برابر خواهد بود با:

$$O(|\mathrm{E}| + |\mathcal{A}|\log|\mathcal{A}|) \qquad (۱۵)$$

البته کاملا واضح است که در شرایط واقعی بسیار بعید است که بدترین شرایط به وجود آیند. زیرا به ندرت پیش می آید که اعضای یک قطعه معنایی با کل کلمات یک دانش پیش زمینه دارای روابط معنایی نزدیک باشند. در نتیجه در واقعیت زمان اجرای الگوریتم بسیار کمتر از مقدار ذکر شده در قضیه ۲ است. برای ارائه یک تحلیل واقعی تر به قضیه ۳ توجه فرمایید.

قضیه ۳: چنانچه تعداد متوسط لغاتی که با یک لغت در دانش پیش زمینه در ارتباط هستند برابر با m و حداکثر وزن یک ارتباط σ و متوسط اندازه ی قطعه معنایی n باشد، آنگاه زمان اجرای الگوریتم۱ به صورت متوسط برابر با:

$$O\left(n \times \frac{m^{\log_\sigma \delta} - 1}{m-1} + n \times \frac{m^{\log_\sigma \delta} - 1}{m-1} \times \log\left(n \times \frac{m^{\log_\sigma \delta} - 1}{m-1}\right)\right) \qquad (۱۶)$$

خواهد بود .

اثبات: بلندترین زنجیره لغات قابل پیمایش در این حالت برابر خواهد بود با $\log_\sigma \delta$. این زنجیره برای یک لغت درختی با همین عمق ایجاد خواهد کرد که تعداد اعضای این درخت برابر خواهد بود با $\frac{m^{\log_\sigma \delta}-1}{m-1}$. با فرض اینکه هر قطعه معنایی دارای n عضو است تعداد کل لغات و ارتباطات پیموده شده هر یک برابر $n \times \frac{m^{\log_\sigma \delta}-1}{m-1}$ می باشد. درنتیجه زمان اجرای الگوریتم برابر با عبارت ۱۶ خواهد بود.

برای درک بهتر این تفاوت فرض کنید تعداد کل لغات یک دانش پیش زمینه ۱،۰۰۰،۰۰۰ عدد باشد. همین طور فرض کنید تعداد روابط تعریف شده نیز ۱۰،۰۰۰،۰۰۰ عدد باشد و تعداد متوسط لغات یک قطعه معنایی ۸ عدد باشد. همین طور شرط پایان الگوریتم مقدار ۰,۶ و حداکثر وزن در دانش پیش زمینه ۰,۹ باشد. در نتیجه با توجه به قضیه ۲ زمان اجرا در بدترین حالت برابر خواهد بود با

$$O(|\mathrm{E}| + |\mathcal{A}|\log|\mathcal{A}|) = 10000000 + 1000000 \times 20 = 30{,}000{,}000$$

در حالیکه با توجه به قضیه ۳ خواهیم داشت

$$n \times \frac{m^{\log_\sigma \delta} - 1}{m - 1} = 8 \times \left(\frac{10^{4.8} - 1}{10 - 1}\right) \simeq 56084$$

و در نهایت زمان اجرا برابر خواهد بود با ۹۴۰۸۰۹ که عددی بسیار کمتر از بدترین شرایط می باشد. حتی همین عدد نیز برای شرایط بد محاسبه شده است. زیرا بسیاری از روابط دارای وزن ۰,۹ نیستند و زنجیره ها دارای طول بسیار کمتری خواهند بود.

۴,۲,۲. بردار محتوا

در الگوریتم ارائه شده در این فصل بردار زمینه نقش اصلی را ایفا می نماید. بردار زمینه در اصل نمایش دهنده معنا در هر نقطه از زمینه می باشد. این بردار براساس دو فرضیه شکل گرفته است که عبارتند از:

- اسناد برای استفاده انسان ها ساخته می شوند و یا توسط انسان ها ساخته می شوند و از این رو عمدتا دارای ماهیت زنجیره ای و یا سری مانند هستند.
- نظریه مرکزیت توانسته است به عنوان یک تابع حدس خوب برای رتبه بندی جملات در متن مورد استفاده قرار گیرد.

با استفاده از بردار زمینه قصاد داریم تا در هر نقطه میزان اهمیت قطعه معنایی جاری را تعیین نماییم. بر این اساس موقعیت در یک زمینه دارای اهمیت خواهد بود زیرا اهمیت یک قطعه معنایی یکسان در موقعیت های مختلف در یک زمینه دارای مقادیر متفاوتی خواهد بود. همچنین واضح است که ترتیب قطعات معنایی نیز در برداشت معنایی از زمینه تأثیر خواهد داشت. تقریبا این شرایط جزء شرایط ایده آلی است که در یک الگوریتم وزن دهی مورد نیاز است. موقعیت و ترتیب هر دو تأثیر به سزایی در تعیین میزان اهمیت قطعات معنایی و به تبع آن لغات خواهد داشت.

در روش ارائه شده در این فصل ایده اصلی نظریه مرکزیت مورد استفاده قرار گرفته است. به عبارت دیگر این پیوستگی است که تعیین کننده میزان اهمیت قطعات معنایی در یک زمینه می باشد. اما برای فائق آمدن بر

مشکلات نظریه مرکزیت قالب کاری جدیدی را در بخش ۴,۱ ارائه کردیم. سپس با استفاده از همین قالب کاری بردارهای مهاد قطعات معنایی ایجاد شد. در کاربردهای معمول با جمع نمودن همین بردارها به بردار معنای زمینه دست می یابیم. اما چنین تجمیعی خواص مورد نیاز ما را دارا نمی باشد.

در روش ارائه شده در این فصل برای تجمیع بردارهای مهاد قطعات معنایی از یک عملگر دیگر به نام ضرب افزودنی[5][6] استفاده می گردد. برای دو بردار نمونه مانند α و β ضرب افزودنی برابر خواهد بود با

$$[\alpha_1, \alpha_2, \dots] \boxplus [\beta_1, \beta_2, \dots] = [\alpha_1\beta_1, \alpha_1\beta_2 + \alpha_2\beta_1, \alpha_1\beta_3 + \alpha_3\beta_1, \alpha_1\beta_4 + \alpha_4\beta_1, \dots] \quad (۱۷)$$

این فرآیند برای بردارهای مهاد قطعات معنایی به صورت زیر انجام می گردد

- بردار مهاد قطعه معنایی جاری با اضافه کردن یک بعد به آن که مقدار این بعد برابر ۱ است گسترش داده می شود. همین طور بردار زمینه با اضافه کردن یک بعد به آن که مقدار این بعد برابر عامل α است گسترش داده می شود. به عبارت دیگر بردار $\overrightarrow{S_i}$ تبدیل خواهد شد به $[1, \overrightarrow{S_i}]$ که این بردار جدید با $\overrightarrow{\mathbb{S}_i}$ نمایش داده می شود. همین طور بردار زمینه به $[\alpha, \overrightarrow{CV_i}]$ تبدیل خواهد شد که با $\overrightarrow{\mathbb{CV}_i}$ نمایش داده می شود. واضح است که در این حالت تعداد ابعاد هر دو بردار برابر خواهد بود با $|\mathcal{A}| + 1$.
- مقدار α برای اولین قطعه معنایی برابر ۱ خواهد بود.
- بردار زمینه $\overrightarrow{CV_i}$ در اصل بردار زمینه پس از پردازش i اُمین قطعه معنایی می باشد و واضح است که $\overrightarrow{CV_1} = \overrightarrow{S_1}$.
- و در نهایت داریم:

$$[\alpha, \overrightarrow{CV_{i+1}}] = \overrightarrow{\mathbb{CV}_i} \boxplus \overrightarrow{\mathbb{S}_{i+1}} \quad (۱۸)$$

کاملا واضح است که عملگر مورد نظر غیر جابجایی پذیر است و برای کاربردهای معنایی کاملا مناسب می باشد. تنها نکته باقی مانده که تا اینجا در مورد آن توضیح داده نشده است نحوه به دست آوردن α می باشد. این عامل نشان دهنده میزان اهمیت قطعه معنایی جاری در زمینه مورد بررسی می باشد. این عامل با استفاده از فرمول زیر محاسبه می شود:

$$\alpha = \alpha_{gct} + \alpha_{Gct} \quad (۱۹)$$

---

[5] Additive Product [6]

که در آن α_{gct} با استفاده از الگوریتمی که به نظریه مرکزیت عمومی شده[6] می گوییم به دست می آید. همچنین α_{Gct} نیز با استفاده از الگوریتمی که به آن نظریه مرکزیت سراسری شده[7] می گوییم به دست می آید. این دو الگوریتم در دو زیر بخش بعدی توضیح داده شده اند.

۴،۲،۲،۱. عمومی سازی نظریه مرکزیت

در این فصل مفاهیم جمله، دانش پیش زمینه و زمینه گسترش داده شدند. در این قسمت ارائه مجددی از نظریه مرکزیت با استفاده از این تعاریف جدید صورت می گیرد که به آن نظریه مرکزیت عمومی شده می گوییم. به همین منظور تمام قوانین جهت پشتیبانی از مفاهیم تعریف شده جدید تغییر شکل داده می شوند.

تعریف ۶ (مجموعه رو به جلوی[8] قطعه معنایی)

مجموعه رو به جلوی یک قطعه معنایی مانند $\mathcal{S}_i$ با توجه به دانش پیش زمینه داده شده با استفاده از $C_F(\mathcal{S}_i)$ نمایش داده می شود و عبارت است از:

$$C_F(\mathcal{S}_i) = \mathcal{S}_i \cap (\bigcup_{e_i \in \mathrm{E}} e_i) \qquad (۲۰)$$

که در آن E از مشخصات دانش پیش زمینه داده شده می باشد.

در یک قطعه معنایی تمام لغات در مجموعه رو به جلو قرار نمی گیرند و تنها کلماتی که دارای معنای مستقل می باشند باید در این مجموعه قرار گیرند که از این جمله می توان به اسامی و فعل ها اشاره نمود. تشخیص این لغات به این صورت است که تنها کلماتی که دارای روابط معنایی در دانش پیش زمینه هستند به عنوان مجموعه رو به جلو انتخاب می شوند که شامل همان اسامی و فعل ها می باشند. کلماتی از مجموعه الفبا که در یک دانش پیش زمینه دارای هیچ رابطه معنایی نباشند نمی توانند نمایندگی معنای مشخصی را بعهده بگیرند و از این رو جزء مجموعه منتخب نخواهند بود. از این طریق بخشی از پیش پردازش های لغوی در نظریه مرکزیت حذف شده و روش تبدیل به یک روش قطعی می گردد.

مجموعه بعدی مجموعه مسندی[9] قطعه معنایی می باشد که با استفاده از الگوریتم ۲ به دست می آید. این الگوریتم در زیر آمده است.

الگوریتم ۲ (ساخت مجموعه مسندی)

---

6 Generalized Centering Theory
7 Globalized Centering Theory
8 Forward Looking Set
9 Predicator Looking Set

شبه کد ۶ - الگوریتم محاسبه مجموعه مسندی

Input: Forward Looking Set $C_F(\mathcal{S}_i)$ and a partial order $\preccurlyeq_i$.

1. $PredicatorSet = C_F(\mathcal{S}_i)$
2. **while** (**True**) **do**
3. $SetIsUpdated =$ **False**
4. **for all elements of** $PredicatorSet$ **like** t **do**
5. **if** $range(\preccurlyeq_i |_t) = Nil$ **then**
6. $SetIsUpdated =$ **True**
7. $PredictorSet = PredicatorSet - \{t\}$
8. **if not** $SetIsUpdated$ **then**
9. **break**
10. $C_P(\mathcal{S}_i) = PredicatorSet$

Output: $C_P(\mathcal{S}_i)$

$C_P(\mathcal{S}_i)$ مجموعه مهم ترین اعضای یک قطعه معنایی داده شده می باشد. آخرین مجموعه باقیمانده مجموعه رو به عقب[۷۰] می باشد که با استفاده از $C_B(\mathcal{S}_i)$ نشان داده می شود. این مجموعه با استفاده از الگوریتم ۳ به دست می آید که در زیر آمده است.

الگوریتم ۳ (ساخت مجموعه رو به عقب)

شبه کد ۷ - محاسبه مجموعه رو به عقب

Input: Forward looking sets $C_F(\mathcal{S}_i)$ and $C_F(\mathcal{S}_{i-1})$

1. $C_B(\mathcal{S}_i) = C_F(\mathcal{S}_i) \cap C_F(\mathcal{S}_{i-1})$
2. $C_B(\mathcal{S}_i) = CreatePredicatorLookingSet(C_B(\mathcal{S}_i), \preccurlyeq_{i-1})$

Output: Backward looking set $C_B(\mathcal{S}_i)$

براساس این مجموعه های تعریف شده جدید باید قوانین مربوط به تعیین انتقال های معنایی نیز تغییر یابند. این قوانین جدید که در اصلی بازنویسی قوانین **Error! Reference source not found.** می با اشند در جدول زیر خلاصه شده اند.

[7] Backward Looking Set [0]

جدول3 - قوانین محاسبه انتقال های معنایی براساس روش ارائه شده

	$C_B(\mathcal{S}_i) \cap C_B(\mathcal{S}_{i-1}) \neq \emptyset$ or $C_B(\mathcal{S}_{i-1})\ is\ undef$	$C_B(\mathcal{S}_i) \cap C_B(\mathcal{S}_{i-1}) = \emptyset$
$C_B(\mathcal{S}_i) \cap C_P(\mathcal{S}_i) \neq \emptyset$	Continue	Smooth-Shift
$C_B(\mathcal{S}_i) \cap C_P(\mathcal{S}_i) = \emptyset$	Retain	Rough-Shift

با استفاده از **Error! Reference source not found.** انتقال معنایی بین دو قطعه معنایی و با توجه به ت عاریف جدید محاسبه می گردد. واضح است که این ارائه جدید از نظریه مرکزیت هم عمومی تر و هم رسمی تر از نسخه اصلی نظریه مرکزیت می باشد. این ویژگی های باعث می شوند تا بتوان پیاده سازی های شخصی سازی شده متناسب با کاربرد مورد نیاز هر فرد در اختیار داشت و در عین حال جنبه های کارایی نیز کاملا در پیاده سازی مد نظر قرار گیرند. برای به دست آوردن α_{gct} با توجه به انتقال معنایی شناسایی شده می توان از **Error! Reference source not found.** استفاده نمود.

جدول4 - مقادیر اهمیت هر یک از انتقال های معنایی

Transition	α_{gct}
Continue	0.5
Retain	0.32
Smooth-Shift	0.41
Rough-Shift	0.3

شهود و دلایل نحوه به دست آوردن این مقادیر در بخش نتایج عملی مورد بررسی قرار گرفته است. در این مرحله توانسته ایم ضریب مربوط به نظریه مرکزیت عمومی شده را به دست آوریم. در گام بعد باید ضریب مرتبط با نظریه مرکزیت سراسری شده محاسبه گردد و پس از آن می توان وزن دهی را انجام داد.

۴،۲،۲،۲. سراسری سازی نظریه مرکزیت

همان طور که پیش از این بیان شد یکی از مشکلات نظریه مرکزیت محلی بودن آن است. برای حل این مشکل در این فصل بردار زمینه ارائه شد که می تواند تا حدود زیادی این مشکل را مرتفع سازد. همان طور که گفته شد بردار زمینه مایانگر معنای متن از ابتدا تا نقطه مورد بررسی می باشد. بنابراین یک معیار محلی نمی باشد و در ضمن می تواند میزان پیوستگی یک قطعه معنایی به بقیه متن را نمایش دهد.

براساس توضیحات فوق مقدار α_{Gct} به روش زیر قابل محاسبه خواهد بود

- یک تابع مشابهت مانند $\odot$ را در نظر بگیرید که بازه برد آن بازه $[0,1]$ می باشد.

- رابطه مستقیمی بین این ضریب و مشابهت حاصل از تابع فوق بین بردار زمینه و بردار قطعه معنایی وجود دارد، به عبارت دیگر:

$$\alpha_{Gct} \simeq CV_i \odot \overrightarrow{S_{i+1}} \qquad (۲۱)$$

- حاصل فوق باید به بازه [0.4, 1.2] مقیاس شود.

نتیجه گام های فوق ضریب α_{Gct} خواهد بود. یک نمونه بسیار ساده از تابع مشابهت تابع مشابهت کسینوسی می باشد. البته توابع مشابهت بسیار زیاد دیگری نیز وجود دارند که می توان از آن ها استفاده نمود و یک نمونه از آن در فصل قبل ارائه گردید که تواسته است مشکلات توابع مشابهتی مانند تابع مشابهت کسینوسی را پوشش دهد.

۴,۲,۳. وزن دهی

آخرین گام وزن دهی به لغات می باشد. این گام بسیار مشابه به فرآیندی است که در نظریه مرکزیت سراسری شده ارائه گردید. گام های لازم برای وزن دهی به لغات به شرح زیر می باشند:

- یک تابع مشابهت مانند $\odot$ را در نظر بگیرید که بازه برد آن بازه $[0,1]$ می باشد.
- رابطه مستقیمی به صورت زیر وجود دارد:

$$\eta_i \simeq CV_i \odot \overrightarrow{S_i} \qquad (۲۲)$$

- نتیجه تابع مشابهت باید به بازه [0.7, 2.2] مقیاس شود.

نتیجه که با η_i نشان داده شده است در اصل وزن کلمات حاضر در قطعه معنایی i-اُم می باشد. در پایان باید بردار زمینه مورد استفاده ساخته شود. این بردار با استفاده از رابطه زیر محاسبه می شود.

$$DV = \sum_{S_i} \eta_i . \overrightarrow{s_i} \qquad (۲۳)$$

نتیجه این عبارت نشان دهنده وزن های محلی است که باید وزن سراسری و ضرایب نرمال سازی نیز برروی آن اعمال گردند. به عبارت دیگر به جای استفاده از فرکانس کلمات از وزن η_i در وزن محلی استفاده می شود. البته باید اصلاحاتی نیز برروی این وزن صورت گیرد که در بخش ۵,۲ مربوط به نتایج عملی مورد بررسی قرار می گیرد. در قطعات معنایی ابتدایی میزان شباهت به دست آمده برای قطعات بالا است، در نتیجه باید نتایج به نحوی اصلاح شوند تا تأثیر قطعات معنایی اولیه در جمله بسیار بالا نرود.

۴,۳. درستی یابی

با توجه به تغییراتی که در روش ارائه شده در این فصل نسبت روش های معمول وجود دارد، شایسته است که روش مورد درستی یابی قرار گیرد. به همین منظور از قالب کاری استاندارد ارائه شده در بخش ۰ به همین منظور استفاده می گردد. اگر روش ارائه شده بتواند ویژگی های مطرح شده در قالب کاری مذکور را برآورده سازد آنگاه می توان از درست بودن استدلال ها و روش ارائه شده اطمینان یافت. البته پیش از بررسی ذکر این نکته خالی از لطف نیست که در بسیاری موارد روش های ارائه شده بدون توجه به این ویژگی ها ارائه شده و دارای خواص اساسی مورد نیاز برای یک کاربرد متن کاوی نمی باشند.

همانطور که در بخش ۰ گفته خواهد شد، این قالب کاری استاندارد را می توان با استفاده از یک پنجگانه مانند $\langle A, \mathcal{A}, \xi, V, \psi \rangle$ نمایش داد. در اصل برای بررسی درستی روش ارائه شده باید بررسی گردد که جبرهای $\mathcal{A}$ و V در روش ارائه شده موجود باشند. در این فصل تنها یک روش ارائه جدید برای اجزاء معنایی ارائه گردید و از این رو به حیطه استنتاج پرداخته نشد. جبر V در اصل درگیر مسئله استنتاج معنایی می باشد. در این فصل از الگوریتم LSI برای انجام استنتاج براساس روش ارائه شده در این فصل استفاده شده است. در [CLA12] نمایش داده شده است که LSI دارای ویژگی های مورد نیاز برای V می باشد. از این رو می توان گفت یک گام از درستی یابی به بررسی روش استنتاج مربوط می شود که روش ارائه شده در این فصل شامل این حوزه نمی شود.

برای بررسی بخش اول نیز باید به تعاریف اشاره شده در بخش ۰ در خصوص جبر و جبر یکایی اشاره نماییم. واضح است که چنین عضوی برای جیر استفاده شده در روش ارائه شده موجود است که همان بردار تماما ۱ می باشد. یک ویژگی مورد نیاز دیگر برای $\mathcal{A}$ غیر جابجایی پذیر بودن عملگر تجمیع برروی این جبر است. براساس آنچه در بخش های قبل ارائه شد واضح است که رو ش تجمیع معنایی ارائه شده غیر جابجایی پذیر است. در نتیجه هر دو ویژگی مورد نیاز برای $\mathcal{A}$ برآورده شده اند. از این رو روش ارائه شده در این فصل براساس قالب کاری استاندارد ارائه شده در [CLA12] یک روش قابل قبول می باشد.

۴,۴.نتیجه گیری

در این فصل یک روش مبتنی بر اطلاعات زبان شناختی برای وزن دهی به لغات در متن ارائه گردید. این روش مسلما نسبت به روش های مرسوم در این حوزه دارای پیچیدگی محاسباتی بالاتری می باشد اما بعضی از مشکلات این روش ها را پوشش داده است. از مهم ترین مشکلات این روش ها که در اینجا به آن دقت شده است توجه به اطلاعات زبان شناختی می باشد.

علاوه براین سعی شد علیرغم استفاده از روش های زبان شناختی یک مدل محاسباتی مشخص برای روش مورد نظر ارائه گردد. این مدل محاسباتی کمک می نماید تا بتوان به صورت دقیق پیچیدگی زمانی اجرای این روش را محاسبه نمود. همچنین سعی شد تا توسعه مفهومی لازم در روش ارائه شده اعمال گردد تا بتوان از این روش برای حالت هایی که متون دارای شکل مرسوم خود نیستند نیز استفاده نمود.

همچنین در استفاده از نظریه مرکزیت به عنوان محور اصلی اطلاعات زبان شناختی مشکلات این روش نیز مد نظر قرار گیرد. یکی از مهم ترین مشکلات این روش توجه بیش از حد به اطلاعات محلی می باشد. از این رو سعی شد در دو حوزه درون متنی و فرامتنی جامعیت بیشتری به روش داده شود که تحت دو عنوان عمومی سازی و سراسری سازی ارائه شدند.

برای انجام مقایسه های دقیق تر به علت تفاوت های بنیادی با روش های موجود مسلما به نتایج عملی نیاز می باشد که در فصل پنجم این نتایج مورد بررسی قرار گرفته اند.

۵. نتایج تجربی

در این بخش به ارائه نتایج عملی پیاده سازی روش های ارائه شده در این پایان نامه می پردازیم. آنچه مسلم است با توجه به وجود فاصله بین راهکارهای ارائه شده به لحاظ روش حل مسئله لازم است تا در ابتدا هر یک از روش های ارائه شده مستقلا مورد بررسی قرار گیرند و سپس هر دو روش در یک آزمون مشترک با هم ترکیب و مورد بررسی قرار گیرند. این آزمون جامع باید بتواند ویژگی های مورد انتظار در هر یک از روش ها را پوشش داده و بیازماید. از این رو مسئله خلاصه سازی برای این موضوع مورد نظر قرار گرفته است. این مسئله به لحاظ میزان پیچیدگی و پوشش بر انواع پیچیدگی های حوزه بازیابی متن نمونه خوبی بوده و می تواند به درستی میزان کارایی روش های پیشنهادی را نمایش دهد. به عبارت دیگر چنانچه این دو روش بتوانند در کنار یکدیگر به خوبی برروی مسئله خلاصه سازی مورد استفاده قرار گیرند می توان انتظار داشت که هر یک به تنهایی نیز می توانند در ترکیب با دیگر روش ها نتایج خوبی را از خود نشان دهند.

به همین منظور در قسمت اول از این فصل به بررسی نتایج عملی پیاده سازی روش ارائه شده برای اصلاح معیارهای مشابهت خواهیم پرداخت. در قسمت دوم نیز نتایج عملی روش ارائه شده برای ورود معنا به وزن دهی را بررسی می نماییم. در هر یک از این دو بخش بررسی ها صرفا براساس اعمال هر یک از این روش ها برروی قالب کلی بازیابی اطلاعات بوده و بقیه اجزاء از الگوریتم های متداول پیروی می نمایند. در پایان در بخش سوم این دو روش به صورت ترکیبی استفاده شده و در کاربرد خلاصه سازی مورد استفاده قرار گرفته است.

۵,۱. اصلاح مشابهت

در این بخش نتایج عملی و مقایسه های حاصل از آن بین روش ارائه شده در خصوص اصلاح مشابهت در فصل ۳ با دیگر روش های محاسبه میزان مشابهت ارائه می گردد. اما پیش از پرداختن به مقایسه ها لازم است تا اندکی در خصوص محیط مقایسه صورت گرفته توضیح داده شود.

مجموعه داده های تست مورد استفاده عبارتند از مجموعه داده های MED, CISI, CRAN و CACM که به صورت استاندارد در اکثر مقالات در حوزه بازیابی اطلاعات مورد استفاده قرار می گیرند. همچنین در حوزه پاسخگویی به جستجو این مجموعه ها به صورت خاص دارای کاربرد بسیار زیادی هستند.

در آزمایش هایی که در این بخش صورت خواهند گرفت علاوه بر انجام مقایسه با روش های موجود به بررسی نقاط مبهم در روش ارائه شده پرداخته خواهد شد. به عبارت دیگر بعضی از جزئیات روش ارائه شده ممکن است علیرغم ارائه مبانی نظری مناسب در عمل دارای کارایی مناسب نباشند، از این رو باید به صورت مناسبی آزموده شوند تا کارایی آن ها نشان داده شود. مسائل اصلی مورد نظر شامل موارد زیر می باشند

- انتخاب یک تابع copula مناسب
- میزان کارایی نمایه گذاری اسناد
- تعیین میزان بهبود در معیار perplexity
- مطالعه تأثیر گسسته سازی مقادیر برروی نتایج حاصل

در گام اول به مقایسه روش LSI و جایگزینی معیار مشابهت مورد استفاده آن با ضرایب همبستگی مستقل از توزیع داده ها پرداخته خواهد شد. علت انتخاب LSI اولا اهمیت آن در بین روش های موجود و همچنین وابستگی آن به تابع توزیع داده ها می باشد که در فصل مربوط به روش اصلاح مشابهت به صورت کامل به توضیح آن پرداخته شده است. الگوریتم مورد استفاده در این مقایسه همان الگوریتم عمومی ارائه شده در بخش ۳,۱ می باشد. اما با این تفاوت که به جای استفاده از توابع copula در این بخش از ضرایب همبستگی Spearman و Kendall استفاده خواهد شد. همچنین هیچ گسسته سازی در مقادیر نیز مورد استفاده قرار نخواهد گرفت. نتایجی که در این بخش ارائه می شوند نیز نشان دهنده میانگین دقت در مجموعه داده های آزمون مختلف می باشد. نتایج این مقایسه در **Error! Reference source not found.** نمایش داده شده است.

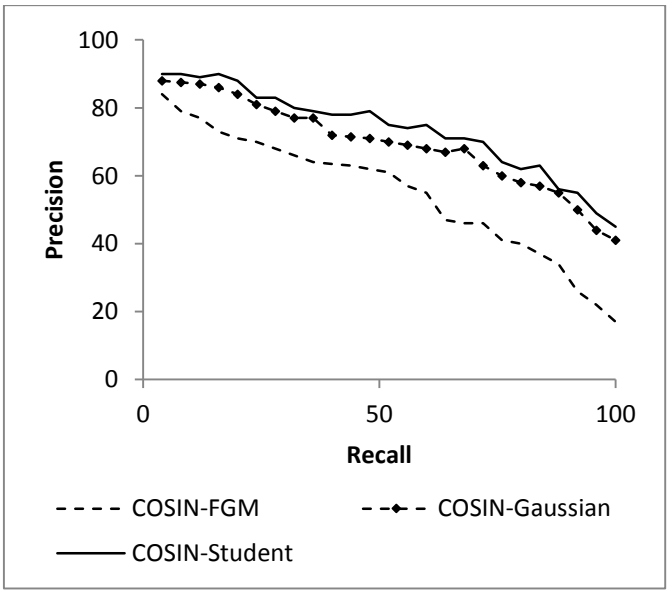
جدول5 - نتایج مقایسه بین LSI و دیگر معیارهای همبستگی مستقل از توزیع داده ها

	MED	CISI	CRAN	CACM
LSI	50.4	12.5	29.2	17.2

Kendall	51.1	12.3	29.4	17.8
Spearman	51.8	12.7	30.3	17.5

همانطور که مشاهده می شود در مقایسه با LSI میران بهبود نتایج قابل توجه نیست. همانطور که در فصل ۳ نیز به آن اشاره شد به علت گسسته بودن مقادیر مورد استفاده و همچنین ساده بودن این معیارهای همبستگی این عدم بهبود در نتایج قابل پیش بینی بوده اند. اما همین عدم افت در دقت نتایج نشان گر این نکته است که این معیارها دارای پتانسیل مناسبی جهت استفاده در کاربردهایی از این دست هستند و همین مطلب راه را برای استفاده از معیارهای همبستگی با قدرت بیان بالاتر را هموار می سازد.

با توجه به نتایج گام قبل و اطمینان از سودمند بودن استفاده از معیارهای همبستگی مستقل از توزیع، گام بعد می تواند انجام آزمون های لازم جهت انتخاب تابع Copula مناسب باشد. توابعی که در این آزمون مورد نظر قرار گرفته اند توابع FGM، Gaussian و Student's t هستند. این سه تابع Copula به ترتیب دارای ویژگی های سادگی، عمومیت و توان بیان بالا می باشند. نتایج این مقایسه در شکل ۵ قابل مشاهده می باشد. براساس این نتایج بهترین تابع در بین این سه تابع Students't می باشد. به همین دلیل همین تابع به عنوان تابع اصلی در مقایسه روش ارائه شده با روش های موجود ملاک عمل قرار خواهد گرفت.

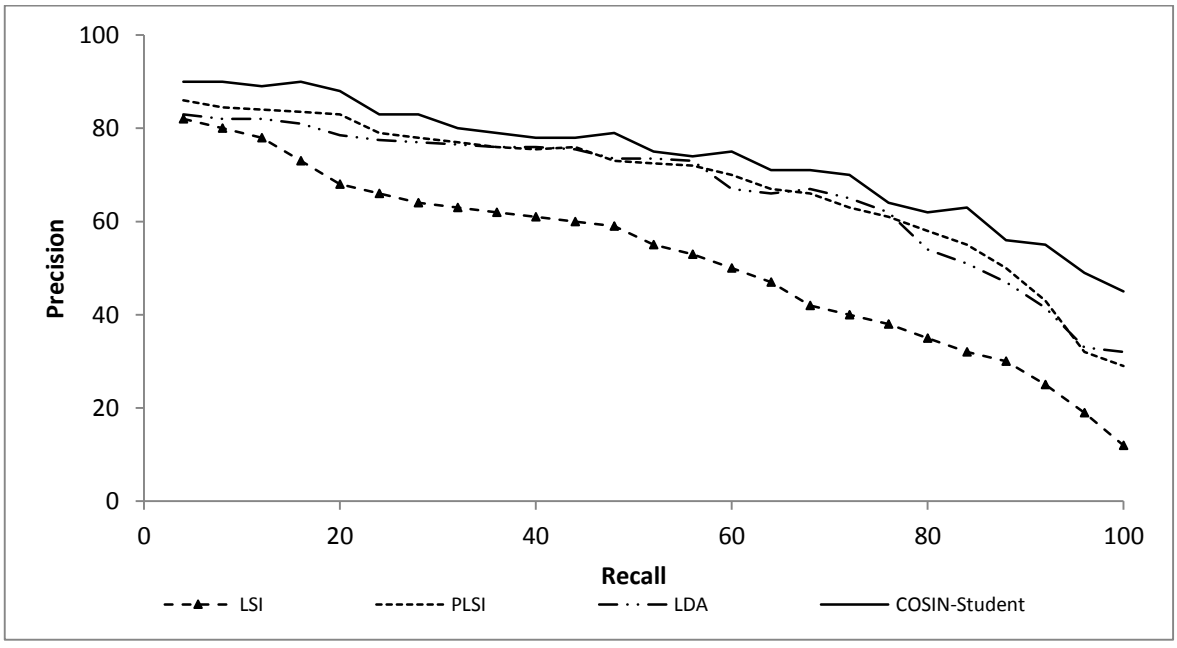


شکل ۵ - نتایج مقایسه منحنی میانگین Precision/Recall برای سه تابع مورد هدف

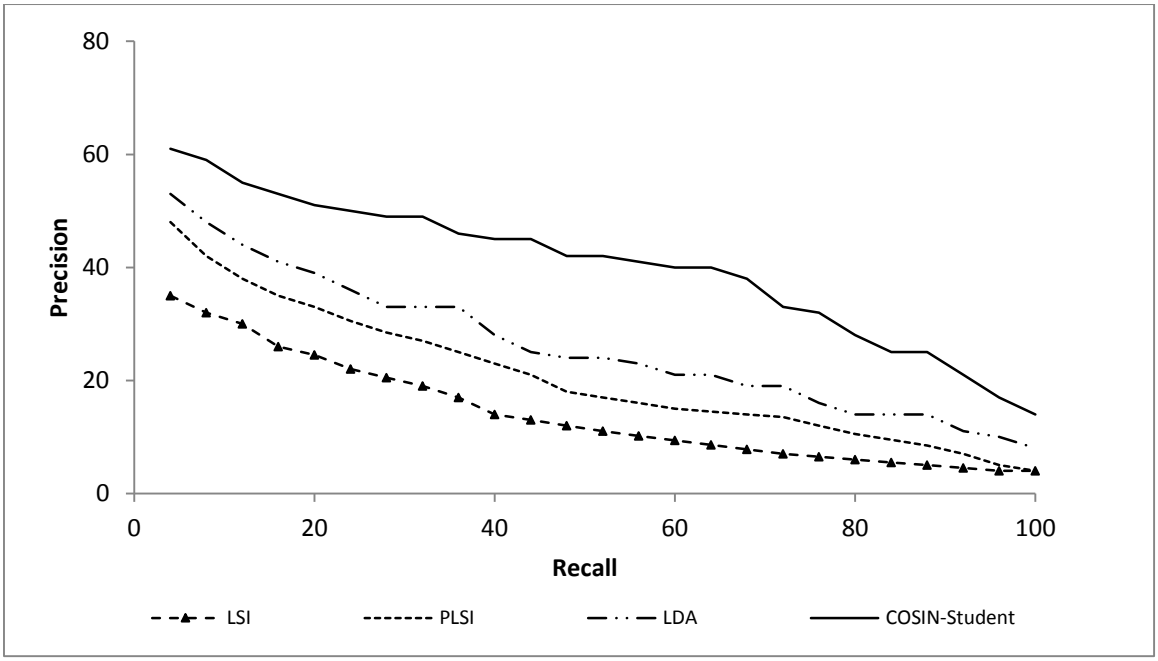
مطابق با نتایج فوق تابع Students't بهترین نتایج را بدست آورده است. همانطور که در فصل ۳ نیز به صورت کامل بررسی شد، علت این تفاوت ها ناشی از تفاوت در ویژگی های ذاتی هر یک از توابع مورد استفاده می باشد. تفاوت بین دو تابع Gaussianو Students't نیز ناشی از قدرت این تابع در پوشش وابستگی های انتهایی بین توابع توزیع می باشد. در این بین نیز FGM به علت سادگی بسیار زیاد دارای بدترین نتایج بوده است.

پس از دو گام ابتدایی می توانیم این نتایج را بگیریم. اولا استفاده از معیارهای همبستگی مستقل از توزیع می توانند به عنوان راه حلی قابل اعتماد مورد بررسی قرار گیرند. ثانیا در مسئله پاسخگویی به پرس و جو، از بین روش های مختلف محاسبه همبستگی مستقل از توزیع استفاده از Students't Copula دارای بیشترین شانس در بهبود نتایج می باشد. حال برای گام سوم که مقایسه COSIN با دیگر روش های متداول تحلیل مدل معنایی می باشد آمادگی لازم ایجاد شده است. مقایسه مبتنی بر منحنی Precision/Recall در مجموعه داده های MED و CISI می باشد. به صورت معمول مجموعه داده MED بهترین نتایج منحنی Precision/Recall و مجموعه داده CISI بدترین نتایج را نمایش می دهد. این نتایج در شکل ۷ قابل مشاهده می باشند.

براساس نتایج ارائه شده در شکل ۷ روش COSIN توانسته نتایج بهتری را نسبت به دیگر روش ها ارائه نماید. همچنین در recall در حدود ۱۰۰ درصد میزان دقت COSIN در حدود ۱,۵ تا ۲ برابر LDA می باشد. این شرایط دقیقا شرایطی است که در مسائل دنیای واقعی با آن مواجه هستیم. به عبارت دیگر در یک مسئله واقعی مانند یک موتور جستجو هدف این است که کاربر بتواند در کمترین تعداد نتایج تمامی جواب های مرتبط با جستجوی خود را بیابد (شرایطی که recall باید برابر ۱۰۰ درصد باشد). این دقیقا همان شرایط توصیف شده در بالا است. پس می توان این طور نتیجه گرفت که هرچند COSIN در Precision برابر با ۱۰۰ درصد تفاوت مهمی با دیگر روش ها ندارد اما در شرایط مسائل واقعی دارای دقتی دو برابر دیگر روش ها می باشد. بنابراین COSIN می تواند در کاربردهای عملی قابلیت های بسیار خوبی از خود نشان دهد.



شکل ۶ - نتیجه مقایسه COSIN با روش های موجود ایجاد مدل معنایی نهان برروی مجموعه داده MED



شکل ۷ – نتیجه مقایسه COSIN با روش های موجود ایجاد مدل معنایی نهان برروی مجموعه داده CISI

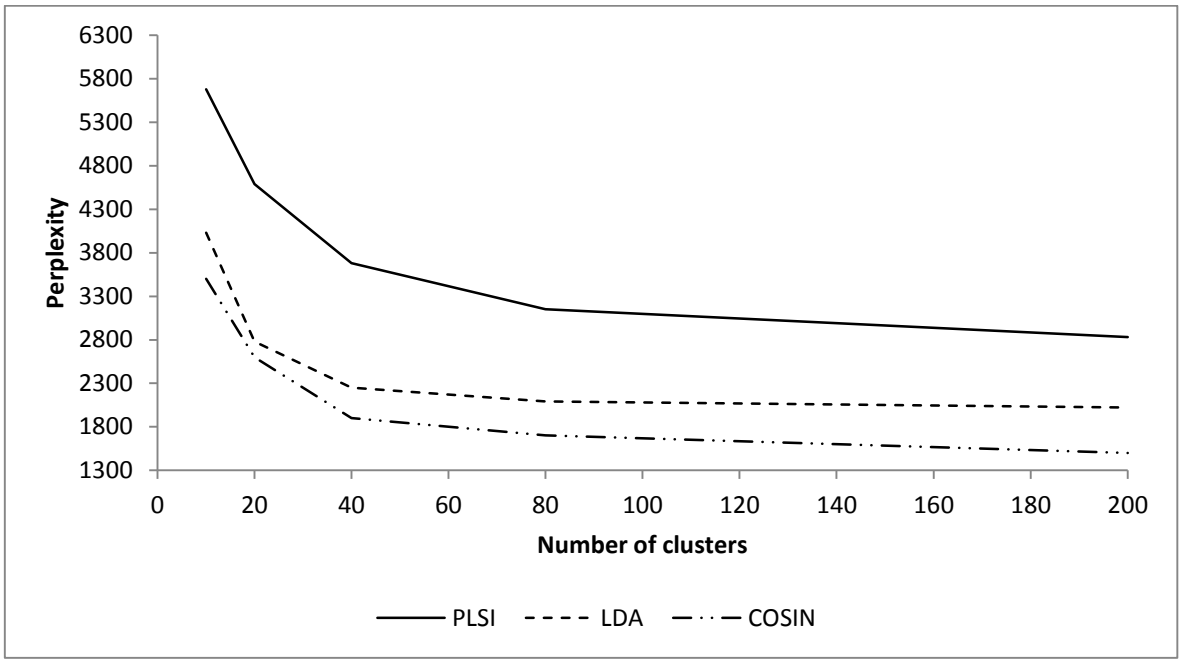
تا این مرحله در شرایط آزمون و برروی داده های آزمون برتری روش ارائه شده نمایش داده شد. اما این برتری نتایج باید قابل توسعه به شرایط واقعی نیز باشد. به همین دلیل در گام بعد مقایسه برروی مقادیر معیار Perplexity صورت گرفته است. Perplexity معیاری برای نمایش کارایی تعمیم یک روش در شرایط مختلف می باشد. هرچه مقدار این معیار برای یک روش پایین تر باشد آن روش می تواند نتایج بهتری را در شرایط مختلف به صورت میانگین از خود نشان دهد. به عبارت دیگر هر چه مقدار این معیار پایین تر باشد رفتار روش مورد بررسی دارای قابلیت پیش بینی بیشتری خواهد بود. برای محاسبه این معیار باید دارای دو مجموعه داده یادگیری و آزمون باشیم. به عبارت دیگر باید نتایج اولیه برروی یک مجموعه داده یادگیری محاسبه شوند و سپس تفاوت این نتایج با نتایج حاصل از مجموعه داده آرمون محاسبه شود تا میزان قابلیت پیش بینی رفتار روش مورد نظر محاسبه شود. به بیان آماری می توان گفت Perplexity عبارت است از میزان شباهت مدل محاسبه شده برای مجموعه داده های یادگیری و آزمون می باشد و یا به عبارت دقیق تر برابر است با عکس میانگین هندسی احتمالات محاسبه شده برای متغیرهای تصادفی مجهول موجود در مسئله می باشد. در زیر این بیان ریاضی به صورت دقیق نمایش داده شده است.

$$\text{perplexity} = 2^{-(\sum_{i=1}^{N} p(D_i))/(\sum_{i=1}^{N} \text{size}_i)} \qquad (۲۴)$$

که در آن $p(D_i)$ عبارت است از میزان احتمال تخصیص داده شده به هر سند و به صورت زیر محاسبه می شود.

$$p(D_i) = \prod_{j=1}^{\text{size } D_i} \sum_{k=1}^{K} p(w_j|z_k)p(z_k) \qquad (۲۵)$$

برهمین اساس مقایسه ای برروی این معیار در بین روش های مختلف صورت گرفته است که نتایج آن را می توان در شکل ۸ مشاهده نمود. براساس این نتایج COSIN توانسته است توان تعمیم بالاتری را در اختیار قرار دهد.

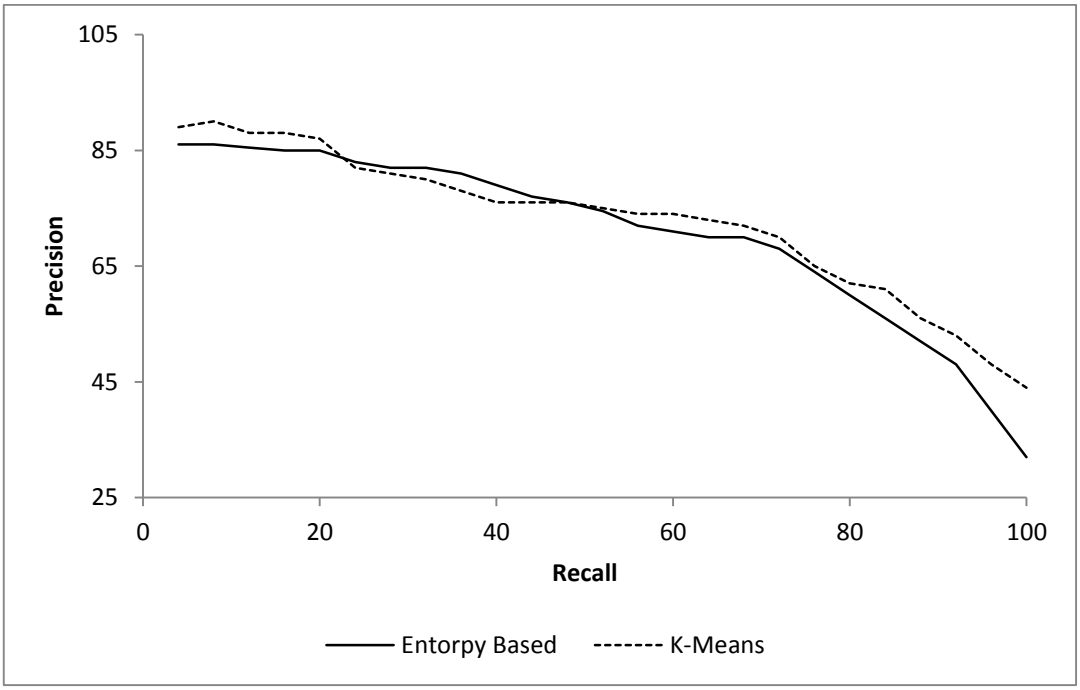


شکل ۸ – مقایسه معیار Perplexity برای روش های مختلف و با تعداد مفاهیم مختلف

براساس نتایجی که تا این مرحله ارائه شده است توانمندی COSIN در حل مسئله پاسخگویی به پرس و جوها و همچنین تعمیم به شرایط مختلف نمایش داده شده است. اما در الگوریتم ارائه شده بعضی مسائل مرتبط با پیاده سازی نیز وجود دارند که باید از طریق آزمون میزان تأثیر آن ها و صحت فرض های ارائه شده بررسی شوند. یکی از این مسائل انتخاب تعداد مفاهیم می باشد. براساس نتایجی که تا کنون ارائه شد و همچنین روش ارائه شده در [GAO11] تعداد مفاهیم مورد استفاده برای COSIN در مجموعه داده های آزمون استفاده شده در اینجا برابر مقدار ۲۰۰ می باشد. روش ارائه شده در [GAO11] تعداد مفاهیم را براساس یک تغییر چشمگیر در مقادیر ویژه ماتریس همبستگی انتخاب می نماید (با فرض مرتب بودن مقادیر ویژه به صورت نزولی). به عبارت دیگر هر گاه یک فاصله چشمگیر بین مقادیر ویژه ماتریس همبستگی مشاهده شود، می توان این طور برداشت نمود که به علت فاصله مقداری زیاد به وجود آمده مقادیر و بردارهای ویژه بعد از این تغییر دارای اهمیت بسیار کمتری نسبت به مقادیر و بردارهای ویژه قبل از این تغییر هستند و از این رو می توان این تغییر را مرز مناسبی برای تعیین تعداد مفاهیم در نظر گرفت.

یکی دیگر از مسائل مورد بررسی در حوزه پیاده سازی، پیشنهاد استفاده از یک گام گسسته سازی در محاسبه مقادیر Copula جهت کاهش پیچیدگی زمان اجرا و بهبود کارایی روش می باشد. برای مطالعه این

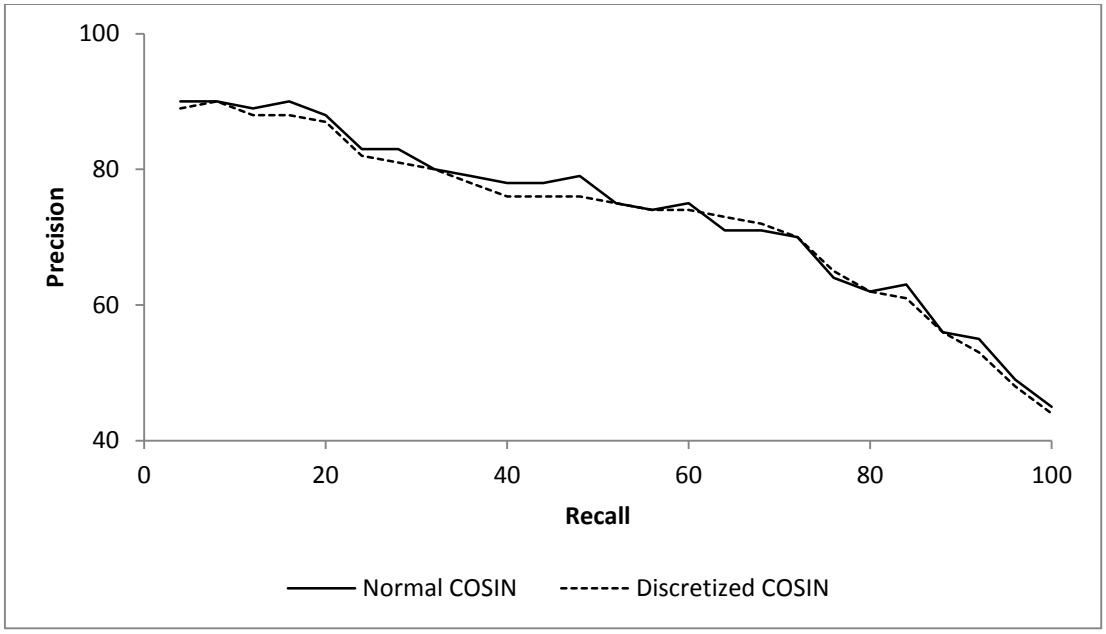
تأثیرات باید دو گام برداشته شود. گام اول انتخاب روش مناسبی برای گسسته سازی و سپس گام دوم مطالعه تأثیر این گسسته سازی می باشد. به همین منظور دو روش گسسته سازی یک روش ارائه شده در [FAY93] که مبتنی بر آنتروپی می باشد و دیگر k-means انتخاب شده و میزان کارایی هر یک مقایسه شده است. این گسسته سازی ها برروی مجموعه داده MED بررسی شده و نتایج آن براساس منحنی Precision/Recall تهیه و در شکل ۹ نمایش داده شده است.



شکل ۹ – مقایسه دو روش گسسته سازی برای تهیه مقادیر از پیش محاسبه شده Copula

این نتایج عملکرد بهتری را برای k-means نمایش می دهند. از این رو k-means در بقیه آزمایش ها به عنوان روش گسسته سازی انتخاب شده است. حال نوبت آن رسیده که نتایج حاصل از این گسسته سازی با شرایط عادی مقایسه گردد. این نتایج نیز مبتنی بر منحنی Precision/Recall ارائه شده اند. نتایج ارائه شده در شکل ۱۰ نشان می دهند که این گسسته سازی تأثیر بسیار مختصری در کارایی روش داشته است و از این رو می توان از این روش برای بهبود پیچیدگی زمانی محاسبات به خوبی استفاده نمود.

یکی دیگر از مسائل مرتبط با گسسته سازی، اندازه این دانش پیش زمینه تولید شده می باشد. در **Error! Reference source not found.** مقایسه ای بین تعداد محاسبات Copula برای دو حالت معمولی و حالت استفاده از مقادیر از پیش محاسبه شده صورت گرفته است. نکته ای که وجود دارد این است که در حالت مقادیر از پیش محاسبه شده این تعداد محاسبه تنها در زمان تهیه مجموعه دانش اولیه انجام می شوند و در هر بار اجرای الگوریتم محاسبه نمی شوند. در حالیکه در حالت معمول به تعداد اعلام شده در **Error! Reference source not found.** محاسبات باید صورت گیرد.



شکل ۱۰ – مقایسه COSIN در دو حالت استفاده از مقادیر از پیش محاسبه شده COSIN و حالت معمول

جدول6 – مقایسه تعداد محاسبات در دو حالت معمولی و تهیه مقادیر Copula از پیش تعریف شده

	MED	CISI	CRAN	CACM
Normal Mode	8.6×10^9	3.5×10^9	2.6×10^9	2.7×10^9
Discretized Mode	8×10^6	8×10^6	8×10^6	8×10^6

در یک نتیجه گیری اولیه از مجموعه گام های برداشته شده می توان به این جمع بندی رسید که استفاده از روش های محاسبه همبستگی مستقل از توزیع نه تنها تأثیر منفی بر نتایج ندارند بلکه بهبود بسیار مناسبی ایجاد می نمایند و همخوانی بیشتری با شرایط مسائل واقعی دارند. از سوی دیگر انتخاب تابع مناسب برای محاسبه همبستگی دارای اهمیت بسیاری می باشد و علاوه بر لزوم همخوانی با شرایط مسئله باید دارای خاصیت مهم پوشش وابستگی های انتهایی نیز باشند. استفاده از ابزارهای قدرتمند مسلما نیاز به توان پردازشی بالاتری دارد اما می توان با کمی دقت و استفاده از روش های گسسته سازی و تهیه مقادیر از پیش محاسبه شده نیاز پردازشی را به حداقل ممکن رساند.

۵,۲.ورود معنا در وزن دهی

در این بخش نتایج عملی و مقایسه های حاصل از آن بین روش ارائه شده در خصوص ورود معنا در وزن دهی در فصل ۴ با دیگر روش های محاسبه وزن ارائه می گردد. اما پیش از پرداختن به مقایسه ها لازم است تا اندکی در خصوص محیط مقایسه صورت گرفته توضیح داده شود. نکته قابل توجه این است که در اینجا به

ویژگی های دیگری از شرایط محیط آزمون که در قسمت قبل کمتر دارای اهمیت بودند نیز پرداخته شده است.

برای آزمون روش ارائه شده و همچنین حفظ استقلال نتایج یک روش معمول بازیابی اطلاعات باید انتخاب گردد. روش انتخاب شده در اینجا LSI می باشد. به عبارت دیگر با استفاده از روش ارائه شده وزن لغات در اسناد محاسبه شده و سپس با استفاده از LSI خوشه های لغات محاسبه شده و پاسخگویی به پرس و جوها صورت می گیرد.

یکی دیگر از نکات مورد توجه انتخاب مجموعه دانش پس زمینه مورد استفاده در روش ارائه شده می باشد. مجموعه دانش مورد استفاده در اینجا WordNet می باشد. با توجه به سهولت دستیابی، جامعیت و همچنین سهولت استفاده از این ابزار می توان گفت که بهترین گزینه حداقل در حوزه زبان انگلیسی برای دانش پس زمینه همین مجموعه دانش می باشد.

همچنین مجموعه داده های آزمون مورد استفاده نیز مجموعه داده های MED, CISI و CRAN می باشند که در **Error! Reference source not found.** مشخصات هر یک بیان شده است. علت بیان ا ین مشخصات در این بخش نیز لزوم توجه به ابعاد و کیفیت داده های آزمون در این روش می باشد.

جدول7 - مشخصات ابعادی هر یک مجموعه های داده آزمون

نام مجموعه داده	تعداد بردارها	متوسط طول هر بردار	انحراف معیار طول بردارها	میانگین فراکنس لغات در بردارها	درصد تعداد لغات با فرکانس یک در بردارها
MED					
اسناد	1033	51.6	22.78	1.54	72.7
پرس و جوها	30	10.1	6.03	1.12	90.76
CISI					
اسناد	1460	46.55	19.38	1.37	80.27
پرس و جوها	112	28.29	19.49	1.38	78.36
CRAN					
اسناد	1398	53.13	22.53	1.58	69.5
پرس و جوها	225	9.17	3.19	1.04	95.69

البته در مواردی که هدف صرفا بررسی های زبانی بوده و بازیابی اطلاعات هدف نبوده است مجموعه DUC2001 از این جهت یک مجموعه داده با اهداف متفاوت بوده و بیشتر از ویژگی های بازیابی اطلاعات ویژگی های زبانی آن مورد نظر بوده اند مورد استفاده قرار گرفته است.

یکی دیگر از ویژگی های محیط آزمون که باید مورد توجه قرار گیرد روش های مشابهی است که در مقایسه مورد استفاده قرار می گیرند. علت مرور جزئیات این روش ها نیز نمایش این نکته است که محاسبه بعضی از این روش ها دارای پیچیدگی های بالایی می باشد. در روش های مختلف و فرمول محاسبه هر یک نمایش داده شده است.

در [LAN06] نشان داده شده است که تفاوت آشکاری بین روش های ذکر شده در این جدول و دیگر روش های وزن دهی لغات در اسناد وجود ندارد. از این رو همین روش ها می توانند نماینده مناسبی برای انجام مقایسه های مورد نیاز باشند.

جدول8 - روش های محاسبه وزن متداول که هر یک نماینده یک دسته خاص هستند

Method	Local Weight	Global Weight
TF-IDf	$f_{i,j}$	Log $\frac{N}{n_i}$
LOGTF-IDF	1+log $f_{i,j}$ if $f_{i,j}$>0 0 if $f_{i,j}$ = 0	Log $\frac{N}{n_i}$
TF-ENPY	$f_{i,j}$	$1+\sum_{j=1}^{N}\frac{\frac{f_{ij}}{gf_i}\log\frac{f_{ij}}{gf_i}}{\log N}$
ATF1-IDFP	0.5 + 0.5 (f_{ij} / x_j) if $f_{ij} > 0$ 0 if $f_{ij} = 0$	log [(N – ni) / ni]
GF-IDF	$\frac{gf_i}{n_i}$	Log $\frac{N}{n_i}$
TF	$f_{i,j}$	None

در این جدول $f_{i,j}$ فرکانس نسبی رخداد لغت t_i در سند d_j است. N برابر تعداد اسناد موجود در مجموعه داده آزمون است. n_i تعداد اسنادی است که حاوی لغت t_i می باشند. gf_i کل فرکانس لغت t_i در مجموعه داده آزمون می باشد. x_j نیز حداکثر فرکانس در بین لغات سند d_j می باشد.

یکی از مسائلی که درخصوص LSI نیز باید مورد توجه قرار گیرد و در بخش قبل نیز به آن اشاره شد تعداد خوشه ها و یا مفاهیم می باشد. در اینجا نیز از همان معیار ارائه شده در بخش قبل استفاده خواهد شد.

گام بعدی در آماده سازی محیط آزمون تخمین پارامترهای اجرا است. پس از تخمین این پارامترها می توان مقایسه با روش های دیگر را انجام داد.

- α_{gct}: این پارامتر در عمومی سازی نظریه مرکزیت مورد استفاده قرار گرفت. این پارامتر در اصل یک چهارگانه است که تعیین کننده وزن هر یک انتقال های معنایی می باشد.
- α_{Gct}: یکی دیگر از بخش ها ترکیب حاصل عمومی سازی نظریه مرکزیت و سراسری سازی نظریه مرکزیت بود. در این قسمت دو حاصل با استفاده از یک میانگین وزنی با هم ترکیب می شوند. وزن مورد نظر از تابع شباهت حاصل خواهد شد. اما ممکن است نقش هر دو حاصل در نتیجه نهایی برابر نباشد. از این رو برای مقادیر حاصل از عمومی سازی نظریه مرکزیت نیاز به یک تغییر مقیاس داریم. این کار را از طریق دو مقدار α_{Gct_min} و α_{Gct_max} انجام خواهیم داد.
- η_i: یکی دیگر از پارامترهای مورد نیاز است. در اصل این پارامتر نیز مانند پارامتر قبل احتیاج به یک نگاشت دارد. اما براساس روابط ارائه شده در فصل ۴ یک نگاشت خطی برروی این پارامتر نمی تواند تأثیری بر نتایج داشته باشد. از سوی دیگر یک نگاشت غیرخطی پیچیدگی های تخمین پارامترهای ورودی بسیار افزایش می دهد و در اینجا از آن صرف نظر می گردد. همان مقادیر اولیه به دست آمده از معیارهای شباهت برای ترکیب میزان اهمیت قطعات معنایی استفاده خواهد شد و نگاشتی برروی این مقدار صورت نخواهد گرفت.

براساس توضیحات فوق تعداد مفاهیم و یا همان خوشه ها در زمان اجرا و براساس مقادیر ویژه ماتریس همبستگی محاسبه می شوند. پارامترهای باقیمانده برای تخمین نیز α_{gct}، α_{Gct_min} و α_{Gct_max} می باشند. که البته α_{gct} یک چهارگانه می باشد. تخمین مقادیر شش مقدار مختلف تعداد مقایسه ها و محاسبات را بسیار افزایش خواهد داد. از این رو از یک تابع حدس برای محاسبه α_{gct} استفاده می نماییم. این تابع حدس عبارت است از تعیین میزان اهمیت هر انتقال معنایی براساس یک ارتباط معکوس با میزان تکرار انتقال مورد نظر در مجموعه داده های آزمون. این میزان تکرار برای مجوعه داده آزمون DUC2001 محاسبه شدند که نتایج آن برای نسخه اصلی نظریه مرکزیت در **Error! Reference source not found.** و برای نسخه عمومی سازی نظریه مرکزیت در **Error! Reference source not found.** نمایش داده شده است.

جدول9 - متوسط تعداد رخداد هر انتقال معنایی در نسخه اصلی نظریه مرکزیت

انتقال معنایی	فرکانس نسبی رخداد
Continue	0.15
Retain	0.24
Smooth-shift	0.19

Rough-shift	0.42

جدول10 - متوسط تعداد رخداد هر انتقال معنایی در نسخه عمومی سازی نظریه مرکزیت

انتقال معنایی	فرکانس نسبی رخداد
Continue	0.18
Retain	0.28
Smooth-shift	0.22
Rough-shift	0.32

همانطور که مشاهده می شود و مورد انتظار نیز بوده است در نسخه عمومی سازی شده نظریه مرکزیت تعداد جفت قطعات معنایی بدون انتقال نسبت به دیگر حالت ها کاهش داشته و بقیه حالت ها دارای افزایش در مقدار بوده اند. از سوی دیگر ترتیب مقادیر در هر دو نسخه از نظریه مرکزیت با هم یکسان است و همچنین ترتیب میزان اهمیت انتقال های معنایی در نسخه اصلی نظریه مرکزیت که در فصل ۴ بیان شد دارای نسبت عکس با میزان تکرار این انتقال ها می باشد. از این رو می توان انتظار داشت که تابع حدس در خصوص تعیین میزان اهمیت انتقال های معنایی بتواند به خوبی عمل نماید. بر این اساس مقادیر اهمیت انتقال های معنایی در ارائه گردیده است.

جدول11 - مقادیر وزن محاسبه شده برای هر انتقال براساس میزان تکرار هر یک در عمومی سازی نظریه مرکزیت

انتقال	وزن انتقال
Continue	1.39
Retain	0.89
Smooth-shift	1.14
Rough-shift	0.78

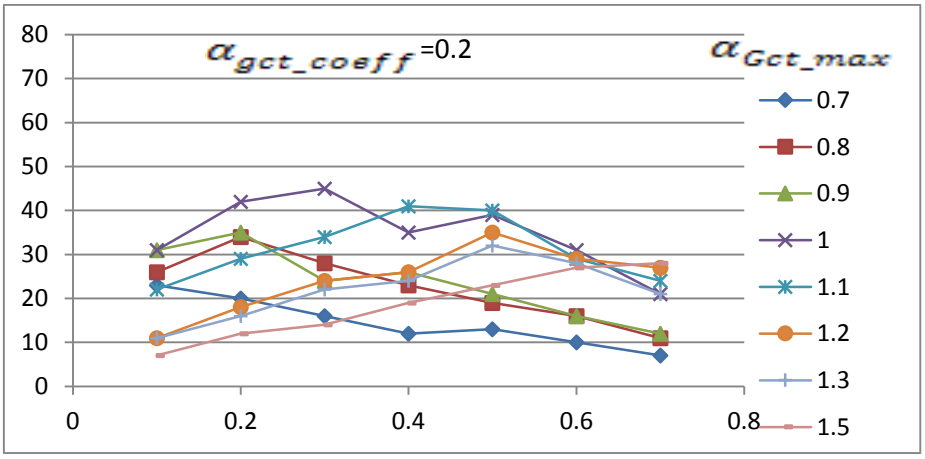
مقادیر این جدول به این صورت محاسبه شدند که حاصلضرب هر یک از مقادیر وزن در فرکانس متناظر برابر با ۰,۲۵ باشد. حال به نظر می رسد برای ترکیب مناسب بین این وزن ها و پارامتر α_{Gct} لازم است یک نگاشت برای مقادیر آن نیز تعریف شود. این پارامتر نگاشت با α_{gct_coeff} نمایش داده می شود. از این رو می توان گفت پارامترهایی که باید تخمین زده شوند عبارتند از α_{gct_coeff}، α_{Gct_min} و α_{Gct_max}. برای تخمین این پارامترها آزمون های متعددی صورت گرفته که نتایج آن در قالب ۸ نمودار در شکل ۱۸ نشان داده شده است.

این آزمون ها به خوبی نشان دهنده مقادیر مناسب برای پارامترهای ورودی هستند. مقادیر انتخاب شده برای α_{gct} در **Error! Reference source not found.** نمایش داده شده اند.

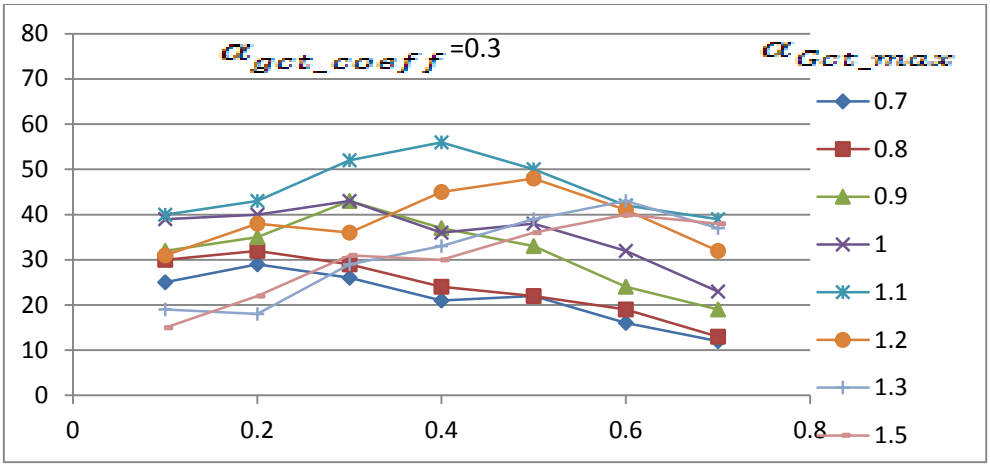
جدول12 - مقادیر مختلف برای وزن هر یک از انتقال های معنایی

انتقال	وزن هر انتقال
Continue	0.6
Retain	0.38
Smooth-shift	0.49
Rough-shift	0.34

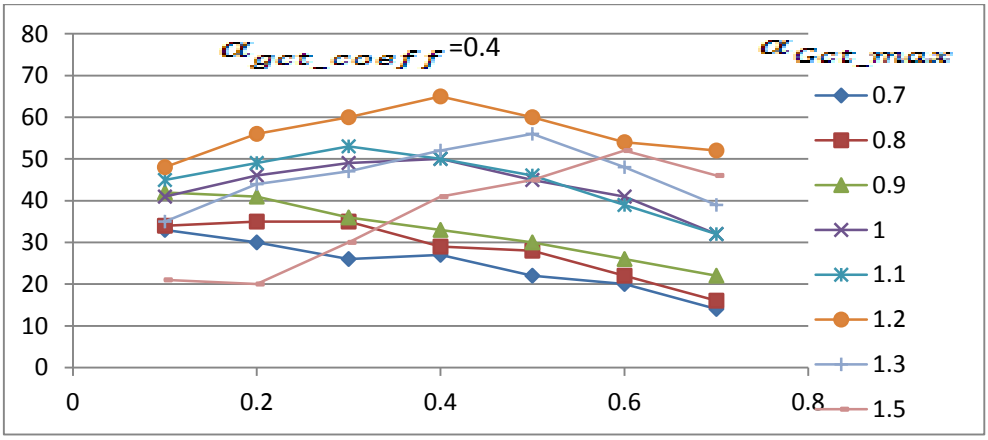
همچنین مقادیر به دست آمده برای α_{gct_coeff}، α_{Gct_min} و α_{Gct_max} نیز به ترتیب ۰,۴۵، ۰,۴ و ۱,۲ می باشند. البته دقت آزمون ها بیش از تعداد نقاط رسم شده در نمودارها بوده و در اینجا جهت سهولت در مشاهده نتایج تعداد کمتری از نتایج در نمودارها نمایش داده شده اند.



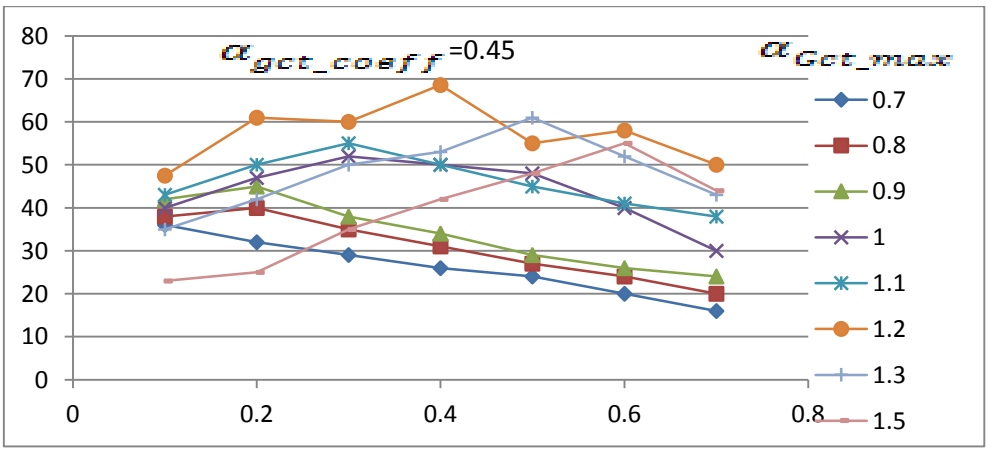
شکل ۱۱ - مقایسه مقادیر مختلف برای پارامترهای ورودی برای α_{gct_coeff} برابر با ۰,۲. محور عمودی نشان دهنده میزان دقت و محور افقی نشان دهنده پارامتر α_{Gct_min} است که بازه ۰,۱ تا ۰,۷ برای آن در نظر گرفته شده است. البته مقادیر بیش از ۰,۷ دارای دقت پایین تری بوده اند که نمایش داده نشده اند.



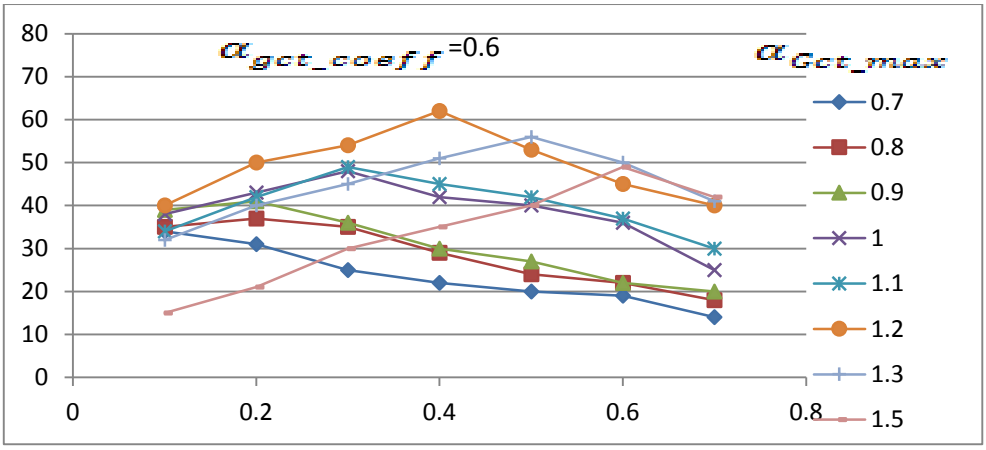
شکل ۱۲ – مقایسه مقادیر مختلف برای پارامترهای ورودی برای α_{gct_coeff} برابر با ۰٫۳. محور عمودی نشان دهنده میزان دقت و محور افقی نشان دهنده پارامتر α_{Gct_min} است که بازه ۰٫۱ تا ۰٫۷ برای آن در نظر گرفته شده است. البته مقادیر بیش از ۰٫۷ دارای دقت پایین تری بوده اند که نمایش داده نشده اند.



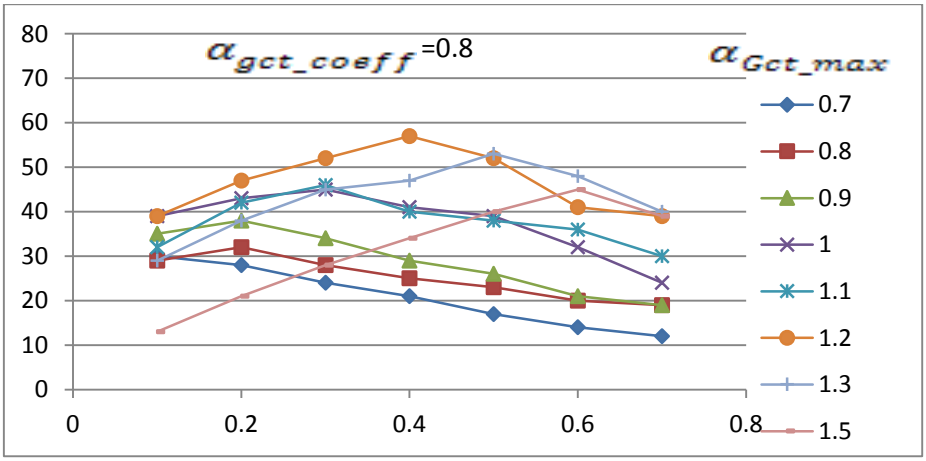
شکل ۱۳ – مقایسه مقادیر مختلف برای پارامترهای ورودی برای α_{gct_coeff} برابر با ۰٫۴. محور عمودی نشان دهنده میزان دقت و محور افقی نشان دهنده پارامتر α_{Gct_min} است که بازه ۰٫۱ تا ۰٫۷ برای آن در نظر گرفته شده است. البته مقادیر بیش از ۰٫۷ دارای دقت پایین تری بوده اند که نمایش داده نشده اند.



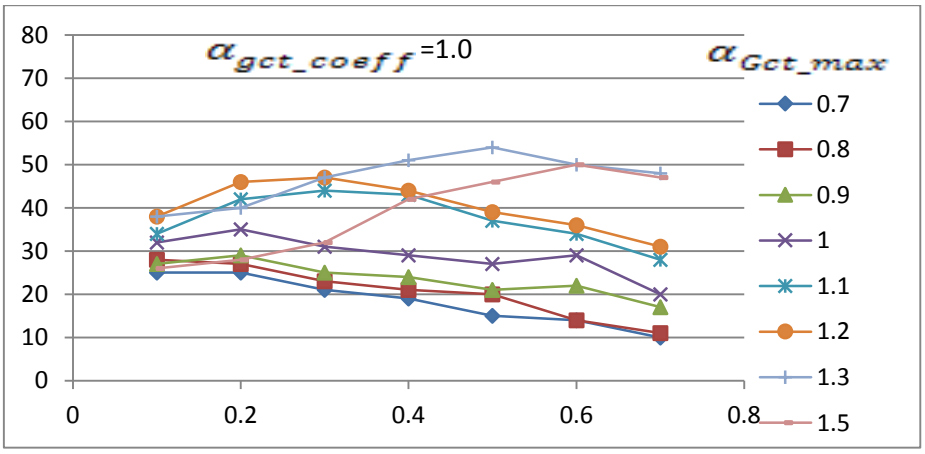
شکل ۱۴ – مقایسه مقادیر مختلف برای پارامترهای ورودی برای α_{gct_coeff} برابر با ۰٫۴۵. محور عمودی نشان دهنده میزان دقت و محور افقی نشان دهنده پارامتر α_{Gct_min} است که بازه ۰٫۱ تا ۰٫۷ برای آن در نظر گرفته شده است. البته مقادیر بیش از ۰٫۷ دارای دقت پایین تری بوده اند که نمایش داده نشده اند.



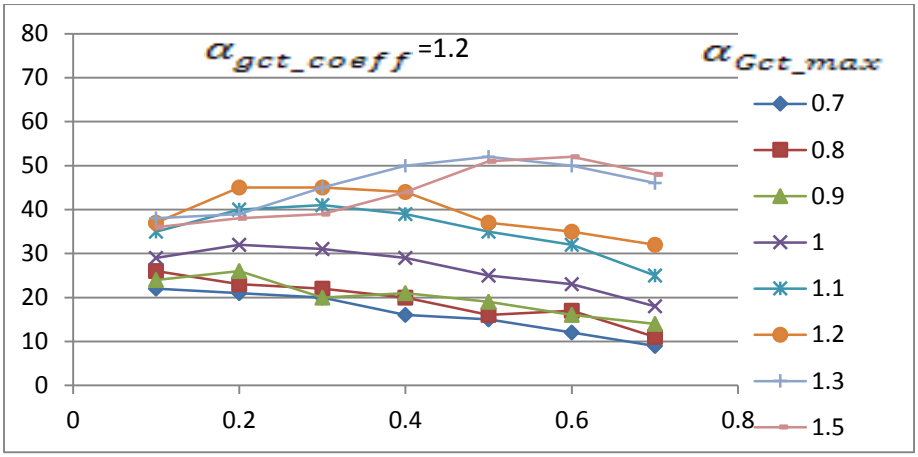
شکل ۱۵ – مقایسه مقادیر مختلف برای پارامترهای ورودی برای α_{gct_coeff} برابر با ۰٫۲. محور عمودی نشان دهنده میزان دقت و محور افقی نشان دهنده پارامتر α_{Gct_min} است که بازه ۰٫۱ تا ۰٫۷ برای آن در نظر گرفته شده است. البته مقادیر بیش از ۰٫۷ دارای دقت پایین تری بوده اند که نمایش داده نشده اند.



شکل ۱۶ – مقایسه مقادیر مختلف برای پارامترهای ورودی برای α_{gct_coeff} برابر با ۰٫۸. محور عمودی نشان دهنده میزان دقت و محور افقی نشان دهنده پارامتر α_{Gct_min} است که بازه ۰٫۱ تا ۰٫۷ برای آن در نظر گرفته شده است. البته مقادیر بیش از ۰٫۷ دارای دقت پایین تری بوده اند که نمایش داده نشده اند.

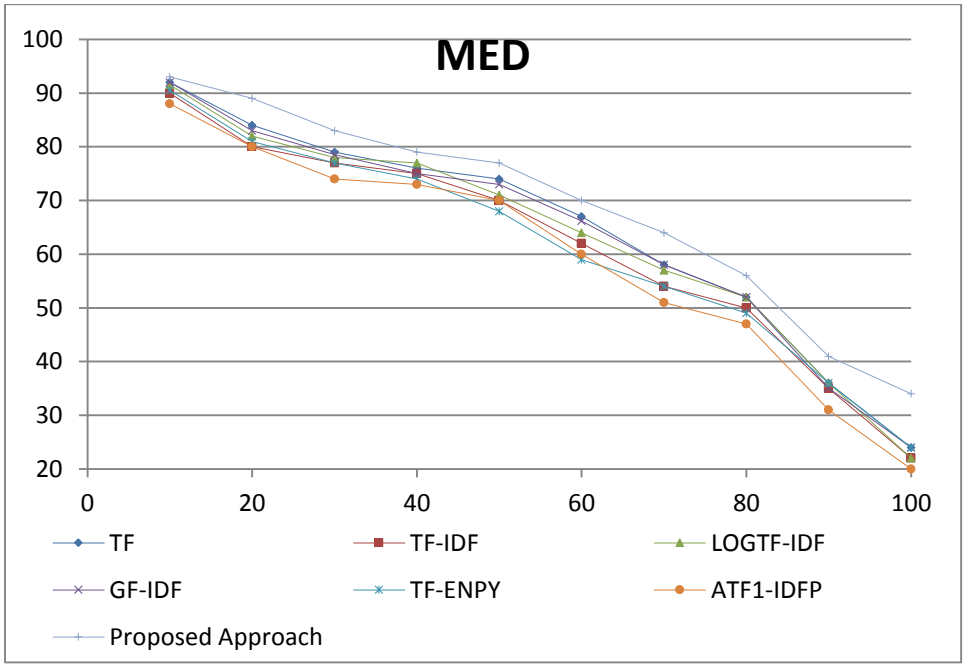


شکل ۱۷ – مقایسه مقادیر مختلف برای پارامترهای ورودی برای α_{gct_coeff} برابر با ۱٫۰. محور عمودی نشان دهنده میزان دقت و محور افقی نشان دهنده پارامتر α_{Gct_min} است که بازه ۰٫۱ تا ۰٫۷ برای آن در نظر گرفته شده است. البته مقادیر بیش از ۰٫۷ دارای دقت پایین تری بوده اند که نمایش داده نشده اند.

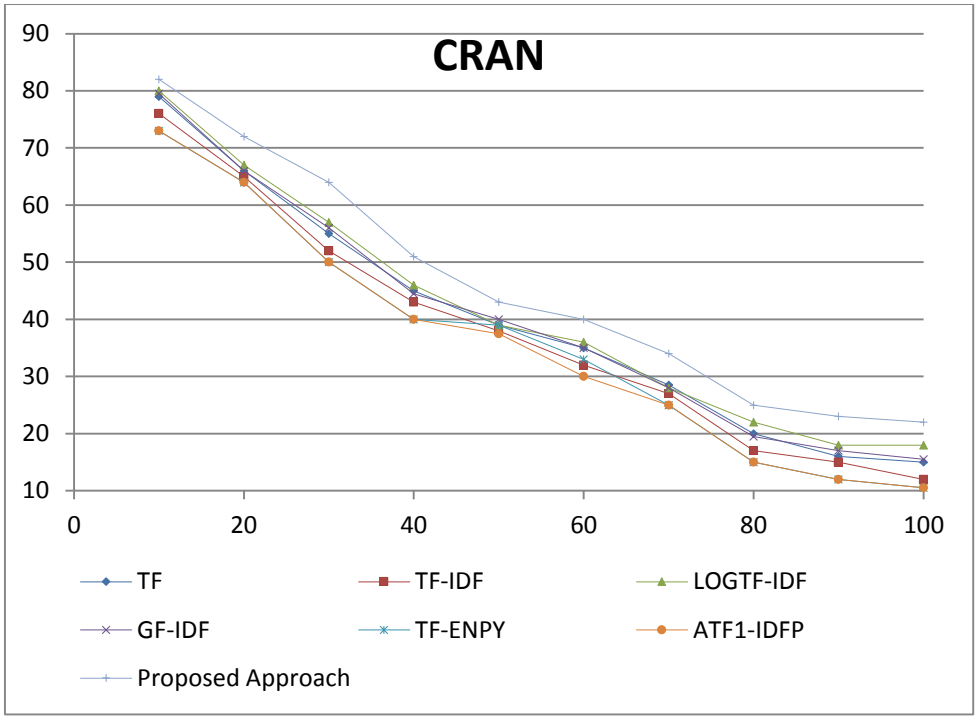


شکل ۱۸ – مقایسه مقادیر مختلف برای پارامترهای ورودی برای α_{gct_coeff} برابر با ۱,۲. محور عمودی نشان دهنده میزان دقت و محور افقی نشان دهنده پارامتر α_{Gct_min} است که بازه ۰,۱ تا ۰,۷ برای آن در نظر گرفته شده است. البته مقادیر بیش از ۰,۷ دارای دقت پایین تری بوده اند که نمایش داده نشده اند.

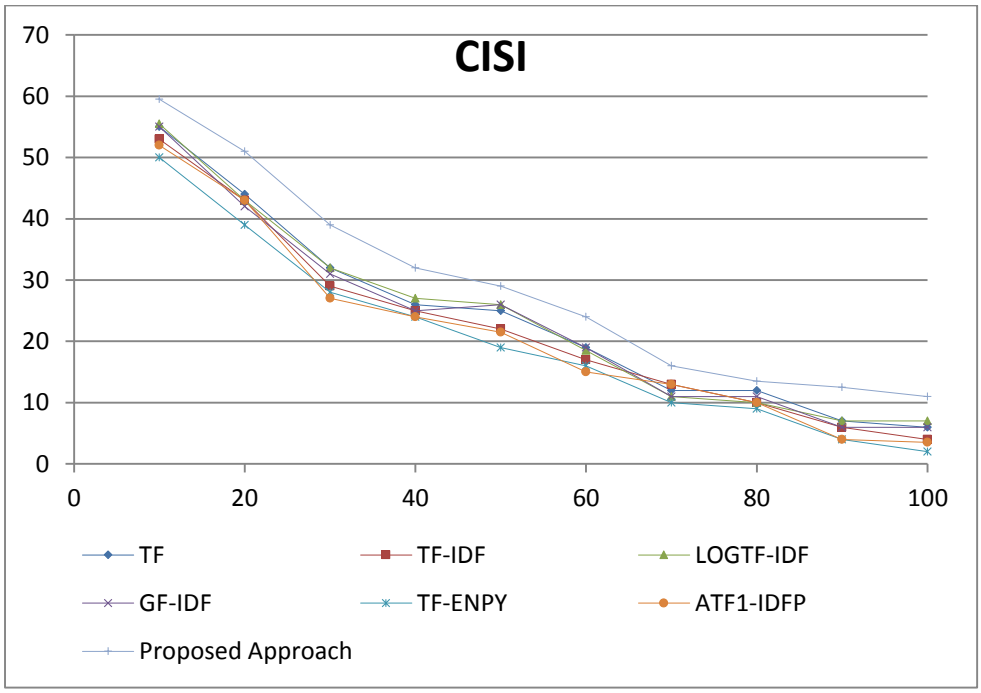
پس از این گام ها همه شرایط محیا برای اجرای آزمون اصلی و مقایسه روش ارائه شده با روش های متداول وزن دهی به لغات می باشد. نتیجه این مقایسه ها در نمایش داده شده است.



شکل ۱۹ – مقایسه منحنی Precision/Recall برای روش های وزن دهی مختلف در مجموعه داده آزمون MED. در نمودارهای زیرین راهنمای نمودار جهت صرفه جویی در فضا حذف شده است.



شکل ۲۰ – مقایسه منحنی Precision/Recall برای روش های وزن دهی مختلف در مجموعه داده آزمون CRAN. در نمودارهای زیرین راهنمای نمودار جهت صرفه جویی در فضا حذف شده است.



شکل ۲۱ – مقایسه منحنی Precision/Recall برای روش های وزن دهی مختلف در مجموعه داده آزمون CISI. در نمودارهای زیرین راهنمای نمودار جهت صرفه جویی در فضا حذف شده است.

براساس نتایج ارائه شده روش پیشنهادی دارای متوسط دقت بالاتری نسبت به روش های موجود است و همین مسئله انگیزه کافی برای قبول پیچیدگی های محاسباتی را فراهم می آورد. البته میزان این بهبود کمتر از میزان مورد انتظار بوده است که از جمله دلایل آن می توان به موارد زیر اشاره نمود:

- در مرحله تقویت یک قطعه معنایی مستقل از محتوای متن این تقویت انجام شده است. به عبارت دیگر تنها مبتنی بر دانش پیش زمینه و مستقل از محتوای کلی متن هر لغت دارای ارتباط معنایی با لغات یک قطعه معنایی به آن اضافه شده اند. در حالیکه می توان با در نظر گرفتن بقیه متن کلمات کمتری را به یک قطعه معنایی اضافه نمود و این امر به خوبی قابل دستیابی می باشد اما به عنوان یک کار آینده در نظر گرفته شده است.
- همچنین یکی دیگر از دلایل این بهبود کم در مقادیر را می توان تأثیر بالاتر قطعات معنایی ابتدایی نسبت به قطعات معنایی میانی و انتهایی در متن دانست. این مسئله نیز می تواند از طریق در نظر گرفتن قاب معنایی و یا دیگر روش های جلوگیری از یکجانبه نگری پوشش داده شود که در اینجا جهت کارهای آتی در نظر گرفته شده است.

در مجموع می توان این طور نتیجه گرفت که استفاده از روش هایی که به معنا اهمیت داده و از توابع حدس انسانی برای این امر استفاده می کنند نه تنها دارای تأثیر مناسبی برروی نتایج می باشد بلکه می توانند کارایی مناسبی را نیز به لحاظ پیچیدگی های محاسباتی داشته باشند.

۵,۳. ارزیابی ترکیب دو روش ارائه شده

در آخرین آزمون هدف بر این است تا ترکیب دو روش ارائه شده نیز برروی مسئله بازیابی اطلاعات آزموده شود. آنچه مسلم است با توجه به اینکه روش ارائه وزن دهی با استفاده از الگوریتم LSI توانسته است نتایج بهتری نشان دهد و همچنین روش COSIN نیز توانسته است نتایج بهتری نسبت به LSI از خود نشان دهد، در نتیجه ترکیب دو روش ارائه شده و استفاده از آن در مسئله پاسخگویی به پرس و جو نیز می تواند نتایج خوبی را نشان دهد. از این رو ارائه نتایج نمی تواند نشان دهنده میزان کارایی ترکیب این دو روش گردد. به همین خاطر در این بخش مسئله دیگری به عنوان مسئله آزمون انتخاب شده است تا قابلیت تعمیم روش های ارائه شده نیز آزموده شود. مسئله مورد هدف در این بخش مسئله خلاصه سازی است.

مسئله خلاصه سازی نه تنها یک مسئله بازیابی اطلاعات است بلکه وابستگی بیشتری به ساختار زبان دارد. همچنین کیفیت روش های ارائه شده تا کنون پایین است و کمی بهبود در نتایج به راحتی قابل شناسایی و همچنین قابل تفسیر خواهد بود. به همین خاطر از جنبه های مختلف می توان این مسئله را مسئله ای مناسب برای آزمون دانست. برای انجام این آزمون باید در خصوص تعدادی از جزئیات توضیح داده شود که در ادامه آمده است.

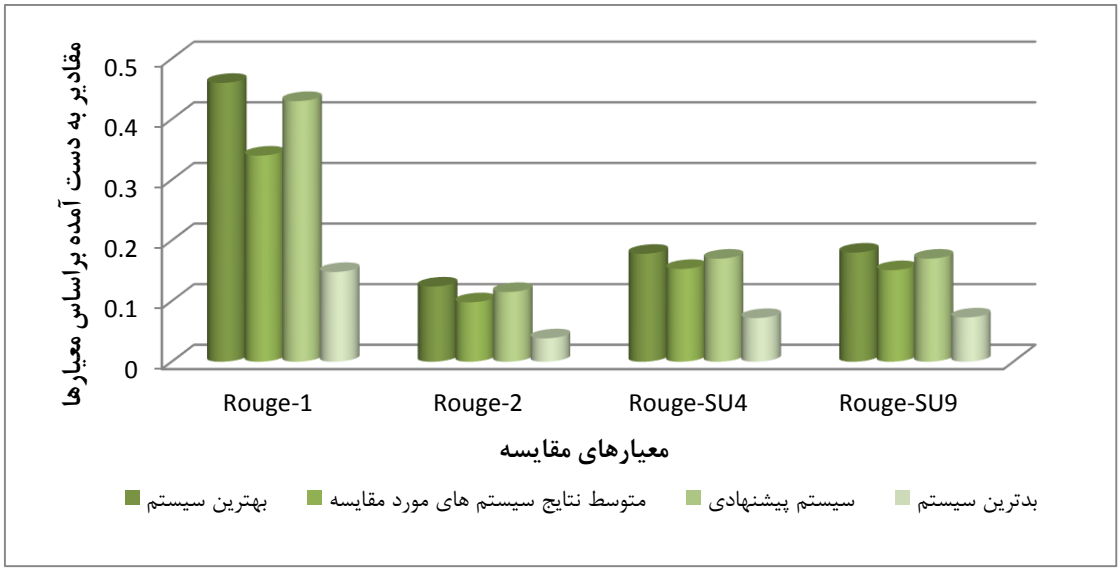
۵,۳,۱. مجموعه داده استاندارد جهت انجام آزمون

در مسئله خلاصه سازی مجموعه داده استاندارد مورد استفاده به صورت معمول مجموعه داده DUC می باشد. نسخه ای از این مجموعه داده که در دسترس بود DUC2007 می باشد.

۵,۳,۲. معیارهای اندازه گیری مورد استفاده در خلاصه سازی

در مسئله خلاصه سازی معیارهای ویژه ای برای اندازه گیری میزان دقت و کارایی روش های ارائه شده وجود دارند. این مجموعه معیارها مبتنی بر مقایسه با روش های مرجع می باشند. به عبارت دیگر در مسئله خلاصه سازی فرض می شود که برای یک مجموعه از اسناد مورد آزمون خلاصه هایی موجودند که این خلاصه ها توسط افراد و یا سیستم های مرجع تولید شده اند. سپس براساس میزان مطابقت نتایج به دست آمده از یک روش با نتایج دیگر روش های موجود مقایسه صورت می گیرد. یک دسته از این معیارها را به صورت کلی ROUGE[۷۱] می نامند [LIN04]. معیار تعیین میزان مطابقت حتی در معیار ROUGE نیز متنوع است و در بخش نمونه هایی از آن ها ذکر شده اند.

معیارهای مورد انتخاب در اینجا برای انجام آزمون عبارتند از ROUGE-1, ROUGE-2, ROUGE-SU4 و ROUGE-SU9 که براساس تجارب موجود می توانند بهترین مقایسه را انجام دهند [LIN04]. عدد موجود پس از ROUGE-SU نشان دهنده حداکثر فاصله مجاز بین دو کلمه در معیار مورد نظر می باشد.



شکل ۲۲ – مقایسه نتایج روش های مختلف توسط معیارهای مختلف برروی آزمون خلاصه صدکلمه ای در مجموعه داده DUC2007

[7] Recall-Oriented Understudy for Gisting Evaluation

۵،۳،۳. نتایج مقایسه

بر همین اساس DUC2007 استفاده شده است و آزمون خلاصه صد کلمه ای انجام شده است. مقایسه نیز برروی معیارهای ذکر شده در فوق انجام شده اند. نتایج این مقایسه در شکل ۲۲ آمده است. در اینجا تنها نتایج مربوط به سیستم پیشنهادی، بهترین سیستم، میانگین سیستم ها و بدترین سیستم ذکر شده اند.

همانطور که در نتایج مشاهده می شود این روش توانسته است تا نزدیک به بهترین سیستم عمل نماید. البته ذکر این نکته ضروری است که سیستم های مورد استفاده در مجموعه داده آزمون DUC2007 به عنوان سیستم های مرجع، سیستم های صرفا تحقیقاتی نیستند بلکه سیستم های تجاری بوده و با ترکیبی از روش های مختلف و بعضا استفاده از تکنیک های پیاده سازی و عملی تولید شده اند. از این رو انتظار نمی رود یک سیستم تحقیقاتی بتواند بهتر از همه این سیستم ها عمل نماید و مقایسه با متوسط روش های موجود در این مجموعه داده آزمون صورت می گیرد.

۶. نتیجه گیری و کارهای آتی

در این پایان نامه به دنبال ارائه راهکاری برای استخراج مفاهیم از متن به صورت خودکار مبتنی بر روش های زبان شناختی بودیم. با توجه به تنوع مشکلات موجود در این حوزه تقسیم بندی انجام شد و این مسئله در دو حوزه اصلاح روش های تعیین مشابهت و ورود معنا در وزن دهی مورد بررسی قرار گرفت. در نهایت نیز پس از ارائه روش های پیشنهادی، هر دو روش آزموده شدند و همچنین در مسئله خلاصه سازی متن به عنوان یک مسئله پیچیده بازیابی اطلاعات به صورت توأمان مورد آزمون قرار گرفتند.

مجموعه نتایج ارائه شده نشان از بهبود بسیار خوب در نتایج داشت. به نحوی که می توان گفت مجموعه هر دو روش ارائه شده می تواند بهبود بسیار مناسبی در نتایج ایجاد نماید. در ادامه سعی شده است به صورت خاص به بررسی دقیق تر نتایج روش های ارائه شده و ارائه مسیر آینده اقدام گردد.

۶,۱. نتیجه گیری

همانطور که گفته شد مسئله از دو دیدگاه تعیین مشابهت و همچنین وزن دهی به لغات مورد بررسی قرار گرفت. عمده مشکلات روش های موجود را می توان در موارد زیر خلاصه نمود:

- عدم استفاده از اطلاعات زبان شناختی در تعیین نقش و اهمیت لغات در سندها
- عدم توجه به ترتیب لغات در سند
- عدم مدل محاسباتی مناسب در روش های وزن دهی مبتنی بر دانش پیش زمینه
- عدم مدل مشخص جهت تعمیم روش های ارائه شده به مسائل بازیابی اطلاعات خاص
- وابستگی روش های ارائه شده در تعیین شباهت به تابع توزیع داده ها
- زمان اجرای بالای روش های ارائه شده برای تعیین شباهت

- حجم زیاد پارامترهای ورودی برای روش های تعیین شباهت
- پیچیدگی پیاده سازی بالای بعضی از روش های ارائه شده و عدم امکان پیاده سازی موازی و یا توزیع شده مناسب از آن ها

مجموعه این مشکلات می توانند تأثیر بسیار زیادی بر کیفیت نتایج ایجاد نمایند. تقریبا تمامی این مشکلات در این پایان نامه مورد توجه قرار گرفتند و سعی شد تا راه حل مناسبی برای آن ها ارائه گردد. در بخش وزن دهی به لغات سعی شد تا با استفاده از نظریه مرکزیت روش مناسبی در این حوزه ارائه گردد. همچنین در بخش تعیین شباهت نیز سعی شد تا با استفاده از مفهوم توابع ارتباط نسبت به رفع معضلات این حوزه اقدام شود. البته در استفاده از هر دوی این روش ها مشکلات دیگری مشاهده شدند که در این پایان نامه راه حل هایی نیز برای فائق آمدن بر این مشلات ارائه گردید.

در خصوص نظریه مرکزیت دو مشکل عمده شامل محلی بودن بیش از اندازه روش ارائه شده و عدم مدل محاسباتی مناسب برای روش مورد نظر وجود داشتند. از این رو سعی شد تا با ارائه تعاریف جدید نسبت به توسعه مفاهیم و همچنین اعمال تغییرات در نظریه مرکزیت و توسعه آن نسبت به رفع این مشکلات اقدام شود.

در حوزه تعیین شباهت نیز محاسبه توابع ارتباط مناسب و کارآمد دارای پیچیدگی محاسباتی زیادی بود که سعی شد تا با استفاده از گسسته سازی نتایج نسبت به بهبود نتایج اقدام شود. همچنین در انتخاب تابع مناسب ملاحظات مختلفی باید مد نظر قرار می گرفتند که سعی شد تا تمامی این ملاحظات بررسی شده و دلایل انتخاب توضیح داده شوند که از آن جمله می توان به اهمیت پوشش وابستگی های انتهایی در مسائل بازیابی اطلاعات اشاره نمود.

می توان اینگونه نتیجه گرفت که مجموعه عوامل اصلی در مسئله و همچنین عوامل تأثیرگذار و مهم در روش های انتخاب شده در حل مسئله بسیار متنوع و هر یک دارای ساختار خاص خود و الزامات مخصوص به خود بودند. اما در این پایان نامه سعی شد تا این گستردگی مسئله و شرایط تأثیرگذار تا حد زیادی پوشش داده شوند و نتایج تجربی ذکر شده در پایان نامه نیز شاهدی بر موفقیت مهار مناسب شرایط مسئله می باشند. در ادامه به توضیح در خصوص کارهای آتی که می توانند مسیرهای تحقیقاتی مناسبی باشند پرداخته شده است.

۶,۲.کارهای آتی

همانطور که در بخش قبل توضیح داده شد مجموعه عوامل تأثیرگذار و شرایط مورد انتظار در این مسئله بسیار متنوع و محدود کننده بودند. از این رو در نگاه اول راه های پیش رو بسیار زیاد هستند اما محدودیت

هایی که این عوامل ایجاد می نمایند از سوی دیگر می توانند این گزینه های پیش رو را تا حد زیادی محدود نمایند.

علیرغم این محدودیت ها در هر یک از دو حوزه تعیین شباهت موجودیت ها و همچنین تعیین میزان اهمیت لغات در صفحه فعالیت هایی به شرح زیر به عنوان کارهای آتی قابل ذکر می باشند. در حوزه تعیین شباهت ها کارهای پیشنهادی آتی به شرح زیر هستند:

- بررسی ویژگی های احتمالی داده ها در حوزه وب معنایی و ارائه یک تابع ارتباط مناسب در این حوزه می تواند یکی از کارهای مناسب برای تحقیقات بعدی باشد. بسیاری بر این باور هستند که داده های وب دارای توزیع پواسون می باشند. بررسی این موضوع و یافتن تابع ارتباطی که بتواند نتایج بهتری را در این حوزه ارائه نماید موضوع مناسبی برای تحقیقات بعدی خواهد بود.
- پیاده سازی موازی مناسب از روش ارائه شده یکی دیگر از موضوعات مناسب برای تحقیقات بعدی خواهد بود. همانطور که از الگوریتم ارائه شده بر می آید این الگوریتم دارای سادگی مناسبی می باشد و از این رو می توان پیاده سازی های موازی مناسبی برای آن ارائه نمود.

همچنین در حوزه روش ارائه شده در خصوص وزن دهی لغات نیز می توان موارد زیر را به عنوان پیشنهادات کارهای آتی ذکر نمود:

- استفاده از شرایط متن در تعیین میزان ارتباط لغات دانش پیش زمینه با هر لغت یکی از کارهای پیشنهادی می باشد. به عبارت دیگر در روش پیشنهادی تنها ارتباط معنایی بر اساس دانش پیش زمینه مورد استناد قرار گرفت. در حالیکه متن در میزان این ارتباط ها تأثیرگذار است و باید مورد توجه قرار گیرد.
- بررسی ترکیب های غیرخطی بین بردارهای زمینه و بردار جمله یکی دیگر از پیشنهادات برای کارهای آتی می باشد. در روش پیشنهادی صرفا ترکیب های خطی مورد نظر بودند، حال آنکه بعضی ترکیب های غیر خطی نیز می توانند نقش مناسبی در این حوزه ایفا نمایند.
- بررسی مجموعه های دانش پیش زمینه دیگر به غیر از WordNet برروی نتایج و همچنین استفاده از روش در مسائل دارای ساختارهای نا منظم و غیرمتداول متن یکی دیگر از حوزه های کاری قابل پیشنهاد برای روش ارائه شده وزن دهی می باشد.
- تعریف پنجره معنایی به جای در نظر گرفتن کل متن نیز یکی دیگر از پیشنهادات برای توسعه روش پیشنهادی می باشد. در حال حاضر کل متن مورد توجه قرار می گیرد و این باعث می شود تا

معنای برداشت شده از متن وابستگی بیشتری به جملات ابتدایی داشته باشد. استفاده از یک پنجره معنایی می تواند این وابستگی را کاهش دهد که البته باید تأثیر آن بررسی و آزموده شود.

مجموعه این پیشنهادات نکاتی هستند که می توانند در توسعه روش پیشنهادی به کار گرفته شوند.

۷. منابع

[AGI09] Eneko Agirre, Enrique Alfonseca, Keith Hall, Jana Kravalova, Marius Pasca, Aitor Soroa, A Study on Similarity and Relatedness Using Distributional and WordNet-based Approaches, Proceedings of NAACL-HLT 2009.

[AHM05] K Ahmad, L Gillam, Automatic ontology extraction from unstructured texts, Lecture notes in computer science, 2005 – Springer.

[AND83] Anderson, A., Garrod, S., Sanford, A. "The accessibility of pronominal antecedents as a function of episode shifts in narrative text". Quarterly Journal of Experimental Psychology, volume 35, pp. 427– 440, 1983.

[AND02] T Andreasen, PA Jensen, JF Nilsson, P Paggio, Ontological extraction of content for text querying, Lecture Notes in Computer Science, 2002, Springer.

[ANG02] Ang, A, Chen, J, "Asymmetric correlations of equity portfolios", Journal of Financial Economics 63 (3): 443–494, 2002.

[ANG05] Angkawattanawit, N., Rungsawang, A., Learnable Crawling: An Efficient Approach to Topic-specific Web Resource Discovery. Journal of Network and Computer Applications, Volume 28 , Issue 2 , April 2005.

[ARD05] Anders, Ardo, Focused crawling in the ALVIS semantic search engine, 2nd European Semantic Web Conference 2005, Heraklion Greece, 29 May- 1 June 2005.

[BAD03] Michael Bada, Robin Mcentire, Chris Wroe, Robert Stevens, GOAT: The Gene Ontology Annotation Tool, Proceedings of the 2003 UK e-Science All Hands Meeting.

[BAD09] Barak, L., Dagan, I., Shnarch, E. "Text categorization from category name via lexical reference". In NAACL ’09: Proceedings of human language technologies: The 2009 annual conference of the north american chapter of the association for computational linguistics, companion volume: Short papers, pp. 33–36, 2009.

[BAE05] R. Baeza-Yates and C. Castillo, M. Martin and A. Rodriguez, Crawling a country: better strategies than breadth-first for web page ordering. In Proc. 14th WWW, pages 864-872, ACM Press New York, NY, USA, 2005.

[BAR97] Barzilay, R., Elhadad, M. "Using lexical chains for text summarization". In Proceedings of the Workshop on Intelligent Scalable Text Summarization, Madrid, Spain, Aug, 1997.

[BAR09] Ziv Bar-Yossef, Idit Keidar, Uri Schonfeld, Do not crawl in the DUST: Different URLs with similar text, ACM Transactions on the Web, vol. 3 (2009), pp. 3.

[BER02] Bergmark, D., Lagoze, C. and Sbityakov, A., Focused Crawls, Tunneling, and Digital Libraries, 2002.

[BHA98] Bharat, K. and Henzinger, M. R., Improved algorithms for topic distillation in a hyperlinked environment, In Proceedings of SIGIR-98, 21st {ACM} International Conference on Research and Development in Information Retrieval.

[BHA04] Bhatt, M. Flahive, A. Wouters, C. Rahayu, W. Taniar, D. Dillon, T. A Distributed Approach to sub-Ontology Extraction, Advanced Information Networking and Applications (AINA), 2004.

[BLE03] Blei DM, Ng AY, Jordan M (2003) Latent Dirichlet allocation. Journal of Machine Learning Research (3): 993–1022.

[BLO04] Bloehdorn, S., Hotho, A. "Boosting for text classification with semantic features". In Proceedings of the MSW 2004 workshop at the 10th ACM SIGKDD conference, pp. 70–87, 2004.

[BLO06] Bloehdorn, S., Basili, R., Cammisa, M., Moschitti, A. "Semantic kernels for text classification based on topological measures of feature similarity". In ICDM ’06, Proceedings of the sixth international conference on data mining, pp. 808–812, 2006.

[BOY04] Boyd S., Vanderberghe L. Convex Optimization, Cambridge University Press, 2004.

[BRA94] De Bra, P., Houben, G., Kornatzky, Y., and Post, R., Information Retrieval in Distributed Hypertexts, In Proceedings of the 4th RIAO Conference, 481 - 491, New York, 1994.

[BRE87] Brennan, S., Friedman, M., Pollard, C. "A centering approach to pronouns". In Proc. of the 25th ACL, pp. 155–162, 1987.

[BRO92] Peter F. Brown, Peter V. Desouza, Robert L. Mercer, Jenifer C. Lai, Class-Based n-Gram Models of Natural Language, Computational Linguistics, 1992.

[CAI05] Cai D, He X, Han J (2005) Document Clustering Using Locality Preserving Indexing. IEEE Transactions on knowledge and data engineering 17(12):1624-1637

[CAR02] Carthy, J. "Lexical chains for topic tracking". Ph.D. Thesis. Ireland: University College Dublin, scalable text summarization, 2002.

[CHA76] Chafe, W. "Givenness, contrastiveness, definiteness, subjects, and topics". In Li, C., editor, Subject and Topic, Academic Press, New York, pp. 25–76, 1976.

[CHA99] Chakrabarti, S., van den Berg, M. and Dom, B., Focused Crawling: A New Approach to Topic-Specific Web Resource Discovery, In Proceedings of the 8th International WWW Conference, Toronto, Canada, May 1999.

[CHA02] Chakrabarti, S., Punera, K., and Subramanyam, M., Accelerated focused crawling through online relevance feedback, In WWW, Hawaii, May 2002. ACM.

[CHA08] Chang, Y. C., Chen, S. M., Liau, C. J. "Multilabel text categorization based on a new linear classifier learning method and a category-sensitive refinement method". Expert Systems with Applications, volume 34, pp. 1948–1953, 2008.

[CHE06] Chen X, Fan Y, Tsyrennikov V (2006) Efficient Estimation of Semiparametric Multivariate Copula Models. Journal of the American Statistical Association 101: 1228-1240

[CHO98] Junghoo Cho, Hector Garcia-Molina, Lawrence Page "Efficient Crawling Through URL Ordering." In Proceedings of the the World Wide Web conference (WWW), Brisbane, Australia, April 1998.

[CHO00] Junghoo Cho, Narayanan Shivakumar, Hector Garcia-Molina "Finding replicated Web collections." In Proceedings of ACM International Conference on Management of Data (SIGMOD), May 2000.

[CHO02] Junghoo Cho, Alexandros Ntoulas "Effective Change Detection Using Sampling." In Proceedings of the International Conference on Very Large Databases (VLDB), September 2002.

[CHO03] Junghoo Cho, Hector Garcia-Molina "Estimating frequency of change." ACM Transactions on Internet Technology, 3(3): August 2003.

[CHO07] Junghoo Cho, Uri Schonfeld "RankMass Crawler: A Crawler with High Personalized PageRank Coverage Guarantee" In Proceedings of the International Conference on Very Large Databases (VLDB), September 2007.

[CHR02] Junghoo Cho, Sridhar Rajagopalan "A Fast Regular Expression Indexing Engine." In Proceedings of the Internatioinal Conference on Data Engineering (ICDE), February 2002.

[CHR05] Christensen D (2005) Fast algorithms for the calculation of Kendall's τ. Computational Statistics 20 (1):51–62

[CLA09] Clauset, A., Shalizi, C. R. and Newman, M. E. J. Power-law distributions in empirical data, SIAM Review, to appear (2009).

[CLA12] Clarke D. (2012) A Context-Theoretic Framework for Compositionality in Distributional Semantics. Computational Linguistics 38(1): 41-71.

[CVE79] D. M. Cvetkovic, M. Doob, and H. Sachs, Spectra of Graphs, Academic Press, 1979.

[DAS06] S. Dasgupta, C. H. Papadimitriou, and U. V. Vazirani, Algorithms, McGraw-Hill 2006.

[DAS09] Das, P., Srihari, R. "Utterance Topic Model for Generating Coherent Summaries". Second Text Analysis Conference (TAC), National Institute of Standards and Technology(NIST), Gaithersburg, Maryland, USA, 2009.

[DEB03] Debole, F., and Sebastiani, F. 2003. Supervised term weighting for automated text categorization. In Proceedings of the 2003 ACM symposium on Applied computing, 784–788. ACM Press.

[DEE90] Deerwester S, Dumais ST, Furnas GW, Landauer TK, Harshman R (1990) Indexing by latent semantic analysis. J. of the American Society for Information Science 41: 391-407.

[DEM77] Dempster P, Laird NM, Rubin DB (1977) Maximum likelihood from incomplete data via the EM algorithm. J Royal Statist Soc 39:1-38

[DEN04] Deng, Z.-H.; Tang, S.-W.; Yang, D.-Q.; Zhang, M.; Li, L.Y.; and Xie, K. Q. 2004. A comparative study on feature weight in text categorization, Advanced Web Technologies and Applications, Lecture Notes in Computer Science Volume 3007, 2004, pp 588-597.

[DEN05] Denuit M, Lambert P (2005) Constraints on concordance measures in bivariate discrete data. Journal of Multivariate Analysis 93:40-57

[DEZ09] Deza M. M., Deza E., Encyclopedia of Distances, Springer-Verlag Berlin Heidelberg 2009.

[DIE98] Di Eugenio, B. "Centering in Italian". In Walker, M. A., Joshi, A. K., Prince, E. F., editors, Centering Theory in Discourse, Oxford, chapter 7, pp. 115–138, 1998.

[DIN04] Li Ding et al., "Swoogle: A Search and Metadata Engine for the Semantic Web", Proceedings of the Thirteenth ACM Conference on Information and Knowledge Management, November 2004.

[DIN05] Li Ding et al., "Finding and Ranking Knowledge on the Semantic Web", Proceedings of the 4th International Semantic Web Conference, November 2005.

[DIN06] Yihong Ding, Using Data-Extraction Ontologies to Foster Automating Semantic Annotation, Proc. PhD Workshop in 22nd International Conference on Data Engineering (ICDE 2006)

[DOU95] Dougherty J, Kohavi R, Sahami M, Supervised and Unsupervised Discretization of Continuous Features, In ML-95: Proceedings of the Twelfth International Conference on Machine Learning, pp. 194-202, San Francisco, CA: Morgan Kaufmann, 1995.

[DUM88] ST. Dumais, G.W. Furnas, T.K. Landauer, and S. Deerwester. Using latent semantic analysis to improve information retrieval. In Proceedings of CHI'88: Conference on Human Factors in Computing, New York: ACM, 281-285, 1988.

[EBA13] Eban, E; Rothschild, R; Mizrahi, A; Elidan, G, "Dynamic Copula Networks for Modeling Real-valued Time Series", in Carvalho, C; Ravikumar, P, Journal of Machine Learning Research 31, 2013.

[EHR03] Ehrig, M., Maedche, A., Ontology-focused Crawling of Web Documents, In Proceedings of the 2003 ACM symposium on Applied computing.

[ERN07] Ernandes, M. et al. "An Adaptive Context Based Algorithm for Term Weighting". Proceedings of the 20th international joint conference on Artifical intelligence, pp. 2748-2753, 2007.

[FAY93] Fayyad UM, Irani KB (1993) Multi-interval discretization of continuous-valued attributes for classification learning. In: Proc. of the 13th Int. Joint Conf. on Artificial Intelligence, Morgan Kauffman: 1022-1027

[FEL98] Christiane Fellbaum, WordNet: An Electronic Lexical Database, MIT Press, 1998.

[FEL07] Feldman D, Monemizadeh M, Sohler C (2007) A PTAS for k-Means Clustering Based on Weak Coresets. In: Proc. of the 23rd annual symposium on Computational Geometry SOCG 2007.

[FEN01] Dieter Fensel & Frank van Harmelen & Ian Horrocks & Deborah L. McGuinness & Peter F. Patel-Schneider. "OIL: An Ontology Infrastructure for the Semantic Web", IEEE Intelligent Systems, Pages 38-45, 2001.

[FIR68] Firth, John R. 1968. A synopsis of linguistic theory, 1930–1955. In John R. Firth, editor, Selected Papers of JR Firth, 1952–59. Indiana University Press, Bloomington, pages 168–205.

[FRE87] Fredman M L, Tarjan R E (1987) Fibonacci heaps and their uses in improved network optimization algorithms. Journal of the ACM (JACM), Volume 34 Issue 3, July 1987, Pages 596 – 615.

[GAO11] Gao K, Hardeman H, Lim E, Potter C, Meyer C, Abbey R (2011) An Approach to Identify the Number of Clusters, NCSU, REU projects

[GIV83] Givon, T. "Topic continuity in discourse : a quantitative cross-language study". J. Benjamins, editor, 1983.

[GOR93] Gordon, P. C., Grosz, B. J., Gillion, L. A. "Pronouns, names, and the centering of attention in discourse". Cognitive Science, volume 17, pp. 311–348, 1993.

[GRO77] Grosz, B. J. "The Representation and Use of Focus in Dialogue Understanding". PhD thesis, Stanford University, 1977.

[GRO83] Grosz, B., Joshi, A., Weinstein, S. "Providing a unified account of definite noun phrases in discourse". In Proc. ACL-83, pp. 44–50, 1983.

[GRO86] Grosz, B. J., Sidner, C. L. "Attention, intention, and the structure of discourse. Computational Linguistics", volume 12(3), pp. 175–204, 1986.

[GRO95] Grosz, B. J., Joshi, A. K., Weinstein, S. "Centering: A framework for modeling the local coherence of discourse". Computational Linguistics, volume 21(2), pp. 202–225, 1995.

[GUN93] Gundel, J. K., Hedberg, N., Zacharski, R. "Cognitive status and the form of referring expressions in discourse". Language, volume 69(2), pp. 274–307, 1993.

[GUP06] Vineet Gupta, Radha Jagadeesan, Prakash Panangaden, Approximate reasoning for real-time probabilistic processes, Logical Methods in Computer Science, vol. 2 (2006).

[HEF99] Heflin, J., Hendler, J., and Luke, S. SHOE: A Knowledge Representation Language for Internet Applications. Technical Report CS-TR-4078 (UMIACS TR-99-71), Dept. of Computer Science, University of Maryland at College Park. 1999.

[HEF01] Jeff Heflin, Towards the Semantic Web: Knowledge Representation in a Dynamic Distributed Environment, Ph.D. Thesis, University of Maryland, College Park, 2001.

[HEI82] Heim, I. "The Semantics of Definite and Indefinite Noun Phrases". PhD thesis, University of Massachusetts at Amherst, 1982.

[HEN00] J. Hendler and D. L. McGuinness. The DARPA Agent Markup Language. IEEE Intelligent Systems. 15 (6) , pp. 67-73. 2000.

[HER98] Hersovici, M., Jacovi, M., Maarek, Y., Pelleg, D., Shtalhaim, M., and Ur, S., The Shark-Search Algorithm – An Application: Tailored Web Site Mapping, In Proceedings of the Seventh International World Wide Web Conference, Brisbane, Australia, April 1998.

[HIR98] Hirst, G., St-Onge, D. "Lexical chains as representations of context for the detection and correction of malapropisms". In Christiane Fellbaum (Ed.), WordNet: An electronic lexical database. Cambridge, MA, The MIT Press, 1998.

[HIT05] Pascal Hitzler, Denny Vr, Resolution-based approximate reasoning for OWL DL, Proc. ISWC, 2005.

[HOF99] Thomas Hofmann, Probabilistic latent semantic indexing, Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval, 1999.

[HOF08] Hofmann, T, Schölkopf, B, and Smola, AJ (2008). Kernel Methods in Machine Learning, Annals of Statistics, 36:1171-1220.

[HOL93] Holte RC (1993) Very simple classification rules perform well on most commonly used datasets. Machine Learning 11: 63-90

[HUA05] Zhisheng Huang, Frank Van Harmelen, Annette Ten Teije, Reasoning with inconsistent ontologies, Proceedings of the International Joint Conference on Artificial Intelligence, IJCAI’05.

[IND08] P. Indyk, and R. Brinde, Sparse recovery using sparse random matrices, MIT-CSAIL-TR-2008-001, April 26, 2008.

[JON72] Karen Sparck Jones, A statistical interpretation of term specificity and its application in retrieval, Journal of Documentation Volume 28 Number 1 1972 pp. 11-21.

[JOS81] Joshi, A. K., Weinstein, S. "Control of inference: Role of some aspects of discourse structure–centering". In Proc. International Joint Conference on Artificial Intelligence, pp. 435– 439, 1981.

[KAM93] Kamp, H., Reyle, U. "From Discourse to Logic". D. Reidel, Dordrecht, 1993.

[KAM98] Kameyama, M. "Intra-sentential centering: A case study". In Walker, M. A., Joshi, A. K., Prince, E. F., editors, Centering Theory in Discourse, Oxford, chapter 6, pp. 89–112, 1998.

[KAN05] Kang, B.-Y., Lee S.J. "Document indexing: a concept-based approach to term weight estimation." Information Processing and Management: an International Journal, volume 41(5), pp. 1065 – 1080, 2005.

[KAR76] Karttunen, L. "Discourse referents". In McCawley, J., editor, Syntax and Semantics 7, Notes from the Linguistic Underground, Academic Press, New York, 1976.

[KER92] Kerber R (1992) ChiMerge: Discretization of numeric attributes. In: Proc. of 10th National Conf. on Artificial Intelligence, MIT press : 123-128

[KIT06] Yoshinobu Kitamura, Naoya Washio, Yusuke Koji, Riichiro Mizoguchi, Towards Ontologies of Functionality and Semantic Annotation for Technical Knowledge Management, New Frontiers in Artificial Intelligence, Lecture Notes in Computer Science, 2006.

[KOL98] Kolda TG, O'Leary DP (1998) A Semi-Discrete Matrix Decomposition for Latent Semantic Indexing in Information Retrieval. ACM Transactions on Information Systems 16: 322-346

[KOT06] Kotsiantis S, Kanellopoulos D (2006) Discretization Techniques: A recent survey. GESTS International Transactions on Computer Science and Engineering, Vol.32 (1), Pages 47-58, 2006.

[KRO93] R.Krovetz ,1993: "Viewing Morphology as an Inference Process", 16th ACM SIGIR Conference,Pittsburgh , June27-July 1, 1993;pp.191-202.

[KUL99] Kulpa T, On Approximation of Copulas, Int. J. of Math. Math. Sci. 22, 1999, 259-269.

[LAN06] Lan M., Tan C. L., Low H. B., Proposing a New Term Weighting Scheme for Text Categorization, In Proceedings of AAAI'06 Proceedings of the 21st national conference on Artificial intelligence, Volume 1, Pages 763-768, 2006.

[LEA03] David B. Leake, Ana Maguitman, Thomas Reichherzer, Topic Extraction and Extension to Support Concept, Proc. of the Sixteenth FLAIRS, 2003.

[LEO02] Leopold E., Kindermann J., Text categorization with support vector machines. How to represent texts in input space?, Journal of Machine Learning, Volume 46 Issue 1-3, 2002, Pages 423-444.

[LEO11] Leon R, Wu B (2011) Copula-based regression models for a bivariate mixed discrete and continuous outcome. Statistics in Medicine 30(2):175-185

[LET97] Letsche TA, Berry MW (1997) Large-Scale Information Retrieval with Latent Semantic Indexing. Information Sciences - Applications 100: 105-137

[LIN98] Lin, D. "An information-theoretic definition of similarity". In Proceedings of the 15th international conference on machine learning, pp. 296–304, 1998.

[LIN01] Lindenstrauss J., Johnson W.B. (eds.) Handbook of the geometry of Banach spaces, Elsevier, 2001.

[LIN04] Lin C. Y. “ROUGE: a Package for Automatic Evaluation of Summaries”. In Proceedings of the Workshop on Text Summarization Branches Out (WAS 2004), Barcelona, Spain, July 25 - 26, 2004.

[LON01] Longin, F, Solnik, B, "Extreme correlation of international equity markets", Journal of Finance 56 (2): 649–676, 2001.

[LUO11] Luo, Q., Chen, E., Xiong, H. "A semantic term weighting scheme for text categorization". presented at Expert Syst, Appl, pp. 12708-12716, 2011.

[LUU06] Toan Luu, Fabius Klemm, Ivana Podnar, Martin Rajman, Karl Aberer, ALVIS peers: a scalable full-text peer-to-peer retrieval engine, Proceedings of the international workshop on Information retrieval in peer-to-peer networks, 2006.

[LUX07] Luxburg UV (2007) A Tutorial on Spectral Clustering. Statistics and Computing 17 (4)

[MAE01] Maedche, A. Staab, S. Ontology learning for the Semantic Web, Journal of Intelligent Systems, IEEE, March-April 2001.

[MAN07] Gurmeet Singh Manku, Arvind Jain, Anish Das Sarma, Detecting near-duplicates for web crawling, WWW 2007 (16th International Conference on the World Wide Web), pp. 141-150.

[MAV07] Dimitrios Mavroeidis, Michalis Vazirgiannis, Stability Based Sparse LSI/PCA: Incorporating Feature Selection in LSI and PCA, Lecture Notes In Artificial Intelligence; Vol. 4701, 2007.

[MAY77] May, R. "The Grammar of Quantification". PhD thesis, MIT, Cambridge, MA, 1977.

[MAY85] May, R. "Logical Form in Natural Language". The MIT Press, 1985.

[MLA04] Mladenic, D.; Brank, J.; Grobelnik, M.; and Milic Frayling, N. 2004. Feature selection using linear classifier weights: interaction with classification models. In Proceedings of the 27th annual international conference on Research and development in information retrieval, 234–241. ACM Press.

[MOR88] Morris, J. "Lexical cohesion, the thesaurus, and the structure of text". Master_s thesis. Department of Computer Science, University of Toronto, 1988.

[MOR91] Morris, J., Hirst, G. "Lexical cohesion computed by thesaural relations as an indicator of the structure of text". Computational Linguistics, volume 17(1), pp. 21–43, 1991.

[MOR02] Mori, T. 2002. Information gain ratio as term weight: the case of summarization of ir results. In Proceedings of the 19th International Conference on Computational Linguistics, 688–694. Association for Computational Linguistics Publisher.

[NAJ01] Marc Najork, Allan Heydon, High-performance web crawling, SRC Research Report 173, Compaq Systems Research, 2001.

[NAW01] M. Najork and J. L. Wiener, Breadth-first crawling yields high quality pages. Proc. 10th WWW, pages114-118, 2001.

[NEL98] Nelsen RB (1998) An Introduction to Copulas, Lectures Notes in Statistics. Springer Verlag, New York

[NTO07] Alexandros Ntoulas, Junghoo Cho "Pruning Policies for Two-Tiered Inverted Index with Correctness Guarantee." In Proceedings of the Annual International ACM SIGIR Conference, July 2007.

[PAG98] L. Page, S. Brin, R. Motwani, T. Winograd, "The PageRank Citation Ranking: Bringing Order to the Web", Technical report, Department of Computer Science, Stanford University, 1998.

[PAP98] Christos H. Papadimitriou, Prabhakar Raghavan, Santosh Vempala, Latent semantic indexing: a probabilistic analysis, Proceedings of the seventeenth ACM SIGACT-SIGMOD-SIGART symposium on Principles of database systems, 1998.

[PAU64] Baran, Paul (1964). On Distributed Communications.

[PER95] Pereira C, Singer Y, Tishby N (1995) Beyond Word N-Grams. In: Proc. of the 3rd Workshop on Very Large Corpora

[POR80] M.F. Porter, An algorithm for suffix stripping, Journal of Program, 14(3) pp 130−137, 1980.

[POS06] Christian Posse, Patrick Paulson, Bob Baddeley, Ryan Hohimer, A White, Automating Ontological Annotation with WordNet, Proceedings of the 3 rd Global WordNet Conference, Jeju Island, South Korea, 2006.

[REE05] Lawrence Reeve, Survey of semantic annotation platforms, Proceedings of the 2005 ACM Symposium on Applied Computing, 2005.

[REI83] Reinhart, T. "Anaphora and semantic interpretation". Croom Helm, London, 1983.

[RUD07] Sebastian Rudolph, Johanna Völker, Pascal Hitzler, Supporting lexical ontology learning by relational exploration, Proceedings of ICCS 2007.

[SAL75] G. Salton , A. Wong , C. S. Yang, A vector space model for automatic indexing, Communications of the ACM, v.18 n.11, p.613-620, Nov. 1975.

[SAN81] Sanford A. J., Garrod, S. C. "Understanding Written Language". Wiley, Chichester, 1981.

[SCH08] Schölzel, C.; Friederichs, P. "Multivariate non-normally distributed random variables in climate research – introduction to the copula approach". Nonlinear Processes in Geophysics 15 (5): 761, 2008.

[SGA67] Sgall, P. "Functional sentence perspective in a generative description". Prague Studies in Mathematical Linguistics, volume 2, pp. 203–225, 1967.

[SHA04] Shawe-Taylor J., Cristianini N., Kernel Methods for Pattern Analysis, Cambridge University Press, 2004.

[SID79] Sidner, C. L. "Towards a computational theory of definite anaphora comprehension in English discourse". PhD thesis, MIT, 1979.

[SLA01] Slaney M, Ponceleon D (2001) Hierarchical segmentation using latent semantic indexing in scale space. In: the Proc. of the IEEE Int. Conf. on Acoustics, Speech, & Signal Processing

[SLE03] D Sleeman, S Potter, D Robertson, M Schorlemmer, Ontology extraction for distributed environments, Knowledge Transformation for the Semantic Web. IOS Press, 2003.

[SON09] Song W, Park SC (2009) Genetic algorithm for text clustering based on latent semantic indexing. Journal of Computers & Mathematics with Applications 57: 1901-1907

[SPA00] Spärck J K, Walker S, Robertson S E (2000) A Probabilistic Model of Information Retrieval: Development and Comparative Experiments (parts 1 and 2). Information Processing and Management, volume 36(6): 779-840.

[STR99] Strube, M., Hahn, U. "Functional centering–grounding referential coherence in information structure". Computational Linguistics, volume 25(3), pp. 309–344, 1999.

[SUN08] Sangsoo Sung, Seokkyung Chung, Dennis McLeod, Efficient Concept Clustering for Ontology Learning using an Event Life Cycle on the Web, Proc. 2008 ACM SYmposium on Applied Computing, pp. 2310-2314.

[SZA97] Szabolcsi, A. "Ways of Scope Taking". Kluwer, Dordrecht, 1997.

[THO11] Thompson, D, Kilgore, R, "Estimating Joint Flow Probabilities at Stream Confluences using Copulas", Transportation Research Record 2262: 200–206, 2011.

[TOD01] A Todirascu, F de Beuvron, D Galea, F Rousselot, Using description logics for ontology extraction, Proc. of ROMAND'2000 Workshop on Robust Parsing, 2001.

[TOU05] J.E. Tougas, H.Stern, R.J. Spiteri, Updating the partial singular value decomposition in latent semantic indexing, Tech. Rep., Department of Computer Science, University of Saskatchewan, 2005.

[TOU08] Jane E. Tougas, Raymond J. Spiteri, Two uses for updating the partial singular value decomposition in latent semantic indexing, Applied Numerical Mathematics, Volume 58 , Issue 4 (April 2008), Pages 499-510.

[TUR98] Turan, U. "Ranking forward-looking centers in turkish: Universal and language-specific properties". In Walker, M. A., Joshi, A. K., and Prince, E. F., editors, Centering in Discourse, Oxford University Press , chapter 8, pp. 139–160, 1998.

[VAL90] Vallduvi, E. "The Informational Component". PhD thesis, University of Pennsylvania, Philadelphia, 1990.

[VIL00] Viligenti, M., Coetzee, F., Lawrence, S., Giles, C. and Gori, M., Focused Crawling Using Context Graphs, In Proceedings of the 26th International Conference on Very Large Databases (VLDB 2000), Cairo, Egypt, September 2000.

[W3C92] http://www.w3.org/History/19921103-hypertext/hypertext/DataSources/WWW/Serve-rs.html

[W3OWL] http://www.w3.org/TR/owl-features/

[WAL94] Walker, M. A., Iida, M., Cote, S. "Japanese discourse and the process of centering". Computational Linguistics, volume 20(2), pp. 193–232, 1994.

[WAL98] Walker, M. A., Joshi, A. K., Prince, E. F. "Centering Theory in Discourse". Clarendon Press, Oxford, 1998.

[WAN08] Wang, P., Domeniconi, C. "Building semantic kernels for text classification using Wikipedia". In KDD ’08: Proceeding of the 14th ACM SIGKDD international conference, pp. 713–721, 2008.

[WEB78] Webber, B. L. "A formal approach to discourse anaphora". Report 3761, BBN, Cambridge, MA, 1978.

[WIKIP] http://en.wikipedia.org/wiki/Search_engine

[WIT53] Wittgenstein, Ludwig. 1953. Philosophical Investigations. Macmillan, New York. G. Anscombe, translator.

[WIT09] Wittek, P., Tan, C., Dar´anyi, S. “An ordering of terms based on semantic relatedness,” in Proceedings of IWCS-09, 8th International Conference on Computational Semantics, H. Bunt, Ed. Tilburg, The Netherlands: Springer, January 2009.

[WOL02] J. L. Wolf, M. S. Squillante, P. S. Yu, J. Sethuraman and L. Ozsen, Optimal crawling strategies for web search engines, In Proc. 11th WWW, pages 136-147, 2002.

[WRDNT] http://wordnet.princeton.edu/.

[WU81] Wu, H., and Salton, G. 1981. A comparison of search term weighting: term relevance vs. inverse document frequency. In SIGIR ’81: Proceedings of the 4th annual international ACM SIGIR conference on Information storage and retrieval, 30–39. New York, NY, USA: ACM Press.

[ZHA99] H. Zha and H. D. Simon. On updating problems in latent semantic indexing. SIAM J. Sci. Comput., 21(2):782–791, 1999.

۸. ضمائم

در این بخش سعی شده است تا نسبت به ارائه مواردی که به عنوان پیش زمینه دانشی مطالعه پایان نامه مورد نیاز هستند پرداخته شود. عمده این مطالب نقش مهمی در تعیین اهمیت و درک دقیق از روش های ارائه شده دارند. به همین علت در این بخش به توضیح کامل در خصوص هر یک از مفاهیم پرداخته شده و در طول پایان نامه نیز در هر قسمت فراخور موقعیت اشاره کوچکی به هر یک از این مطالب نیز شده است.

۸,۱. نمونه هایی از معیارهای شباهت

در جداول زیر نمونه هایی از معیارهای شباهت آورده شده اند. هدف در اینجا نمایش شباهت این معیارها با هم می باشد.

جدول ۱۳ - چند نمونه از معیارهای خانواده Lp

نام معیار	فرمول معیار
فاصله اقلیدسی L_2	$d = \sqrt[2]{\sum_{i=1}^{d} \lvert P_i - Q_i \rvert^2}$
فاصله قطعه شهر L_1	$d = \sum_{i=1}^{d} \lvert P_i - Q_i \rvert$
فاصله مینکوسکی[۲] L_p [TOU74]	$d = \sqrt[p]{\sum_{i=1}^{d} \lvert P_i - Q_i \rvert^p}$

[7] Minkowski [2]

فاصله چبیشف[۷۳] L_∞ [TAX04]	$d = \max_i \lvert P_i - Q_i \rvert$

جدول ۱۴ - چند نمونه از معیارهای خانواده L1

نام معیار	فرمول معیار
Sørensen [SOR48]	$d = \frac{\sum_{i=1}^{d} \lvert P_i - Q_i \rvert}{\sum_{i=1}^{d} (P_i - Q_i)}$
[GOW71] Gower	$d = \frac{1}{d} \sum_{i=1}^{d} \lvert P_i - Q_i \rvert$
[MON04] Soergel	$d = \frac{\sum_{i=1}^{d} \lvert P_i - Q_i \rvert}{\sum_{i=1}^{d} \max(P_i, Q_i)}$
Kulczynski [DEZ06]	$d = \frac{\sum_{i=1}^{d} \lvert P_i - Q_i \rvert}{\sum_{i=1}^{d} \min(P_i, Q_i)}$
Canberra [GOR99]	$d = \sum_{i=1}^{d} \frac{\lvert P_i - Q_i \rvert}{P_i + Q_i}$
Lorentzian [DEZ06]	$d = \sum_{i=1}^{d} \ln(1 + \lvert P_i - Q_i \rvert)$

جدول ۱۵ - چند نمونه از معیارهای خانواده اشتراک

نام معیار	فرمول معیار
اشتراک[۷۴] [DUD01]	$s = \sum_{i=1}^{d} \min(P_i, Q_i)$ $d = \frac{1}{2} \sum_{i=1}^{d} \lvert P_i - Q_i \rvert$
موج شکن[۷۵] [HED76]	$d = \sum_{i=1}^{d} \frac{\lvert P_i - Q_i \rvert}{\max(P_i, Q_i)}$
Czekanowski [GOR99]	$s = 2 \frac{\sum_{i=1}^{d} \min(P_i, Q_i)}{\sum_{i=1}^{d} P_i + Q_i}$

7 Chebyshev 3
7 Intersection 4
7 Wave Hedges 5

$d = \frac{\sum_{i=1}^{d} \lvert P_i - Q_i \rvert}{\sum_{i=1}^{d} P_i + Q_i}$	
$s = \frac{\sum_{i=1}^{d} \min(P_i, Q_i)}{\sum_{i=1}^{d} P_i + Q_i}$ $d = \frac{\sum_{i=1}^{d} \max(P_i, Q_i)}{\sum_{i=1}^{d} P_i + Q_i}$	[DEZ06] Motyka
$s = \frac{1}{d} = \frac{\sum_{i=1}^{d} \min(P_i, Q_i)}{\sum_{i=1}^{d} \lvert P_i - Q_i \rvert}$	Kulczynski [DEZ06]
$s = \frac{\sum_{i=1}^{d} \min(P_i, Q_i)}{\sum_{i=1}^{d} \max(P_i, Q_i)}$	[DEZ06] Ruzicka
$d = \frac{\sum_{i=1}^{d} (\max(P_i, Q_i) - \min(P_i, Q_i))}{\sum_{i=1}^{d} \max(P_i, Q_i)}$	Tanimoto [DUD01]

جدول ۱۶ - چند نمونه از معیارهای خانواده ضرب داخلی

نام معیار	فرمول معیار
ضرب داخلی[۷۶]	$s = \sum_{i=1}^{d} P_i . Q_i$
میانگین هماهنگ[۷۷]	$s = 2 \sum_{i=1}^{d} \frac{P_i . Q_i}{P_i + Q_i}$
فاصله کوسینوسی	$s = \frac{\sum_{i=1}^{d} P_i . Q_i}{\sqrt{\sum_{i=1}^{d} {P_i}^2} \sqrt{\sum_{i=1}^{d} {Q_i}^2}}$
Kumar-Hassebrook [VIJ90]	$s = \frac{\sum_{i=1}^{d} P_i . Q_i}{\sum_{i=1}^{d} {P_i}^2 + \sum_{i=1}^{d} {Q_i}^2 - \sum_{i=1}^{d} P_i . Q_i}$
[JAC01] Jaccard	$s = \frac{\sum_{i=1}^{d} P_i . Q_i}{\sum_{i=1}^{d} {P_i}^2 + \sum_{i=1}^{d} {Q_i}^2 - \sum_{i=1}^{d} P_i . Q_i}$ $d = \frac{\sum_{i=1}^{d} (P_i - Q_i)^2}{\sum_{i=1}^{d} {P_i}^2 + \sum_{i=1}^{d} {Q_i}^2 - \sum_{i=1}^{d} P_i . Q_i}$

[7] Inner Product [6]
[7] Harmonic Mean [7]

تاس[8] [CHA06]	$s = \frac{2\sum_{i=1}^{d} P_i.Q_i}{\sum_{i=1}^{d} {P_i}^2 + \sum_{i=1}^{d} {Q_i}^2}$ $d = \frac{\sum_{i=1}^{d} (P_i - Q_i)^2}{\sum_{i=1}^{d} {P_i}^2 + \sum_{i=1}^{d} {Q_i}^2}$

جدول ۱۷ - چند نمونه از معیارهای خانواده مجذور وتر

نام معیار	فرمول معیار
وفاداری[9] [DEZ06]	$s = \sum_{i=1}^{d} \sqrt{P_i.Q_i}$
Bhattacharyya [BHA43]	$d = \ln \sum_{i=1}^{d} \sqrt{P_i.Q_i}$
Hellinger [DEZ06]	$d = 2\sqrt{1 - \sum_{i=1}^{d} \sqrt{P_i.Q_i}}$
Matusita [MAT55]	$d = \sqrt{2 - 2\sum_{i=1}^{d} \sqrt{P_i.Q_i}}$
مجذور وتر[10] [GAV03]	$d = \sum_{i=1}^{d} (\sqrt{P_i} - \sqrt{Q_i})^2$ $s = 2(\sum_{i=1}^{d} \sqrt{P_i.Q_i}) - 1$

جدول ۱۸ - چند نمونه از معیارهای خانواده مجذور L_2

نام معیار	فرمول معیار
مجذور اقلیدسی	$d = \sum_{i=1}^{d} (P_i - Q_i)^2$
Pearson $\mathcal{X}^2$ [PEA00]	$d(P,Q) = \sum_{i=1}^{d} \frac{(P_i - Q_i)^2}{Q_i}$

[7] Dice
[8] Squared-Chord
[7] Fidelity
[8]
[9]
[0]

$d(P,Q)=\sum_{i=1}^{d}\frac{(P_i-Q_i)^2}{P_i}$	Neyman $\mathcal{X}^2$ [NEY49]
$d=\sum_{i=1}^{d}\frac{(P_i-Q_i)^2}{P_i+Q_i}$	مجذور $\mathcal{X}^2$ [GAV03]
$d=2\sum_{i=1}^{d}\frac{(P_i-Q_i)^2}{P_i+Q_i}$	احتمالی متقارن $\mathcal{X}^2$ [DEZ06]
$d=2\sum_{i=1}^{d}\frac{(P_i-Q_i)^2}{(P_i+Q_i)^2}$	واگرایی [COX01]
$d=\sqrt{\sum_{i=1}^{d}\left(\frac{\lvert P_i-Q_i\rvert}{P_i+Q_i}\right)^2}$	[DEZ06] Clark
$d=\sum_{i=1}^{d}\frac{(P_i-Q_i)^2(P_i+Q_i)}{P_i.Q_i}$	افزایشی متقارن $\mathcal{X}^2$ [ZEZ06]

جدول ۱۹ - چند نمونه از معیارهای خانواده بی نظمی شانون

نام معیار	فرمول معیار
Kullback-Leibler [KUL51]	$d=\sum_{i=1}^{d}P_i\ln\frac{P_i}{Q_i}$
[TAN01] Jeffreys	$d=\sum_{i=1}^{d}(P_i-Q_i)\ln\frac{P_i}{Q_i}$
واگرایی K [GAV03]	$d=\sum_{i=1}^{d}P_i\ln\frac{2P_i}{P_i+Q_i}$
[GAV03] Topsøe	$d=\sum_{i=1}^{d}\left(P_i\ln\frac{2P_i}{P_i+Q_i}+Q_i\ln\frac{2Q_i}{P_i+Q_i}\right)$
Jensen-Shannon [JIA91]	$d=\frac{1}{2}\sum_{i=1}^{d}\left(P_i\ln\frac{2P_i}{P_i+Q_i}+Q_i\ln\frac{2Q_i}{P_i+Q_i}\right)$
فاصله Jensen [TAN01]	$d=\sum_{i=1}^{d}\left(\frac{P_i\ln P_i+Q_i\ln Q_i}{2}-\frac{P_i+Q_i}{2}\ln\left(\frac{P_i+Q_i}{2}\right)\right)$

جدول ۲۰ - چند نمونه از معیارهای ترکیبی

نام معیار	فرمول معیار
[TAN95] Taneja	$d = \sum_{i=1}^{d} \frac{P_i + Q_i}{2} \ln \frac{P_i + Q_i}{2\sqrt{P_i . Q_i}}$
Kumar-Johnson [KUM05]	$d = \sum_{i=1}^{d} \frac{\left({P_i}^2 - {Q_i}^2\right)^2}{2(P_i . Q_i)^{\frac{3}{2}}}$
متوسط L_1 و L_∞ [KRA87]	$d = \frac{\left(\sum_{i=1}^{d} \lvert P_i - Q_i \rvert\right) + \left(\max_i \lvert P_i - Q_i \rvert\right)}{2}$

۸,۲. متغیرهای احتمالی

یکی از حوزه های مهم که در حوزه روش های استخراج مفاهیم مورد اشاره قرار گرفته اند انواع متغیرهای تصادفی هستند. در اینجا توضیح مختصری در خصوص هر یک از انواع متغیرهای تصادفی مورد استفاده داده شده است.

۸,۲,۱. متغیر تصادفی چند جمله ای[۸۱]

یکی از مهم ترین انواع متغیرهای تصادفی، متغیرهای چند جمله ای هستند. یک متغیر تصادفی چند جمله ای برروی یک بردار از رویدادها تعریف می شود. متغیر تصادفی چند جمله ای دارای یک پارامتر ورودی با اندازه تعداد رویدادها است که تعیین می کند به ازای هر یک از رویدادها متوسط احتمال موفقیت به چه میزان می باشد. این نوع از متغیرها در کاربردهای بسیار زیادی قابل استفاده می باشند. به عنوان نمونه در یک کاربرد خوشه بندی در اصل هر یک از رویدادها یک خوشه محسوب می شود و هر نمونه داده مشاهده شده باید در یکی از این رویدادها جای گیرد. همچنین در انواع کاربردهای دیگری نیز می توان از این متغیرها استفاده نمود. مهم ترین نکته در اینجا امکان ایجاد یک مدل منطقی از مسئله مورد بررسی برروی متغیرهای تصادفی چندجمله ای می باشد.

تعریف رسمی تابع توزیع چگالی این نوع از متغیرهای تصادفی به صورت زیر می باشد:

$$f(x_1, \dots, x_k; n, p_1, \dots, p_k) = P(X_1 = x_1 \; and \dots and \; X_k = x_k)$$
$$= \begin{cases} \frac{n!}{x_1! \dots x_k!} p_1^{x_1} \dots p_k^{x_k} & when \sum_{i=1}^{k} x_i = n \\ 0 & otherwise \end{cases} \qquad (۲۶)$$

در این تعریف باید مقادیر x_i غیر صفر باشند. که با توجه به شهود مطرح شده در فوق منطقی می باشد.

[8] Multinomial [1]

۸,۲,۲. متغیر تصادفی Dirichlet

این متغیر تصادفی حالتی توسعه یافته از متغیر تصادفی چندجمله ای می باشد. این نوع متغیر تصادفی دارای یک پارامتر ورودی مانند α می باشد. این متغیر تصادفی نیز مانند قبل برروی تعدادی رویداد عمل می نماید و برای یک بردار ورودی از مقادیر مانند x تعیین می نماید که احتمال این که هر یک از رویدادها با احتمال x_i رخ دهند به شرط اینکه هر یک از آن ها پیش از این $\alpha_i - 1$ بار مشاهده شده باشند به چه میزان است. به عبارت دیگر با توجه به یک دانش پیش زمینه از رخداد رویدادهای مورد نظر که با استفاده از متغیر α بیان می شود احتمال رخداد رویدادها بر اساس احتمال های مشخص شده توسط x به چه میزان است. این شهود بیان شده به میزان زیادی در درک تعاریف بعدی و اهمیت این تابع توزیع تأثیر دارد. به همین خاطر باید درک مناسبی از این متغیر تصادفی ایجاد شود.

تعریف رسمی از تابع توزیع چگالی این متغیر تصادفی در زیر آمده است.

$$
\begin{aligned}
&f(x_1, \dots, x_{k-1}; \alpha_1, \dots, \alpha_k) = \frac{1}{\beta(\alpha)} \prod_{i=1}^{k} x_i^{\alpha_i - 1} \\
&where \quad \beta(\alpha) = \frac{\prod_{i=1}^{k} \Gamma(\alpha_i)}{\Gamma(\sum_{i=1}^{k} \alpha_i)} \\
&and\ \Gamma(z) = \int_0^{\infty} t^{z-1} e^{-t} dt \\
&and\ \ x_k = 1 - \sum_{i=1}^{k-1} x_i
\end{aligned}
\qquad (۲۷)
$$

این نوع از متغیرهای تصادفی نیز دارای اهمیت زیادی در کاربردهایی مانند خوشه بندی می باشند. مهم ترین ویژگی آن ها خاصیت پیش گویانه آن ها می باشد که پیش از این تا حدودی توضیح داده شده است.

۸,۲,۳. توابع پیشین[۸۲]

در احتمالات یک تابع توزیع مانند P را تابع پیشین تابع توزیع دیگری مانند P' می گوییم اگر بتواند میزان عدم قطعیت در مورد مقادیر متغیر تصادفی با توزیع P' را پیش از رخداد آن متغیر تصادفی تعیین نماید. به عبارت دیگر باید بتواند تعیین کند دانش ما در خصوص متغیر تصادفی هدف تا چه میزان با واقعیت تطابق خواهند داشت.

---

[8] Prior [2]

۸,۲,۴. توابع پسین[۸۳]

در احتمالات یک تابع توزیع مانند P را تابع پسین تابع توزیع دیگری مانند P' می گوییم اگر بتواند میزان عدم قطعیت در مورد مقادیر متغیر تصادفی با توزیع P' را پس از رخداد آن متغیر تصادفی تعیین نماید. به عبارت دیگر باید بتواند تعیین کند دانش ما در خصوص متغیر تصادفی هدف تا چه میزان با واقعیت تطابق داشته است. به عبارت دیگر پس از مشاهده مقادیر واقعی باید این میزان عدم انطباق را بتوان با استفاده از یک تابع توزیع که به آن تابع پسین می گوییم توصیف نماییم.

۸,۲,۵. توابع پیشین مجتمع[۸۴]

چنانچه توابع توزیع پیشین و پسین یک تابع توزیع مانند P' دارای جنس یکسانی باشند آنگاه آن تابع توزیع تابع پیشین مجتمع تابع توزیع P' محسوب خواهد شد. ویژگی مهم این نوع توابع آن است که امکان ایجاد فرآیند یادگیری برای آن ها وجود خواهد داشت. به عبارت دیگر پیش از رخداد متغیر تصادفی رفتار آن را پیش بینی می نماییم. سپس بعد از رخداد رفتار آن را مدل می نماییم و در نهایت با توجه به یک جنس بودن دو تابعی که رفتار را پیش و پس از رخداد توصیف نموده اند اقدام به اصلاح پارامترهای تابع پسین و پیشین می نماییم.

این ویژگی به ایجاد یک سامانه یادگیرنده کمک خواهد نمود. حال با شناسایی توابع پیشین مجتمع برای انواع توابع توزیع احتمالی می توان برای مسائل مختلف یک مدل یادگیری را ایجاد نمود. به عنوان نمونه تابع توزیع پیشین مجتمع برای تابع توزیع چندجمله ای تابع توزیع Dirichlet است. با توجه به آنچه در خصوص تعریف این دو تابع نیز بیان نمودیم کاملا مشخص است که تابع توزیع Dirichlet رفتار تابع توزیع چندجمله ای را پیش بینی می نماید. از سوی دیگر تابع توزیع چندجمله ای همان طور که توضیح داده شد دقیقا توصیف کننده شرایط مسائلی مانند خوشه بندی می باشد.

۸,۳. فضاهای باناخ[۸۵]

یک فضای باناخ در ریاضی یک فضای برداری دارای تابع طول کامل[۸۶]می‌باشد. به عبارت دیگر یک فضای باناخ، فضایی است که مجهز به یک تابع طول باشد و نسبت به این تابع طول کامل باشد. در شکل رسمی‌این تعریف به صورت زیر خواهد بود.

---

[3] [8] Posterior
[4] [8] Conjugate Prior
[5] [8] Banach
[6] [8] Complete normed vector space

تعریف ۱: یک دنباله مانند $x_1, x_2, x_3, \ldots$ یک دنباله کوشی[۴۷]گفته می‌شود اگر برای هر عدد حقیقی مثبت مانند ε یک عدد مثبت صحیح مانند N وجود داشته باشد که به ازای تمام اعداد طبیعی مانند $m, n > N$ داشته باشیم

$$|x_m - x_n| < \varepsilon \qquad (۲۸)$$

تعریف ۲: یک فضای دارای طول[۴۸]مانند V یک فضای باناخ گفته می‌شود اگر برای هر دنباله کوشی مانند $\{v_n\}_{n=1}^{\infty} \subset V$ یک عضو مانند $v \in V$ وجود داشته باشد که $\lim_{n\to\infty} v_n = v$.

مهمترین مزیت این فضا جامعیت آن نسبت به انواع معمول در کاربردها می‌باشد. به عنوان مثال فضای برداری شامل اعداد حقیقی و تابع طول شناخته شده اقلیدسی نمونه ای از این فضا محسوب می‌شوند. همچنین فضای هیلبرت[۴۹]نیز نمونه ای از این فضا محسوب می‌شود. فضاهای هیلبرت بستر اصلی برای معیارهای شباهتی مانند فاصله کسینوسی محسوب می‌شوند که کاربرد فراوانی در داده کاوی دارند. از این رو این فضا به عنوان بستر اصلی طراحی روش‌های پیشنهادی مورد استفاده قرار گرفته است.

۸,۴. توابع هسته

یکی از مهم ترین ابزارها برای شناسایی روابط غیر خطی بین داده ها توابع هسته[۵۰]می باشند. یکی از ایده های اولیه در این پایان نامه جهت بهبود دقت روش های استخراج مفاهیم استفاده از همین روش ها بوده است. البته برای بیان دقیق نحوه کاربرد این ابزار باید توضیحات مختصری در خصوص تعاریف مرتبط داده شود. این تعاریف به صورت دقیق در [HOF08] و [SHA04] آورده شده اند که در اینجا به خلاصه ای از آن اشاره شده است.

یک تابع هسته برروی یک فضای ورودی مانند $\mathcal{X}$ تعریف می گردد. برروی چنین فضایی یک تابع هسته به صورت $\kappa: \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ نمایش داده می شود و باید یک نگاشت مثبت شبه قطعی بوده و ویژگی $\kappa(x, y) = \kappa(y, x)$ را نیز دارا باشد. به عبارت دیگر یک نگاشت باید دارای ویژگی به صورت زیر باشد تا به عنوان نگاشت مثبت شبه قطعی شناخته شود:

$$for\ any\ n \in N, \{c_i\}_{i=1}^{n} \in \mathbb{R}^n and\ \{x_i\}_{i=1}^{n} \in \mathcal{X}^n: \sum_{i=1}^{n}\sum_{j=1}^{n} c_i c_j \kappa(x_i, x_j) \geq 0 \qquad (۲۹)$$

---

[8] Cauchy Sequence [7]
[8] Normed [8]
[8] Hilbert Space [9]
[9] Kernel Functions [0]

همچنین براساس قضیه مرسر[۹۱]داریم، $\kappa: \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ یک تابع هسته است اگر و تنها اگر یک فضای ویژگی مانند $\mathcal{F}$ وجود داشته باشد که توسط یک تابع ضرب داخلی مانند $\langle .,. \rangle$ تجهیز شده باشد و نگاشتی مانند $\phi: \mathcal{X} \to \mathcal{F}$ وجود داشته باشد که به ازای همه $x, y \in \mathcal{X}$ داشته باشیم:

$$\kappa(x, y) = \langle \phi(x), \phi(y) \rangle \qquad (۳۰)$$

در اصل همین ویژگی دوم که از قضیه مرسر نتیجه می شود محور اصلی ویژگی های توابع هسته برای حل مسائل غیرخطی می باشد. به فضای ویژگی $\mathcal{F}$ اصطلاحا یک فضای قابل تکثیر هسته هیلبرت[۹۲]گفته می شود. چنین فضایی امکان انجام محاسبات غیرخطی در فضای $\mathcal{X}$ با استفاده از الگوریتم های خطی شناخته شده در فضای $\mathcal{F}$ را فراهم می آورد. به عبارت دیگر برای تبدیل محاسبات غیرخطی به محاسبات خطی، داده ها از فضایی مانند $\mathcal{X}$ به فضایی با ابعاد بالاتر مانند $\mathcal{F}$ نگاشت می شوند. در فضای مقصد با استفاده از خواصی مانند تنک[۹۳]بودن داده ها پیچیدگی محاسبات کاهش می یابد. اصطلاحا به چنین روشی فن هسته[۹۴]گفته می شود. یکی از نمونه های پرکاربرد توابع هسته تابع هسته گاوسی است که برروی فضایی که مجهز به سنجه ای مانند $d(x, y)$ باشد تعریف شده و برابر است با:

$$\kappa(x, y) = \exp(-\alpha d(x, y)^2) \qquad (۳۱)$$

البته پرداختن به توابع هسته احتیاج به فرصت بیشتری دارد اما در اینجا تنها برای پوشش تعاریف اولیه و ارجاعات بعدی در پایان نامه به همین میزان اکتفا گردیده است.

۸,۵. توابع ارتباط

به زبان ساده می‌توان توابع ارتباط را توابعی دانست که توابع توزیع حاشیه ای[۹۵]چندین متغیر تصادفی را به یکدیگر متصل می‌نماید. اما این تعریف نمی‌تواند به درستی مشخصه‌های توابع ارتباط را نمایان سازد. در نگاهی دقیق‌تر می‌توان گفت به ازای هر دو متغیر تصادفی مانند X و Y که دارای توابع توزیع مانند $F(x)$ و $G(y)$ و همچنین تابع توزیع دو جمله ای $H(x, y)$ می‌باشند، می‌توان نگاشتی را از فضای دو بعدی $[0,1] \times [0,1]$ به $[0,1]$ متصور شد که نگاشت بین توابع توزیع حاشیه ای این دو متغیر و تابع توزیع دو جمله ای آن‌ها می‌باشد. می‌توان نشان داد که این نگاشت یک تابع است که به چنین توابعی تابع ارتباط می‌گوییم.

---

[1] [9] Mercer's Theorem
[2] [9] reproducing kernel Hilbert space
[3] [9] Sparsity
[4] [9] Kernel trick
[5] [9] Marginal Distribution Function

برای تعریف رسمی از توابع ارتباط باید تعدادی تعریف و نماد را ابتدا تعریف نماییم. این تعاریف همگی از [NEL98] برداشت شده‌اند. فرض کنید خط اعداد حقیقی شامل $(-\infty,\infty)$ را با $\mathbb{R}$ نمایش می‌دهیم. همچنین خط اعداد حقیقی توسعه یافته شامل $[-\infty,\infty]$ با $\overline{\mathbb{R}}$ نمایش داده می‌شود. همچنین $\overline{\mathbb{R}}^2$ عبارت است از صفحه توسعه یافته اعداد حقیقی و برابر است با $\overline{\mathbb{R}} \times \overline{\mathbb{R}}$. یک مستطیل در $\overline{\mathbb{R}}^2$ عبارت است از ضرب کارتزین دو بازه بسته مانند $B = [x_1, x_2] \times [y_1, y_2]$. مربع واحد نیز با $\mathbb{I}^2$ نشان داده می‌شود و عبارت است از $[0,1] \times [0,1]$. یک تابع دو متغیره حقیقی مانند H، تابعی است که دامنه آن $DomH \subseteq \overline{\mathbb{R}}^2$ و برد آن $RanH \subseteq \mathbb{R}$ باشد.

تعریف ۳: فرض کنید S_1و S_2 دو مجموعه غیرتهی برروی $\overline{\mathbb{R}}$ باشند و هچنین H تابعی دومتغیره حقیقی است که دامنه آن $DomH = S_1 \times S_2$ است. همچنین یک مستطیل مانند $B = [x_1, x_2] \times [y_1, y_2]$ را در نظر بگیرید که تمام رئوس آن در دامنه H قرار دارند. در این صورت حجم B براساس تابع H عبارت است از:

$$V_H(B) = H(x_2, y_2) - H(x_2, y_1) - H(x_1, y_2) + H(x_1, y_1) \qquad (۳۲)$$

تعریف ۴: یک تابع دو متغیره حقیقی مانند H ۲-صعودی گفته می‌شود اگر به ازای تمام مستطیل‌هایی مانند B که در دامنه H قرار می‌گیرند داشته باشیم $V_H(B) > 0$.

تعریف ۵: اگر تابع دو متغیره حقیقی مانند H برروی $S_1 \times S_2$ تعریف شده باشد و a کوچک ترین عضو S_1 و b کوچک ترین عضو S_2 باشد آنگاه H را یک تابع مستقر[۹۶]می‌گوییم اگر به ازای هر x, y داشته باشیم:

$$H(a, y) = 0 = H(x, b) \qquad (۳۳)$$

تعریف ۶: یک تابع زیر ارتباط دو بعدی[۹۷]تابعی است مانند $\acute{C}$ با خواص زیر

- دامنه $\acute{C}$ برابر است با $S_1 \times S_2$ که مجموعه‌های S_1, S_2 زیر مجموعه $\mathbb{I}$ هستند.
- $\acute{C}$ تابعی مستقر و ۲-صعودی است.
- برای هر u در S_1 و v در S_2 داریم:

$$\acute{C}(u, 1) = u, \acute{C}(1, v) = v \qquad (۳۴)$$

تعریف ۷: یک تابع ارتباط دو بعدی، یک تابع زیر ارتباط دوبعدی است که دامنه آن $\mathbb{I}^2$ باشد.

---

[6] Grounded [9]

[7] 2-dimensional Sub-Copula [9]

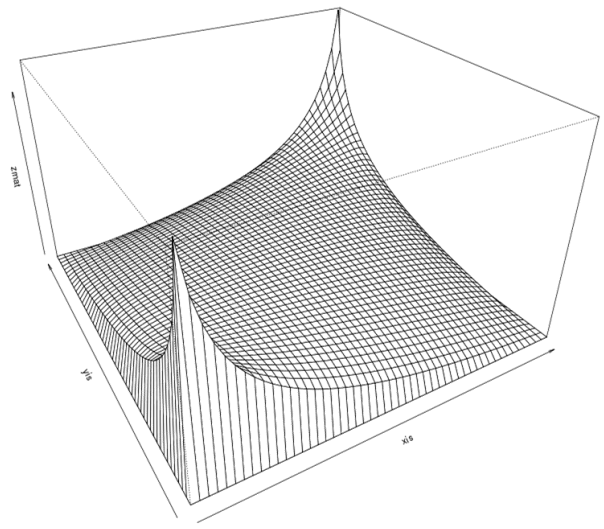
در اینجا تنها قضیه اساسی توابع ارتباط را به علت اهمیت فراوان آن بیان می‌نماییم.

قضیه ۱: (قضیه اسکالار)[۹][۸]فرض کنید H یک تابع توزیع دو جمله ای با توابع توزیع حاشیه ای F و G باشد. آنگاه حتما یک تابع ارتباط مانند C وجود دارد که به ازای هر x, y داشته باشیم:

$$H(x,y) = C(F(x), G(y)) \qquad (۳۵)$$

می توان گفت که تا اینجا تعریف دقیقی از توابع ارتباط ارائه گردید. اما آنچه بسیار اهمیت دارد خواص این توابع می‌باشد. مهم ترین خاصیت توابع ارتباط عدم وابستگی آن‌ها به نوع توزیع توابع حاشیه ای می‌باشد [NEL98]. این خاصیت باعث می‌شود تا توابع ارتباط ابزارهای بسیار مناسبی جهت استفاده در محیط‌های ناهمگن مانند وب باشند.

یکی دیگر از خواص مهم توابع ارتباط این است که وابستگی‌ها را در رخداد و عدم رخداد متغیرهای تصادفی شناسایی می‌نمایند. به عبارت دیگر اگر دو متغیر تصادفی دارای رخدادهای مشابه باشند به عنوان دو متغیر وابسته شناسایی می‌شوند. از سوی دیگر حتی عدم رخدادهای مشابه نیز می‌تواند نشان دهنده تشابه دو متغیر باشد. در حالی که روش‌های معمول تعیین مشابهت در داده کاوی نمی توانند چنین امکانی را در اختیار قرار دهند. این رفتار را می‌توان در شکل ۲۳ که نمودار تابع ارتباط گاوسی می‌باشد مشاهده نمود.



شکل ۲۳ – نمونه ای از تابع ارتباط گاوسی

همانطور که در شکل ۲۳ نمایان است تابع ارتباط گاوسی به ازای مقادیر نزدیک به یک و صفر مقادیر بزرگی را اختیار می‌کند.

---

[9] Sklar's Theorem [8]

همانطور که گفته شد توابع ارتباط تابع توزیع دوجمله ای دو متغیر تصادفی داده شده را محاسبه می‌نمایند. اما کاربرد این تابع در تعیین میزان وابستگی و یا همبستگی دو متغیر به چه صورت است. براساس توابع ارتباط می‌توان همبستگی دو متغیر تصادفی را به راحتی با استفاده از یکی از دو فرمول زیر محاسبه نمود:

$$correlation = \int_0^1 \int_0^1 C(u_1, u_2) dC(u_1, u_2) - 1$$
$$correlation = \int_0^1 \int_0^1 \{C(u_1, u_2) - u_1 u_2\} du_1 du_2 \qquad (۳۶)$$

شهود فرمول دوم کاملا مشخص می‌باشد. همانطور که می‌دانید برای دو متغیر مستقل تابع توزیع دو جمله ای همان ضرب توابع توزیع حاشیه ای آن‌ها می‌باشد. از این رو فاصله بین تابع ارتباط با تابع توزیع دوجمله ای ضرب می‌تواند نشان دهنده میزان وابستگی و یا استقلال دو متغیر نسبت به هم باشد.

علاوه بر آن نکته مهم بعدی نحوه به دست آوردن یک تابع ارتباط می‌باشد. به صورت کلی راهکارهای زیر برای به دست آوردن یک تابع ارتباط وجود دارند.

استفاده از توابع ارتباط عمومی: تعدادی از توابع ارتباط تا کنون در کاربردهای مختلف طراحی شده‌اند و به علت عمومیت و همچنین کارایی مناسب مورد توجه کاربردهای مختلف واقع شده‌اند. اولین و ساده ترین راهکار برای به دست آوردن یک تابع ارتباط استفاده از همین توابع عمومی می‌باشد. در جدول ۱ نمونه‌هایی از این توابع ارتباط مشاهده می‌شود.

جدول21 - نمونه هایی از توابع ارتباط متداول

نام تابع	فرمول تابع	مقدار پارامتر ورودی
FGM	$u_1 u_2(1+\theta(1-u_1)(1-u_2))$	$\theta = 3\rho$
Gaussian	$\int_{-\infty}^{\Phi^{-1}(u_1)} \int_{-\infty}^{\Phi^{-1}(u_2)} \frac{1}{2\pi(1-\theta^2)^{\frac{1}{2}}} \times e^{\left\{\frac{-(s^2-2\theta st+t^2)}{2(1-\theta^2)}\right\}} dsdt$	$\rho = \frac{6}{\pi}\arcsin(\frac{\theta}{2})$
Student's t-Copula	$\int_{-\infty}^{t_{\theta_1}^{-1}(u_1)} \int_{-\infty}^{t_{\theta_2}^{-1}(u_2)} \frac{1}{2\pi(1-\theta_2^2)^{1/2}} \times \left\{1+\frac{s^2-2\theta_2 st+t^2}{v(1-\theta_2^2)}\right\}^{-(\theta_1+2)/2} dsdt$	$\rho = \frac{6}{\pi}\arcsin(\frac{\theta}{2})$

هر یک از این توابع ارتباط دارای یک پارامتر ورودی می‌باشند که در اصل پارامتری برای همخوانی بیشتر تابع با داده‌ها می‌باشد و براساس پارامترهای همبستگی داده‌ها محاسبه می‌شوند. به عنوان مثال برای تابع

FGM مقدار پارامتر ورودی ۳ برابر ضریب همبستگی داده‌ها می‌باشد. از جمله دیگر توابع ارتباط عمومی می‌توان به Clyton، Frank، Gumbel، Ali-Mikhail-Haq و ... اشاره نمود [NEL98].

روش معکوس سازی: در این روش فرض بر این است که تابع توزیع دوجمله ای متغیرهای تصادفی مورد نظر موجود می‌باشد و از این طریق قصد داریم تابع ارتباط را به دست آوریم. به همین منظور از گزاره زیر استفاده می‌کنیم:

$$F(y_1, y_2) = C\big(F_1(y_1), F_2(y_2)\big) \Rightarrow C(u_1, u_2) = F(F_1^{-1}(y_1), F_2^{-1}(y_2)) \qquad (۳۷)$$

با استفاده از این گزاره که نتیجه مستقیم قضیه اسکالار می‌باشد می‌توان با استفاده از یک تابع توزیع دو جمله ای داده شده تابع ارتباط مورد نظر را به دست آورد.

مثال ۱: تابع توزیع دوجمله ای $F(y_1, y_2) = \exp\{-[e^{-y_1} + e^{-y_2} - (e^{-\theta y_1} + e^{-\theta y_2})^{-1/\theta}]\}$ را در نظر بگیرید. با توجه به این تابع توزیع دوجمله ای برای توابع توزیع حاشیه ای خواهیم داشت:

$$\begin{aligned} &\lim_{y_2 \to \infty} F(y_1, y_2) = F_1(y_1) = \exp(e^{-y_1}) \equiv u_1 \\ &\lim_{y_1 \to \infty} F(y_1, y_2) = F_2(y_2) = \exp(e^{-y_2}) \equiv u_2 \end{aligned} \qquad (۳۸)$$

در نتیجه خواهیم داشت:

$$y_1 = -\log(-\log(u_1)), y_2 = -\log(-\log(u_2)) \qquad (۳۹)$$

و با جایگزاری خواهیم داشت:

$$C(u_1, u_2) = u_1 u_2 \exp\{[(-\log u_1)^{-\theta} + (-\log u_2)^{-\theta}]^{-1/\theta}\} \qquad (۴۰)$$

این مثال به خوبی جزئیات این روش را نمایش می‌دهد. البته می‌توان به جای توابع نمایی و لگاریتم از نماد نیز استفاده نمود که این تابع ارتباط را تبدیل به یک تابع عمومی‌تر می‌نماید.

روش‌های جبری: در این دسته از روش‌ها کا با یک رابطه مفروض بین توابع توزیع حاشیه ای و با فرض مستقل بودن این توابع توزیع از هم شروع می‌شود. سپس با به دست آمدن روابط نهایی ارتباط بین این توابع حاشیه ای که پارامتر وابستگی به این رابطه اضافه شده و تابع ارتباط حاصل می‌شود. به علت پیچیدگی این روش از ذکر مثال در اینجا خودداری می‌نماییم. از جمله توابعی که از این طریق به دست آمده‌اند می‌توان به توابع Plackett و Ali-Mikhail-Haq اشاره نمود.

روش مجموع محدب: در این روش دو رویکرد وجود دارد. رویکرد اول ایجاد یک تابع ارتباط جدید از طریق جمع محدب بین توابع ارتباط کمینه، بیشینه و حالت استقلال متغیرهای تصادفی می‌باشد. در این

رویکرد مسئله مهم تعیین ضرایب دخیل در این تجمیع می‌باشد. یک نمونه از این رویکرد تابع ارتباط Frechet می‌باشد که در زیر آمده است:

$$C^F = \pi_1 C_L + (1 - \pi_1 - \pi_2)C^{\perp} + \pi_2 C_U \qquad (۴۱)$$

رویکرد دوم مبتنی بر انتگرال گرفتن از خانواده ای از توابع ارتباط می‌باشد که دارای پارامتری مانند η هستند و این پارامتر نیز دارای تابع توزیع دلخواه $\Lambda_\theta(\eta)$ می‌باشد. بر این اساس تابع ارتباط حاصل از رابطه زیر محاسبه می‌شود:

$$C_\theta(\vec{u}) = \int_{R(\eta)} C_\eta(\vec{u}) d\Lambda_\theta(\eta) \qquad (۴۲)$$

خانواده Archimedean: این خانواده از توابع ارتباط دارای شکل کلی زیر می‌باشند:

$$C(u_1, u_2; \theta) = \varphi^{-1}(\varphi(u_1), \varphi(u_2)) \qquad (۴۳)$$

و برای استفاده از این خانواده باید جایگزین مناسبی برای پارامترها یافت. بسیاری از توابع ارتباط را می‌توان عضوی از این خانواده دانست.

استفاده از روش‌های تخمین: یک دسته از روش‌ها مبتنی بر تخمین توابع ارتباط می‌باشند. از جمله این روش‌ها می‌توان به چندجمله ای‌های Bernstein اشاره نمود [KUL99]. اما محاسبه این تخمین‌ها دارای پیچیدگی محاسباتی بالایی می‌باشد.

روش‌های ذکر شده در فوق مهم ترین روش‌های ایجاد توابع ارتباط می‌باشند. همانطور که مشهود است این روش‌ها همگی مبتنی بر اطلاع از توابع ارتباط موجود و یا خواص داده می‌باشند. در خصوص داده‌های دارای ناهمگونی فراوان می‌توان گفت که چنین اطلاعاتی موجود نمی باشند، از این رو شاید یکی از بهترین راه‌ها برای شروع استفاده از توابع ارتباط عمومی باشد که در کاربردهای فراوانی توانسته‌اند نتایج مناسبی را نشان دهند.

۸,۵,۱. کاربردهای توابع ارتباط

توابع ارتباط دارای کاربردهای فراوانی در حوزه‌هایی که به محاسبه ارتباط بین متغیرهای تصادفی مختلف می‌پردازند می‌باشد. از این جمله می‌توان به موارد زیر استفاده نمود:

- **تخمین سرمایه گذاری:** یکی از کاربردهای اصلی توابع ارتباط در حوزه اقتصاد می‌باشد و یکی از نمونه‌های بارز آن محاسبات مربوط به تخمین سرمایه گذاری می‌باشد. در این حوزه نیز تنوع کاربردها از تعیین مدیریت ریسک سرمایه گذاری تا تعیین میزان اعتبار افراد در دریافت اعتبار

خرید می‌باشد. میزان استفاده از این تکنیک تا حدی گسترش یافته است که عده ای یکی از دلایل بروز بحران مالی این سال‌های کشورهای توسعه یافته را استفاده از تابع ارتباط گاوسی در محاسبه اعتبار افراد جهت اخذ وام و یا اعتبار خرید می‌دانند [LON01, ANG02].

- **مهندسی عمران:** یکی از کاربردهای توابع ارتباط محاسبات مربوط به بزرگراه‌ها و همچنین قابلیت اطمینان[9]سازه‌ها می‌باشد [THO11].
- **پزشکی:** در حوزه عصب شناسی کاربردهای موفقی از توابع ارتباط رخ داده است [EBA13].
- **هواشناسی:** در حوزه پیش بینی وضعیت هوا براساس تعیین وابستگی بین متغیرهای مختلف استفاده‌های زیادی از توابع ارتباط صورت گرفته است [SCH08].

۸,۶. نظریه مرکزیت

این نظریه به عنوان یک نظریه زبان شناختی که دارای قابلیت پیاده سازی محاسباتی را دارد، در این پایان نامه مورد توجه قرار گرفته است. در ادامه سعی شده است تا توضیحی مختصر در خصوص این نظریه بیان گردد.

۸,۶,۱. شهود اصلی نظریه مرکزیت

تئوری مرکزیت [JOS81، GRO83، GRO95] و [WAL98] یک کامپوننت از تئوری کلی تمرکز و انسجام گفتاری [GRO86] است [GRO77، SID79] و [GRO86] که حول انسجام و برجستگی محلی مطرح می‌شود. انسجام و برجستگی در یک سگمنت از گفتار مورد نظر می‌باشد. همانطور که در ادامه نیز خواهیم دید، تمامی تحقیقات انجام شده بر روی نظریه‌ی مرکزیت تاکنون، در مورد انسجام و برجستگی محلی، تأکید فراوانی دارند. این بدان معنی نیست که احتمالاً این نظریه توانایی خاصی در تشخیص انسجام و برجستگی در سطح مجموعه ای از اسناد مرتبط با هم و به صورت سراسری ندارد. بلکه به لحاظ بحث‌های تئوریک، این نظریه ادعایی در رابطه با تشخیص انسجام سراسری در سطح چندین و یا یک سند ندارد. این تئوری تنها در مورد ارتباط معنایی میان جملات متوالی ادعاهای قوی ارائه می‌دهد و مثلاً در مورد ارتباط معنایی میان جمله‌ی دوم و بیستم متن ادعایی ندارد.

از آنجایی که بسیاری از تلاش‌ها بر روی نظریه‌ی مرکزیت حول مباحث تئوریک آن بوده و کاربردهای تجربی آن کمتر مورد تحقیق قرار گرفته است، می‌توان ادعا نمود که بسیاری از توانایی‌ها بالقوه‌ی این نظریه مورد

---

[9] Reliability [9]

غفلت قرار گرفته است. به عنوان مثال تاکنون هیچ کار مبتنی بر یادگیری پیرامون توانایی یا عدم توانایی این نظریه در تشخیص انسجام و برجستگی سراسری صورت نگرفته است.

یک ویژگی اساسی مرکزیت، این است که، این تئوری شباهت بیشتری به یک تئوری زبانی دارد تا یک تئوری محاسباتی. بر این اساس اولین هدف این نظریه این است که، ادعاهای زبانی صحیحی، در مورد اینکه کدام قسمت گفتار از لحاظ پردازش ساده‌تر می‌باشد، ارائه دهد.

در نتیجه یک تئوری بسیار متفاوت با آنچه به طور معمول در زبان شناسی محاسباتی وجود دارد به دست می‌آید. نکته‌ی نگران کننده‌ی خاص در بسیاری از تئوری‌های زبان شناسی محاسباتی این حقیقت است که، مقالات مرکزیت مانند [GRO95] هیچ الگوریتمی برای محاسبه‌ی مفاهیمی مانند جمله و جمله‌ی قبلی و رتبه بندی و شناسایی که نقش مهمی در این تئوری بازی می‌کنند را ارائه نکرده‌اند. محققینی که بر روی مرکزیت کار می‌کنند، می‌گویند تا زمانی که این مفاهیم نقش مرکزی در تئوری‌های انسجام و مرکزیت بازی می‌کنند، بهتر است که ویژگی‌های مشخص آن‌ها برای تحقیقات متوالی دست نخورده باقی بماند. در واقع تعدادی از این مفاهیم مانند رتبه بندی با یک روش متفاوت برای هر زبان تعریف می‌شوند [WAL94]. به عبارت دیگر این مفاهیم باید به عنوان مفاهیم مرکزیت دیده شوند. این ویژگی تئوری الهام بخش یک تلاش تحقیقاتی قابل توجه برای خصوصی سازی پارامترهای مرکزیت برای زبان‌های مختلف است [KAM98، WAL94، DIE98، STR99] و [TUR98]. نسخه‌های رقابتی تئوری مرکزیت و ادعاهای مطرح در آن معمولاً به وسیله‌ی مؤلفین مشابهی ارائه شده است، مثلاً تعاریف مختلف برای CB در [GRO83، GRO95] و [GOR93] قابل مشاهده می‌باشد. در نتیجه محققی که می‌خواهد، تا یک پیش بینی در مورد مرکزیت را تست کند، یا از مرکزیت برای کاربردهای تمرینی استفاده کند، با تعداد زیادی از نمونه‌های ممکن این تئوری روبرو خواهد بود. این وضعیت برای یک تئوری زبانی غیر معمول نیست - به عنوان مثال، وضعیتی که در نحو و صرف با آن مواجه می‌شویم، زمانی که تعداد زیادی تعاریف متفاوت برای مفهوم مشابه دستور ارائه می‌شود - اما باعث می‌شود که مرکزیت با تئوری‌هایی که در زبان شناسی محاسباتی با آن روبرو می‌شویم، نسبتاً متفاوت باشد.

این کمبود جزئیات الگوریتمی نباید منجر به این نتیجه شود که مرکزیت فقط یک تلاش برای تعیین یک زمینه‌ی تحقیقاتی است، بدون اینکه ادعای خاصی را ارائه کند. در یک تضاد این تئوری ادعاهای بسیار قوی را ایجاد می‌کند که به زودی در مورد آن بحث می‌شود. این به این معنی است که، دو نمونه متفاوت از فریم ورک مرکزیت ممکن است، با ادعاهای مرکزی فریم ورک سازگار باشند، در عین حال پیش بینی‌های متفاوتی را ایجاد کنند، درست مثل راه‌های مختلف برای مشخص کردن دستور که ممکن است نتیجه‌ی آن، پیش بینی‌های متفاوت درباره‌ی نقض محدود سازی یا محدوده‌ی محتمل متفاوت برای خواندن یک جمله

می‌باشد [REI83، SZA97، MAY77] و [MAY85]. این آزادی در پر کردن خلأها الهام بخش محققانی است که، خود را وقف تهیه‌ی چنین تعاریفی، برای زبان‌های خاص یا به طور عمومی کرده‌اند، آنقدر که فریم ورک‌های مفهومی که از روی مقالات اخیر بر روی مرکزیت تولید شده‌اند، پایه‌ی بسیاری از کارها بر روی برجستگی محلی در زبان شناسی محاسباتی؛ حتی زبان شناسی در ۱۰ سال اخیر هستند. اما تا حدودی به خاطر وجود تعداد زیاد نمونه‌های رقابتی، تعدادی از محققین تردید شدیدی در مورد وضعیت تجربی این تئوری دارند: در مورد اینکه کدام ادعاها می‌توانند توسط ملاک‌های تجربی پشتیبانی شوند، و اینکه محققین چگونه می‌توانند پارامترهای ویژه را تحت تأثیر قرار دهند. برای اینکه این مقایسه با معنی باشد باید از مجموعه داده‌های مشابهی استفاده کرد.

۸،۶،۲. نظریه مرکزیت و پارامترهای آن

در این گردآوری فرصت آن نیست که به تئوری مرکزیت با تمام جزئیات آن پرداخته شود. در این جا کلیات این تئوری به تفصیل مورد اشاره قرار می‌گیرد. برای جزئیات بیشتر می‌توان به [GRO95] و [WAL98] مراجعه کرد.

۸،۶،۳. تمرکز محلی، مرکز رو به جلو و جمله‌ها

یک فرض بنیادی که تحت لوای مرکزیت وجود دارد، این است که پردازش یک گفتار دربرگیرنده‌ی بروزرسانی‌های متوالی وضعیت‌های محلی است، که به آن تمرکز محلی گویند. تمرکز محلی شامل مجموعه ای به نام CF می‌باشد که متناظر با کانون پتانسیل گفتار در [SID79] است و می‌تواند به صورت موجودیت‌های گفتار دیده شود [KAR76]، [WEB78]، [HEI82] و [KAM93]. تمرکز محلی همچنین شامل اطلاعاتی درباره‌ی برجستگی‌های مرتبط یا رتبه بندی موجودیت‌های CF است. این تمرکز محلی بعد از هر جمله بروز می‌شود: در این بروزرسانی CF فعلی با CF جدید و تغییرات CF جایگزین می‌شود.[۱۰۰] مجموعه‌ی CF که در تمرکز محلی با جمله‌ی U_{i} در سگمنت گفتار DS تعریف می‌شود به صورت $\mathrm{CF}(U_{\mathrm{i}}, \mathrm{DS})$ دیده می‌شود، که به طور کلی با مخفف $\mathrm{CF}(U_{\mathrm{i}})$ نمایش داده می‌شود. ارتباط بین جمله‌ها و CF[۱] بنا[۲] وسیله‌ی یکی از شرایط[۳] معروف آنها فرموله می‌شود[۱۰۳] [BRE87].

---

[۱۰۰] این فرضیه که پردازش گفتمان شامل بروز رسانی‌های ادامه دار مدل گفتمان است به عنوان قلب تئوری‌های معروف به گفتمان معنایی پویا نیز قرار دارد [KAR76]، [KAM93] و [HEI82].

[1] ‘Forward-Looking Center’ [0] [1]

[1] ‘Constraints’ [0] [2]

[۱۰۳] ترتیب مربوط به نمایش قوانین و محدودیت‌های مربوط به مرکزیت در این گردآوری با نمایش معمول آن‌ها در دیگر مقالات مربوط به مرکزیت تفاوت دارد. این به این خاطر است که در این جا سعی شده تا میان تعاریف و ادعاها با سه محدودیت ارائه شده توسط Brennan و دیگران تمایز ایجاد

محدودیت (۲): هر عضو از لیست CF(U,DS) باید در U قابل فهم و درک باشد (قابل شناسایی باشد).

۸,۶,۴. رتبه بندی مرکز مسندی و مرکز رو به جلو

اکنون دو ادعای مهم این تئوری را خاطر نشان می‌کنیم: یک این که لیست CF رتبه بندی شده است و دوم این که به خاطر این رتبه بندی بعضی از CFها به برجستگی ویژه ای دست می‌یابند. تابع رتبه بندی تنها می‌تواند به صورت نسبی اعمال شود، اما بالاترین رتبه در CF باید توسط جمله قابل شناسایی باشد (البته اگر وجود داشته باشد) و به اسم مرکز دارای ترجیح یا CP[۴]شناخته می‌شود. رتبه بندی همچنین در ساخت یکی از CFها که به CB مشهور است، استفاده می‌شود. CB در حقیقت نزدیک ترین مفهوم در مرکزیت به اندیشه‌ی عرفی و کلی موجود در موضوع متن [SGA67]، [CHA76]، [SAN81]، [AND83]، [GIV83]، [REI85]، [VAL90] و [GUN93] می‌باشد و نقش مرکزی و اصلی در کلیه‌ی ادعاهای این تئوری در مورد انسجام و برجستگی را بازی می‌کند. اگرچه در مقاله‌ی اصلی مرکزیت CB [GRO83] تنها در مورد کلمات شهودی توصیف شده است، اما بسیاری از کارهای زیر مجموعه ای که در فریم ورک مورد بحث در این گردآوری انجام شده، بر مبنای تعریف زیر از CB یک جمله و در رابطه با رتبه بندی [GRO95] کلمات قرار دارد:

محدودیت (۳): $\mathrm{CB}(U_{\mathrm{i}})$ جمله‌ی U_{i} است که بالا رتبه ترین عضو $\mathrm{CF}(U_{\mathrm{i}-1})$ است که در U_{i} قابل شناسایی می‌باشد [BRE87].

قابل ذکر است که بر طبق این تعریف محاسبه‌ی CB صرفاً وابسته به رتبه بندی و بیان قبلی است، که بر این اساس این پارامترها را برای فریم ورک مطرح شده بسیار مهم می‌شوند.

۸,۶,۵. گذارها

این فرضیه که می‌گوید گفتمان‌ها در صورتی که در قالب چنین روشی به صورت عبارت‌های پی در پی بسته بندی شوند، بسیار منسجم‌تر از زمانی که، عبارات با موجودیت گفتار منحصر بفرد شروع می‌شوند، هستند (مثال ۴ را نگاه کنید)، در تئوری مرکزیت فرموله شده است. این فرمول به عنوان یک اولویت برای روشهای معین بروزرسانی تمرکز محلی ارائه شده است. این اولویت به منظور کلاسه بندی عبارات بر طبق نوع گذار[۵](بروز سانی) آنها فرموله می‌شود که از طریق تمرکز محلی استنتاج می‌شود. تعدادی از این کلاسه

شود. این سه محدودیت وضعیت یکسانی ندارند: محدودیت دو می‌تواند به عنوان فیلتر حاکم بر متغیرهای مشخص دیده شود، محدودیت سه یک تعریف است و محدودیت یک یک ادعای تجربی می‌باشد.

[۱] - ‘Predicator-Looking Center’

[۱] Transition

بندی‌های گذارها ارائه شده است. [GRO95] بر اساس اینکه آیا مرکز لیست پویش عقبگرد عبارت Ui-1 در عبارت Ui دیده شده یا نه، و اینکه آیا CB(ui) بالا رتبه ترین (cp) موجودیت در Ui است یا نه، بین ۳ نوع از گذارها تفاوت قائل می‌شود.

ادامه مرکز (CON): $CB(U_i) = CB(U_{i-1})$ و $CB(U_i)$ بالا رتبه ترین عضو CF(CP) از عبارت U_i می‌باشد. (به عنوان مثال $CP(U_i) = CB(U_i)$)

حفظ مرکز (RET): $CB(U_i) = CB(U_{i-1})$ اما $CP(U_i) \neq CB(U_i)$.

شیفت مرکز (SHIFT): $CB(U_i) \neq CB(U_{i-1})$.

در ادامه چندین شمای متفاوت از کلاسه بندی‌ها مورد بررس قرار گرفته است. سپس در مورد اینکه چگونه از این کلاسه بندی‌ها برای فرموله کردن ادعاهای اصلی تئوری مرکزیت - قانون دو- استفاده می‌شود، بحث می‌کنیم.

۸,۶,۶. ادعاهای اصلی

طبق گفته‌ی [GRO95] بنیادی ترین ادعا در تئوری مرکزیت به این شرح است: " به میزانی که یک گفتار به قیدهای مرکزیت پایبند باشد، انسجام آن افزایش می‌یابد و بار استنتاجی که به شنونده وارد می‌شود کاهش می‌یابد" [GRO95]. آنها ۷ عدد از این قیدها را لیست کرده‌اند، که ۳ تای آن مستقیماً قابل ارزیابی می‌باشند. اگرچه در اینجا به دنبال تحلیل تفاوت میان قید و قانون که در [BRE87] آمده است نیستیم، اما می‌خواهیم از اسامی پیشنهادی [BRE87] برای این ۳ ادعا استفاده کنیم. که در حال حاضر با این اسامی شناخته شده‌تر هستند:

قید یک (Strong): همه‌ی عبارات یک سگمنت به جز اولین آنها دقیقاً یک CB دارند.

قانون یک (GJW95): اگر هر CF ای به ضمیر وابسته است، CB نیز به ضمیر وابسته خواهد بود.

قانون دو (GJW 95): یک توالی از ادامه‌ها بر یک توالی از حفظ‌ها ارجحیت دارد و به همین ترتیب یک توالی از حفظ‌ها بر یک توالی از شیفت‌ها ارجح می‌باشد.

۸,۶,۷. انتقال‌های معنایی نظریه مرکزیت

بر اساس پارامترهای تعریف شده در بخش ۲-۴-۳، انتقال‌های تعریف شده در بخش ۲-۴-۴ و قوانین و قیدهای تعریف شده در ۲-۴-۵، انتقال‌های معنایی چهارگانه نظریه مرکزیت که بر اساس آنها، جریان انتقال

معنا در متن تشخیص داده می‌شوند، تعریف می‌شوند. این انتقال‌ها در شکل ۲-۳ مشاهده می‌شوند. طبق ادعای [STR99] این چهار انتقال، مورد قبول قریب به اتفاق محققین در این زمینه قرار دارد.

جدول22 - قوانین تعیین گذارهای معنایی بین جملات

	$C_B(U_n) = C_B(U_{n-1})$ or $C_B(U_{n-1})\ is\ undef$	$C_B(U_n) \neq C_B(U_{n-1})$
$C_B(U_n) = C_P(U_n)$	Continue	Smooth-Shift
$C_B(U_n) \neq C_P(U_n)$	Retain	Rough-Shift

۸,۷. قالب کاری استاندارد کاربردهای مرتبط با مفهوم

در [CLA12] یک قالب کاری استاندارد در حوزه الگوریتم‌ها و روش‌های مرتبط با متن کاوی ارائه شده است. این قالب کاری با مطالعه عمومی روش‌های موجود و خواص ذاتی متن سعی کرده است تا حداکثر عمومیت را داشته و در عین حال مشخصات مورد نیاز برای یک الگوریتم در حوزه متن کاوی را بیان نماید. به همین دلیل در این پایان نامه سعی شده است تا از این قالب کاری به عنوان معیاری برای درستی‌یابی راه‌کارهای ارائه شده استفاده گردد. در ادامه جزئیات این قالب کاری و همچنین نمونه‌هایی از کاربرد آن بیان گردیده‌اند. تعریفی که برای نظریه زمینه[۶]ذ [CLA12] ارائه گردیده است به شرح زیر می باشد.

تعریف ۱ (نظریه زمینه)

یک **نظریه زمینه** یک پنجگانه است که به صورت $\langle A, \mathcal{A}, \xi, V, \psi \rangle$ نشان داده می شود و در آن:

- A مجموعه الفبای مورد استفاده می باشد. در اصل این مجموعه الفبا، مجموعه کلمات مورد استفاده در کاربرد مورد نظر می باشد.
- $\mathcal{A}$ یک جبر یکایی[۷]بزروی اعداد حقیقی می باشد. یک جبر برروی یک میدان مانند اعداد حقیقی در اصل عبارت است از یک فضای برداری برروی میدان مورد نظر که برروی آن یک عملگر خطی دوسویه[۸]مانند $(a, b) \mapsto ab$ تعریف شده است و داریم:

$$a(\alpha b + \beta c) = \alpha ab + \beta ac$$
$$(\alpha a + \beta b)c = \alpha ac + \beta bc \qquad (۴۴)$$

همچنین باید عضوی به نام عضو یکایی که با **1** نشان داده می شود داشته باشد که به ازای آن داریم:

---

[1] Context Theory [0] [6]
[1] Unital Algebra [0] [7]
[1] Bilinear Operator [0] [8]

$$1x = x1 = x \qquad (۴۵)$$

یکی دیگر از ویژگی های مورد نیاز در این جبر غیر جابجاپذیر بودن عملیات تجمیع می باشد. در حالت معمول برای تجمیع دو بردار از جمع معمول و یا ضرب داخلی استفاده می شود. این عملگرها جابجاپذیر می باشند اما در حوزه معنا تجمیع داده ها جابجاپذیر نیست. به این معنا که چند کلمه می توانند ترکیب های مختلفی داشته باشند و هر یک از این ترکیب ها دارای معانی مختلفی می باشند. حال آنکه در روش تجمیع معمول نتیجه همه این عبارت ها یکسان است.

- ξ یک نگاشت از A به $\mathcal{A}$ می باشد. به عنوان مثال فرض کنید که مجموعه الفبا شامل کلمات $\{\text{cat, pet, eat}\}$ می باشد. همچنین فرض کنید ۳ سند d_1، d_2 و d_3 را در اختیار داریم. آنگاه به حاصل نگاشت برای کلمه cat می تواند به صورت زیر باشد.

$$\xi(cat) = (1,2,5)$$

این نگاشت می تواند تفسیرهای مختلفی داشته باشد. به عنوان مثال می تواند نشان دهد که کلمه cat، ۱ بار در سند d_1، ۲ بار در سند d_2 و ۵ بار در سند d_3 استفاده شده است.

- V نیز یک فضای abstract Lebesgue می باشد. یک فضای abstract Lebesgue در اصل یک فضای باناخ[۹]است که برروی R^n می باشد و مجهز به $p\text{-}norm$ است. علت استفاده از این فضای اضافه این است که ممکن است فضای مورد استفاده برای ارائه و تجمیع معنایی خواص لازم برای استنتاج را نداشته باشد (ممکن است چنین خاصیتی در فضای ارائه وجود داشته باشد و از این رو این دو فضا یکی خواهند بود). از این رو خواص مورد نیاز برای استنتاج برای این فضای جدید تعریف گردیده و یک نگاشت نیز بین این دو فضا تعریف می گردد.
- ψ نیز یک نگاشت از $\mathcal{A}$ به V می باشد که پیش از این شرح آن رفت.

با توجه به این تعریف در صورت انطباق یک الگوریتم ارائه شده با نیازهای آن، می توان گفت که الگوریتم ارائه شده دارای همخوانی با مجموعه روش های دیگر در این حوزه دارد. برای درک بهتر از این قالب کاری به ارائه دو مثال پرداخته می شود. مثال اول در حوزه ارائه و مثال دوم در حوزه استنتاج می باشد.

مثال ۱ (جبر تنسور)[۱۰]

اگر V یک فضای بردار باشد، آنگاه فضای $T(V)$ تعریف می گردد که در اصل یک جبر آزاد از جبر تنسور تولید شده با V استفاده از می باشد و داریم:

---

[1] Banach Spaces [9]

[1] Tensor Algebra [10]

$$T(V) = \mathbb{R} \oplus V \oplus (V \otimes V) \oplus (V \otimes V \otimes V) \oplus \dots \qquad (۴۶)$$

به عبارت دیگر این فضای جدید عبارت است از مجموع تمام توان های تنسور فضای V. برای ترکیب دو بردار در این فضا از ضرب خارجی[۱] استفاده می گردد که نتیجه آن زیرمجموعه ای از فضای تعرریف شده در فوق می باشد. با توجه به اینکه هر بردار را می توان به صورت ترکیبی از ضرب های خارجی و جمع بین آن ها نوشت، از این رو می توان این فضا را به عنوان جبر مورد نیاز محسوب نمود.

همچنین عملگر تجمیع همان ضرب خارجی تعریف می گردد و با توجه به خطی دوسویه بودن این عملگر فرض های دیگر مورد نیاز برای جبر مورد استفاده جهت ارائه را دارا می باشد. به عنوان یک مثال ساده فضای $V = \mathbb{R}^3$ را در نظر بگیرید. آنگاه به ازای $\xi(cat) = (1,2,5)$ و $\xi(eat) = (2,0,3)$ خواهیم داشت.

$$\widehat{cat\ eat} = (1,2,5) \otimes (2,0,3) = \big(1(2,0,3), 2(2,0,3), 5(2,0,3)\big) = (2,0,3,4,0,6,10,0,15)$$

نتیجه در اصل به جای بیان در قالب یک ماتریس به صورت ارائه در فضای $\mathbb{R}^9$ بیان گردیده است. این روش ارائه می تواند به خوبی برای ترکیب بردارهای مختلف و سپس تعریف فرآیندهای مقایسه و استنتاج بردارها با هم مورد استفاده قرار گیرد.

مثال ۲ (تطابق LDA)

در این مثال سعی می شود تا تطابق روشی مانند LDA با قالب کاری ارائه شده بررسی گردد. البته جزئیات روش LDA قبلا در بخش ۲،۱،۳ توضیح داده شده است. در اینجا تنها به ارائه یک مدل که مطابق با قالب کاری ارائه شده و LDA باشد می پردازیم.

نظریه زمینه به صورت $\langle A, B\big(l^\infty(A^*)\big), \xi, l^\infty(A^*), \bar{p}\rangle$ را در نظر بگیرید که در آن

- A همان مجموعه الفبا و یا همان مجموعه کلمات مورد استفاده در کاربرد می باشد.
- $B(U)$ مجموعه تمام عملگرهای محدود[۲] بر روی U می باشد. یک عملگر محدود، عملگری است که نگاشتی را بین دو فضای مجهز به مقیاس[۳] انجام دهد و نسبت بین مقیاس نقاط نگاشت شده از عدد محدودی بیشتر نباشد. همچنین $l^\infty(A^*)$ مجموعه تمام توابع محدود حقیقی بر روی A^* می باشد.

---

[1] Tensor Product [1] [1]
[1] Bounded Operator [1] [2]
[1] Norm [1] [3]

- $\xi: A \to B\left(l^{\infty}(A^*)\right)$ عبارت است از تصویر کلمات برروی اسناد و مقادیر این نگاشت احتمال رخداد کلمات در اسناد می باشد که می توان آن را با $\xi(u) = P_u$ نمایش داد.
- و در نهایت $\bar{p}: B\left(l^{\infty}(A^*)\right) \to l^{\infty}(A^*)$ نگاشتی است که می توان به صورت $\bar{p}(T) = Tp$ نمایش داد و در آن $p \in l^{\infty}(A^*)$ است به صورتی که به ازای هر کلمه مانند u عبارت $\|P_u p\|_1$ برابر با احتمال رخداد کلمه u می باشد. با توجه به مشخصاتی که برای LDA گفته شد $\|P_x p\|_1$ برای یک عبارت مانند x به صورت زیر قابل محاسبه می باشد:

$$\|P_x p\|_1 = \int_{\theta} \left(\prod_{a \in x} p_{\theta}(a) \right) p(\theta) d\theta \qquad (۴۷)$$

که در آن $p_{\theta}(a)$ احتمال مشاهده کلمه ای مانند a در سندی است که با استفاده از پارامتر θ تولید شده باشد. این مقدار به صورت زیر قابل محاسبه می باشد:

$$p_{\theta}(a) \simeq 1 - \left(1 - \sum_{z} p(a|z) p(z|\theta)\right)^{N} \qquad (۴۸)$$

این نظریه مرکزیت اولا کاملا مطابق با قالب کاری ارائه شده بیان شده است و ثانیا همخوان با LDA می باشد. از این رو می توان گفت که LDA دارای همخوانی با قالب کاری ارائه شده می باشد. تنها مشکل از آنجا ناشی می شود که LDA یک روش جابجاپذیر است که از این نظر دارای همخوانی نیست و می توان گفت فاصله زیادی با واقعیت دارد.

۸,۱. نمونه هایی از چند معیار ROUGE

۸,۱,۱ ROUGE-N

این معیار براساس تعداد N-Gramهای مشترک بین داده های آزمون و داده های مرجع محاسبه می شود. به عبارت دیگر برای یک عدد ثابت مانند ۳ معیار ROUGE-3 به این صورت محاسبه می شود که معیار Recall برروی n-Gram ها بین نتایج روش پیشنهادی و خلاصه های مرجع تعیین می شود.

۸,۱,۲ ROUGE-L

این معیار براساس طولانی ترین زیر دنباله مشترک عمل می کند. در زیر عبارت کلی نحوه محاسبه این معیار آمده است:

$$R_{lcs} = \frac{LCS(X,Y)}{m}$$
$$P_{lcs} = \frac{LCS(X,Y)}{n} \qquad (۴۹)$$
$$F_{lcs} = \frac{(1+\beta^2)R_{lcs}P_{lcs}}{R_{lcs}+\beta^2 P_{lcs}}$$

که در آن $\beta = P_{lcs}/R_{lcs}$ و $LCS(X,Y)$ طول بزرگترین دنباله مشترک بین X و Y می باشد و m طول X و n طول Y می باشد که X خلاصه مرجع است و F مقدار معیار مورد نظر خواهد بود. در اینجا منظور از طول بزرگترین دنباله مشترک بیشترین تعداد کلمات مشترکی است که ترتیب رخداد آن ها در دو متن یکسان است. در زیر طول بزرگترین دنباله مشترک ۳ است.

جمله اول: «حسن به علی گفت به خانه برو.»

جمله دوم: «حسن با همراهی علی وارد خانه شد.»

۸،۱،۳. ROUGE-W

این معیار مشابه با معیار ROUGE-L است با این تفاوت که یک تابع وزن دهی به طول بزرگ ترین دنباله مشترک نیز اختصاص داده می شود. به عبارت دیگر با هر یک واحد افزایش در طول بزرگ ترین دنباله مشترک مقدار معیار به صورت خطی افزایش نمی یابد و متناسب با تابع وزن دهی تغییر می نماید.

۸،۱،۴. ROUGE-S

این معیار براساس skip-bigram ها تعریف شده است. یک skip-bigram یک دنباله دو کلمه ای از متن اصلی محسوب می شود. طبیعی است که برای متنی با ۱۰ کلمه تعداد skip-bigram ها برابر است با ترکیب ۱۰ از ۲. بر این اساس و مشابه با ROUGE-L این معیار به صورت زیر محاسبه می شود:

$$R_{skip2} = \frac{SKIP2(X,Y)}{C(m,2)}$$
$$P_{skip2} = \frac{SKIP2(X,Y)}{C(n,2)} \qquad (۵۰)$$
$$F_{skip2} = \frac{(1+\beta^2)R_{skip2}P_{skip2}}{R_{skip2}+\beta^2 P_{skip2}}$$

۸٫۱٫۵ ROUGE-SU

این معیار با اضافه نمودن یک کلمه مجازی به نام علامت شروع جمله به هر جمله و منطبق بر ROUGE-S به دست می آید. به عبارتی علاوه بر حالت های قابل پوشش توسط ROUGE-S تک کلمه های مشترک نیز توسط این معیار مورد بررسی قرار می گیرند.